Alma Mater Studiorum · University of Bologna

SCHOOL OF SCIENCE
Master degree in Computer Science

Detecting social patterns within 20th century documentary photos: a deep learning based approach

Thesis supervisor:
Gustavo Marfia

Thesis co-supervisor:
Giuseppe Lisanti

Presented by:
Lorenzo Stacchio

Session II
Academic Year 2019/2020

Contents

1 Introduction

The job of a historian is to understand what happened in the past, resorting in many cases to written documents as a firsthand source of information. Text, however, does not amount to the only and sole possible source of knowledge. Pictorial representations, in fact, have also accompanied and witnessed the main events of the historical timeline. In particular, the opportunity of visually representing key circumstances has bloomed since the invention of photography, with the possibility of capturing in real-time the occurrence of a specific events.

Technology keeps shaping the way of how different times and events are recorded for the benefit of future generations: between the 20th and 21st century we have been witnessing a sudden shift from the use of analog to digital photography devices with an unprecedented production of digital photos and videos [1]. Thanks to the widespread use of digital technologies (e.g. smartphones and digital cameras), networking capabilities and consequent availability of multimedia content, the academic and industrial research communities have developed artificial intelligence (AI) paradigms with the aim of inferring, transferring and creating new layers of information from images, videos, etc. [2]. For example, image data-sets are used to train deep learning algorithms, which are being increasingly successful in pursuit of many tasks (e.g., object recognition and segmentation).

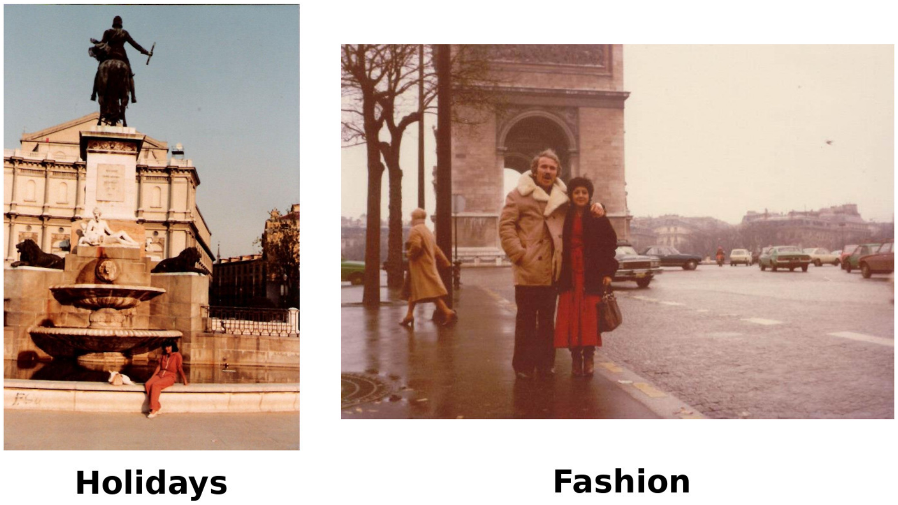
Now, while AI communities are devoting much of their attention to the processing and analysis of contemporary digital images, from a historical research standpoint the most interesting results may be obtained digitizing and analyzing corpus of analog images representing the pre-digital era. The main reasons why working with analog images is challenging are: (i) the photographs may be scattered in numerous public and private collections, (ii) of variable quality, and/or, (iii) damaged due to hard or continued use or exposure. Moreover, the analysis by means of modern artificial intelligence algorithms of such collections also requires their digitization, a process which may potentially introduce an additional amount of noise to the original photo.

Within the aforementioned scenario, the aim of this work is to address the analysis of collections of **analog documentary photographs**, building upon state-of-the-art deep learning techniques. Documentary photography, in particular, can be defined as the popular form of photography used to chronicle events or characteristic environments significant and relevant both to historical events as well as to everyday life. This class of photography has represented the recent history of mankind: first appearing in the first half of the nineteenth century, it spans over the entire twentieth century, i.e., the century where of today's phenomena and tendencies are rooted. Bodies of photos captured with cameras can hence be the key for better understanding the evolution in time of lifestyles, social conventions and customs and could also give us a hint of where they are going. Not only people but also objects within photos (e.g., cars, cycles) could have an important role to identify such kind of evolution.

In particular, the analysis carried out in this thesis focuses aims at producing the two following results: (a) produce the date of an image, and, (b) classify it recognizing its background socio-cultural context.

The dating task could be described as a **multi-classification problem** in which the model, given a photo, has to guess the date (usually the year) in which it was taken, analyzing the image itself. This problem was faced from different points of view from many researches [3, 4, 5, 6]. However, none of these methods has exclusively used analog pictures.

The socio-cultural context classification task could also be described as a **multi-classification problem**. In this case the model, given a photo, is required to return its socio-cultural context, as defined by a group of historical-sociological researchers [7]. This is a quite complicated task, since different labels, defined from a sociological and a cultural point of view, could be assigned to very similar images. For example in Fig.1 we can see that the picture here labeled as belonging to a **Fashion** class is conceptually close to another picture instead labeled belonging to a **Holidays** one, and viceversa. To the best of our knowledge, such type of analysis is performed for the first time in this work.



Holidays **Fashion**

Figure 1: Similarity between different socio-cultural classes.

Given these premises, the contribution of this work amounts to: (i) the introduction of an historical dataset including images of “Family Album” among all the twentieth century, (ii) the introduction of a new classification task regarding the identification of the socio-cultural context of an image, (iii) the exploitation of different deep learning architectures to perform the image dating and the image socio-cultural context classification.
This work is organized as follows. In this Section, a review is provided of the basic deep learning and computer vision concepts. In Section 2, the dataset that has been analyzed in this thesis is introduced. Sections 3 and 4 first carry out literature reviews concerning dating and ”socio-cultural context classification“, respectively, and then present the models devised for the analysis of the dataset considered in this thesis. Finally, Section 5 sets the stage for possible improvements and developments of this work.

1.1 Analog photography

The term “Analogue Photography” refers to photography using an analogue camera and film. A roll of film is loaded into the camera and the magic begins once you start clicking: light interacts with the chemicals in the film and an image is recorded. The pictures collected in your film roll come to life when the film is processed in a photo lab [8]. What we expect from analog photos, differently from digital once is the presence of “photography mistakes” like light leaks (white or red streaks on film created by stray light that enters a camera body) and shoot errors. Moreover, we have to keep in mind that different film labs do not use the same chemicals and calibrations, so you’ll get varied results with your photos (this could not happen with digital cameras of course).
From this notions we could infer that analog pics will probably be full of noise and that cannot perfectly describe the scene that the photographer wanted to portray. Again, the are only few ways to collect digitally analog pics which add even more noise (i.e. scanner, photo taken with smartphone). So far, what is clear is that analyzing this kind of photo from a purely Computer Science point of view seems a very difficult challenge.

1.2 Documentary (analog) photography

Documentary photography usually refers to a popular form of photography used to chronicle events or environments both significant and relevant to history and historical events as well as everyday life. It is typically covered in professional photojournalism, or real life reportage, but it may also be an amateur, artistic, or academic pursuit [9].
Indeed, documentary photographers, are expected to capture the world or everyday life, as it exists, without stage managing or directing or editing the scene. The origins of documentary photography are rooted in the very human desire to affect social change

[10]. First examples of documentary photography date back to **1850** when archaeologist John Beasly Greene traveled to Nubia to photograph the major ruins of the region (2a) and again in **1861** when many photographic publisher-distributors like Mathew Brady and Alexander Gardner took trace of the progress of the American Civil War by means of documentary photography (2b).

(a) John Beasly Greene's photo of the Abu Simbel temples, 1854

(b) New York State Militia in front of a tent, 1861

Documentary photography have been made with analog devices until digital cameras entered the scene. Nowadays, documentary photography **is almost exclusively made with digital cameras**, even if some lovers of analog photography are trying to bring it back into vogue because of his unique style. The reason why **analog and documentary photography** concepts have been explored will be more clear in the Section 2.

1.3 Analyze an enormous amount of images in the Big data era

In this section will be explored the steps that took Computer Scientists to develop tools which are able to extract and analyze knowledge from a big amount of images and then set the stage to connect the world of Big Data with both analog photography and machine learning ones.

1.3.1 Images in the Big data era

The amount of digital content that is daily generated and stored has greatly expanded within a short period of time [11]. This fact is nowadays simply experienced by all those who carry a smartphone in their pocket: the advancements in mobile platforms, fostered by the growth in terms of performance of hardware (e.g., sensors, communications, computing, storage) and application support of software solutions (e.g., social networks and cloud ecosystems), naturally provide the means to produce, collect and upload great quantities of data [12].

This means that quite a lot of digital images are produced and uploaded to Internet every day and most of them are uploaded in social network platforms (i.e. Instagram users upload more than 100 million pics per day in the platform [13]).

Such a big quantity of images had a main role in a quite large number of research in Computer science and in particular in the Machine learning field. However, before describing the usage of images in research field we must fulfill some mandatory steps: **image retrieval, processing and analysis**.

1.3.2 Image retrieval and automatic image annotation

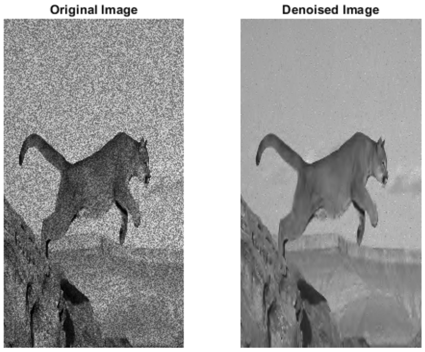
An **image retrieval system** is a digital system for browsing, searching and retrieving images from a large database of digital images. Most traditional and common methods of image retrieval utilize some techniques to **adding metadata** such as captioning, keywords, title or descriptions to the images so that retrieval can be performed over the annotation words[14].

Image annotation is a fundamental step in a machine learning model developing process and, because manual image annotation is time-consuming, laborious and expensive, there has been a large amount of research done on automatic image annotation.

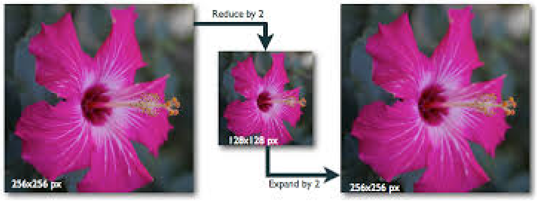
Automatic image annotation is the process by which a computer system automatically assigns metadata in the form of captioning or keywords to a digital image. The latter could be carry out through computer vision techniques which is used in image retrieval systems to organize and locate images of interest from a database [15]. This particular task can be referred as a **type of multi-class image classification** with a very large number of classes - as large as the vocabulary size. Typically, we face this task exploiting images which are already annotated (but eventually we can avoid labels) and machine learning techniques. A machine learning algorithm usually learn the correlations between images and training annotations, or at least cluster regions within them. The fact that **automatic image annotation** use machine learning models often implies that a start-up step of manual annotation is mandatory; there are cases in which manual annotation is unnecessary (unsupervised learning). However, automatic image annotation have improved performance of image retrieval system and have reduced time spent tagging image manually.

1.3.3 Image processing

Image Processing is the field of enhancing the images by tuning many parameter and features of the images [16]. Precisely, **image processing is a subset of Computer Vision** in which many transformations are applied to an input image and an the resultant output image is returned. Such kind of transformations includes denoising, up-scaling, down-scaling and cropping (those will be better described in 1.4). In Fig.3a and 3b there are some graphical examples of these kind of image transformations:



(a) Denoising



(b) Downscaling and Upscaling

1.3.4 Image Analysis

There are a lot of ways to do image analysis and all of them are typically computer-based. Image analysis is usually defined as the **extraction of meaningful information from images**, mainly from digital images by means of digital image processing techniques. The fact that this kind of analysis is computer-based allow us to make thousands of them almost instantly. Image analysis field includes a lot of simple and complex tasks (e.g., recognize a number vs recognized a person identity). Many people always interchange the term Image Analysis with Computer Vision (that will be explore in the next sections) even if Computer Vision is a sub-field of Image Analysis [17].

1.3.5 The path to analyze a big amount of images

Until now, we understood that to gain knowledge from images we have to take a specific path:

1. Generate images;

2. Save in a persistent images annotating them with metadata;

3. Image retrieval;

4. Image processing;

5. Image analysis.

However, this thesis will not deal with the first 3 steps that were considered done but will focus on Image processing and analysis. Indeed, we will see in Section 2, that we had a good number of available historical documentary photos on which we did some analysis exploiting Computer Vision and Machine learning techniques.
In order to take an overall understanding of this work we need to introduce some notions about Computer Vision and how it is linked with Artificial Intelligence.

1.4 Computer Vision & Deep Learning

Computer vision is an interdisciplinary scientific field that deals with how computers can gain high-level understanding from digital images or videos.
Computer vision tasks include methods for **acquiring, processing, analyzing and understanding digital images** and **extraction of high-dimensional data from the real world** in order to produce numerical or symbolic information (i.g. in the forms of decisions).
The term **understanding**,in this context, means the transformation of visual images (the input of the retina) into descriptions of the world that make sense to thought processes and can elicit appropriate action. This image understanding can be seen as the disentangling of symbolic information from image data using models constructed with the aid of geometry, physics, statistics, and learning theory.
Computer vision has a lot of applications that range from industrial machine vision systems to research into artificial intelligence and computers or robots that can comprehend the world around them. A list of typical computer vision applications:

- Automatic inspection;

- Assisting humans in identification tasks;

- Controlling industrial processes;

- Modeling objects or environments;

- Navigation (i.e. for a mobile robot or autonomous vehicle);

- Tactile Feedback to support visual effects;
- Classification;
- Infer Semantic Context;

Despite Computer Vision is not only made with artificial intelligence, many of these applications involve it and, nowadays, a particular branch of artificial intelligence called **Machine Learning** took the reins of this field. In the following section we will explore the notions of Machine Learning and its usage in Computer vision focusing on the subfield of Deep Learning.

1.4.1 Machine Learning

There is a forest of Machine Learning definition and, within them, one fit quite well with my personal take:
"Machine learning is a sub-field of artificial intelligence that provides systems the ability to automatically learn and improve from experience without being explicitly programmed.
Machine learning **focuses on the development of computer programs that can access data and use it to learn**. The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. The primary aim is to allow algorithms to learn automatically without human intervention or assistance and adjust actions accordingly" [18]. Machine learning algorithms could also be defined through the way they learn: **supervised, unsupervised and semi-supervised learning**.

Supervised Learning With **supervised learning**, the machine is trained using a set of labeled data, where each element is composed of given input–outcome pairs. The machine learns the relationship between the input and the outcome, and the goal is to predict behavior or make a decision based on previously given data [19]. For example, we can provide the machine as input pictures of cats and dogs with specified labels:

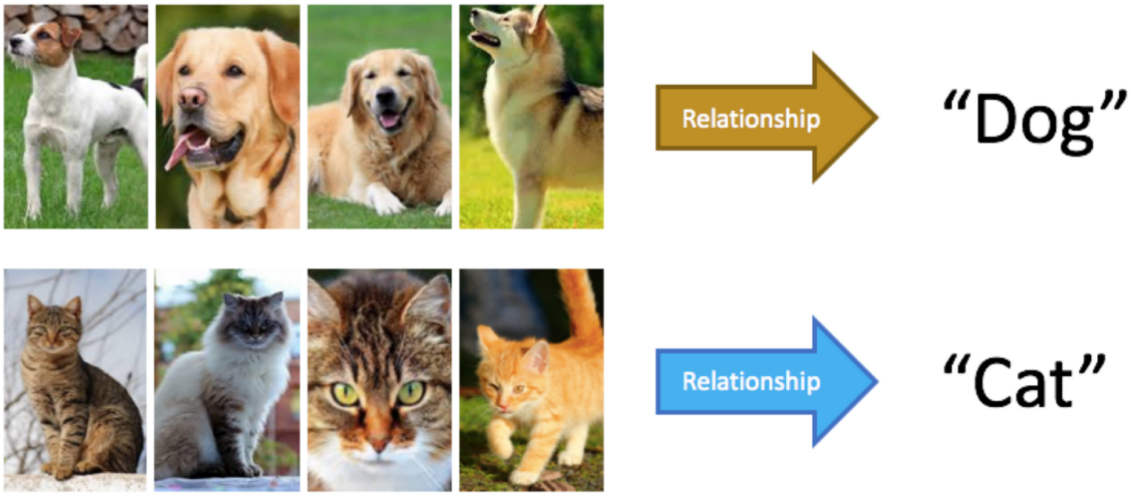


Figure 4: Example of supervised learning

There are many other examples in which we provide the machine with an input to get specific outcomes:

- Pictures of animals with their names, and then train it to identify a given animal;
- Pictures of people with the date in which the photo was taken, and then train it to guess the date given a new unseen photo;
- A number of e-mails received in your inbox, and then train it to distinguish spam messages from legitimate ones;
- Many other examples.

So the nature of Supervised Learning requires data to be labeled (often manually) which has, sometimes, a very high human-time cost.

Unsupervised Learning With unsupervised learning, **the machine is trained with unlabeled data and the goal is to group elements based on similar characteristics or features that make them unique**. These groups are often referred to as clusters. Here we are not searching for a specific, right, or even approximate single answer. Instead, the accuracy of the results is given by the similarities in the characteristics or behavior between members of the same group when compared one to another,

and the differences with the elements of another group [19].
For example, we can provide the machine as input pictures of cats and dogs and see the groups extracted by ML algorithm :

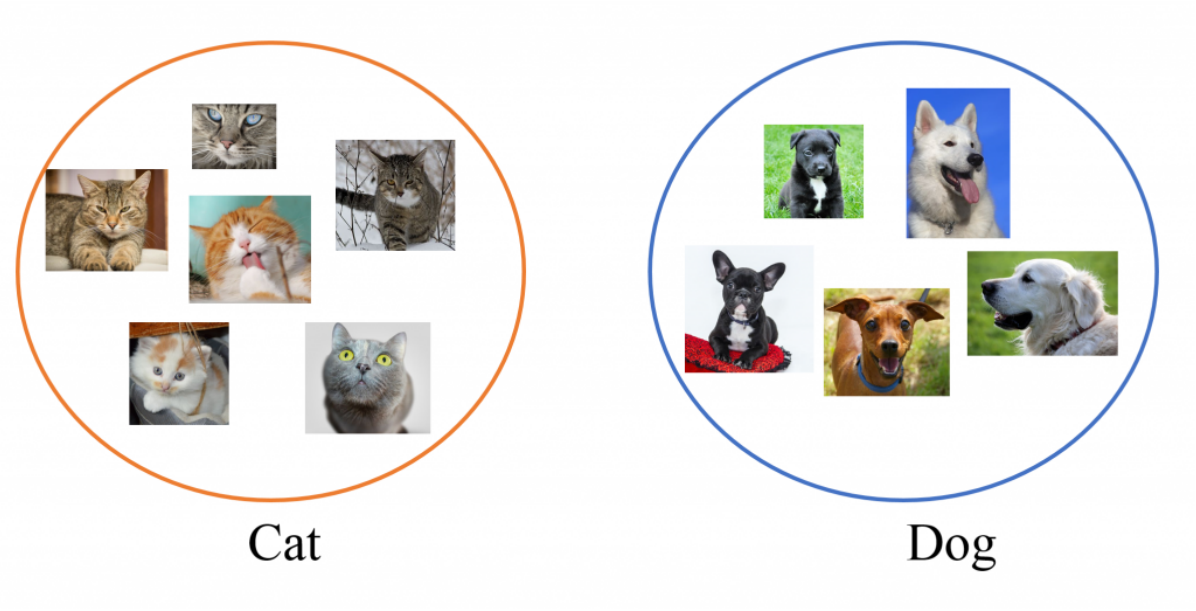


Figure 5: Example of unsupervised learning

There are many other examples in which we provide the machine with an input to get specific specific outcomes:

- A number of pictures with only one person in each, it can help you group them based on ethnicity, hair or eye color, and so on
- A list of items bought from an online store, it can help you determine the shopping habits and group them by geographical location or age
- Pictures of people. it can help you to determine similar wearing styles, presence of some particular landscape and so on.

The nature of Unsupervised Learning doesn't require data to be labeled, this means that unsupervised algorithm usually doesn't require human time. However, results from the application of an Unsupervised learning algorithm has to be interpreted and no clear indication is given about the number of clusters and what the provided data actually represents. Also, the names of the categories are not given at first, and all you can do in the very beginning is determine the boundaries between them.

Semi supervised Learning The semi-supervised learning approach tries to find an halfway between the supervised and unsupervised learning to overcome the problem of both this learning categories (data labeling and data interpretation).
As first example, imagine you want to train a model to classify text documents but you want to give your algorithm a hint about how to construct the categories. You want to use only a very small portion of labeled text documents because every document is not labeled and at the same time you want your model to classify the unlabeled documents as accurately as possible based on the documents that are already labeled [20].
Another example is a presence-face classifier: you want to train a model to classify if there is a face in the photo but you want to give algorithm only two hint labels "there is a face" and "there isn't a face". We can train the model only with few tagged examples and then check the accuracy of the model (Fig. 6).

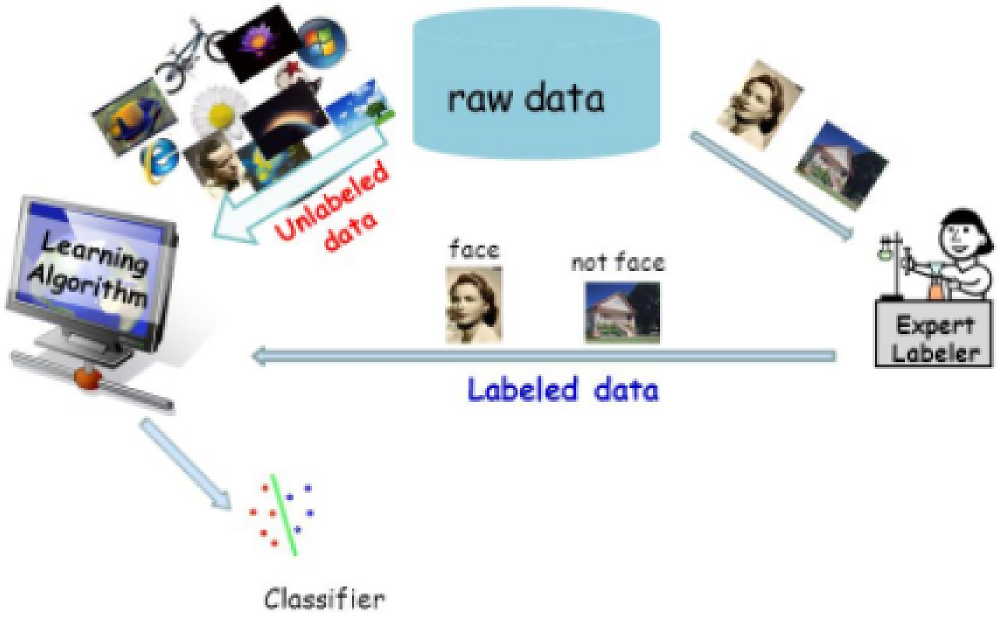


Figure 6: Face classifier Semi-supervised example

These were only simple examples, indeed semi-supervised learning algorithms have a lot of applications in complex fields [21] such as:

- Speech Analysis: Speech analysis is a classic example of the value of semi-supervised learning models. Labeling audio files typically is a very intensive tasks that requires a lot of human resources. Applying SSL techniques can really help to improve traditional speech analytic models;

- Protein Sequence Classification: Inferring the function of proteins typically requires active human intervention;

- Web Content Classification: Organizing the knowledge available in billions of web pages will advance different segments of AI. Unfortunately, that task typically requires human intervention to classify the content.

As these models are very easy to build (from a data construction point of view), more and more companies and researchers are investing time and money to develop semi-supervised learning models.
Finally, we can state this thesis work is based on the Supervised Learning approach.

1.4.2 Revolution in Machine Learning: the advent of Deep Learning

The introduction of machine learning enabled computers to tackle problems involving knowledge of the real world and make decisions that appear subjective. A simple machine learning algorithm called logistic regression can determine whether to recommend cesarean delivery and another simple learning algorithm called naive Bayes can separate legitimate e-mail from spam e-mail. However, the performance of these simple machine learning algorithms depends heavily on the representation of the data they are given. For example, when logistic regression is used to recommend cesarean delivery, the AI system does not examine the patient directly. Instead, the doctor tells the system several pieces of relevant information, such as the presence or absence of a uterine scar. Each piece of information included in the representation of the patient is known as a feature. From this example, we can infer that many artificial intelligence tasks can be solved by designing the right set of features to extract for that task, then providing these features to a simple machine learning algorithm.
For many tasks, however, it is difficult to know what features should be extracted. For example, suppose that we would like to write a program to detect cars in photographs. We know that cars have wheels, so we might like to use the presence of a wheel as a feature. Unfortunately, it is difficult to describe exactly what a wheel looks like in terms of pixel values. A wheel has a simple geometric shape, but its image may be complicated by shadows falling on the wheel, the sun glaring off the metal parts of the wheel, the fender of the car or an object in the foreground obscuring part of the wheel, and so

on. One solution to this problem is to use machine learning to discover not only the mapping from representation to output but also the representation itself. This approach is known as **representation learning**. Learned representations often result in much better performance than can be obtained with hand-designed representations. They also enable AI systems to rapidly adapt to new tasks, with minimal human intervention. A representation learning algorithm can discover a good set of features for a simple task in minutes, or for a complex task in hours to months.

When designing features or algorithms for learning features, our goal is usually to separate the factors of variation that explain the observed data. A major source of difficulty in many real-world artificial intelligence applications is that many of the factors of variation influence every single piece of data we are able to observe. The individual pixels in an image of a red car might be very close to black at night. Of course, it can be very difficult to extract such high-level, abstract features from raw data and **Deep learning** solves this central problem in representation learning by introducing representations that are expressed in terms of other, simpler representations. Deep learning enables the computer to build complex concepts out of simpler concepts [22]. An exhaustive visual example is in Fig. 7.

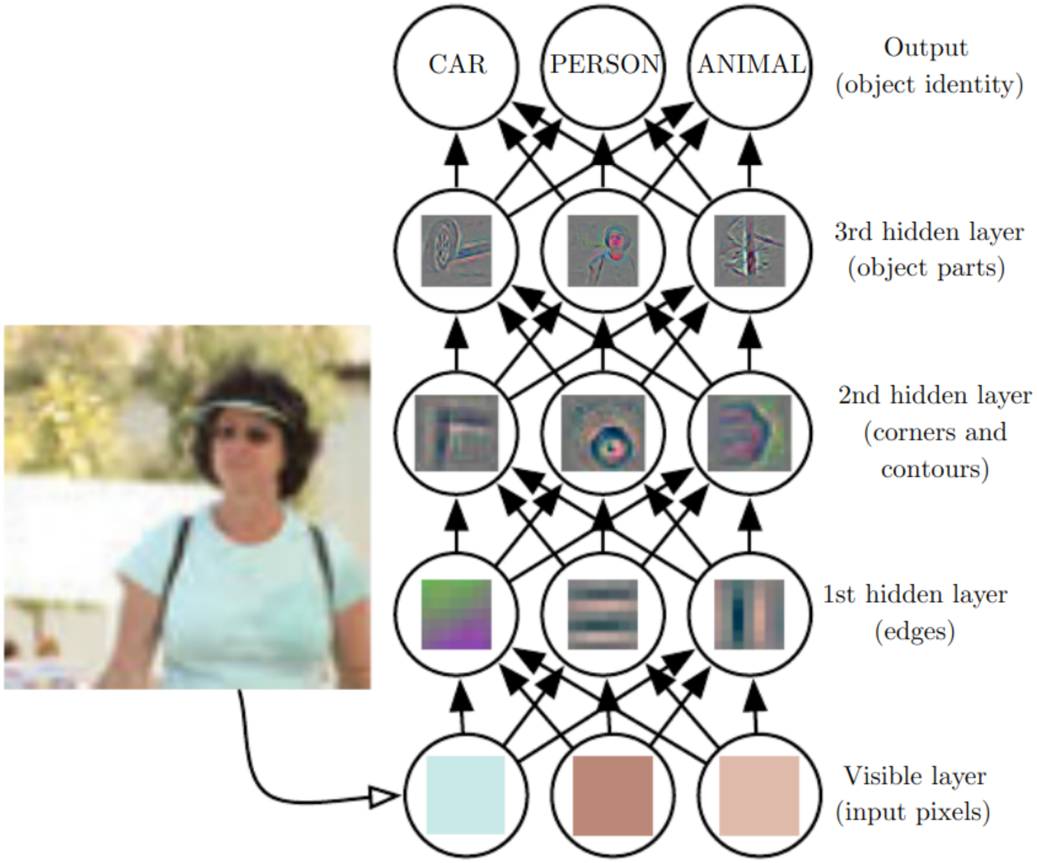


Figure 7: Illustration of how a deep learning model learn to introduce more and more complex features from simpler ones.

The main example of a deep learning model is the feed forward deep network, or multilayer perceptron (MLP). A multilayer perceptron is just a mathematical function mapping some set of input values to output values. The function is formed by composing many simpler functions. We can think of each application of a different mathematical function as providing a new representation of the input. The idea of learning the right representation for the data provides one perspective on deep learning. Another perspective on deep learning is that depth enables the computer to learn a multi-step computer program. Each layer of the representation can be thought of as the state of the computer's memory after executing another set of instructions in parallel. Networks with greater depth can execute more instructions in sequence. Sequential instructions offer great power because later instructions can refer back to the results of earlier instructions.

According to this view of deep learning, not all the information in a layer's activations necessarily encodes factors of variation that explain the input. The representation also stores state information that helps to execute a program that can make sense of the input. This step is fundamental because, in order to make predictions, the discovery of rules that underlie a phenomenon and outcoming all the considering all possible interfering factors is required. From this description, seems that deep learning is the new way to achieve some goals that couldn't be reached from classical machine learning algorithm and to reach higher performances in know tasks.

1.4.3 Deep learning modern application in real world

In recent years, there has been an explosion of deep learning applications in many fields. This fact was possible only because the Big Data era starts. Indeed Deep Learning models work well with an high quantity of data, just because the more data there is, the more patterns the model learns and the more accurate it will be. In this context, the central role of paradigms like ioT, distributed systems and products like social networks, blogs as mean to collect data is highlighted [23]. Exploiting this data, we could create technologies that were impossible until a few years ago, carrying out a certain kind of task. The types of tasks performed by a neural network are not only many but are also very varied (Fig.8).

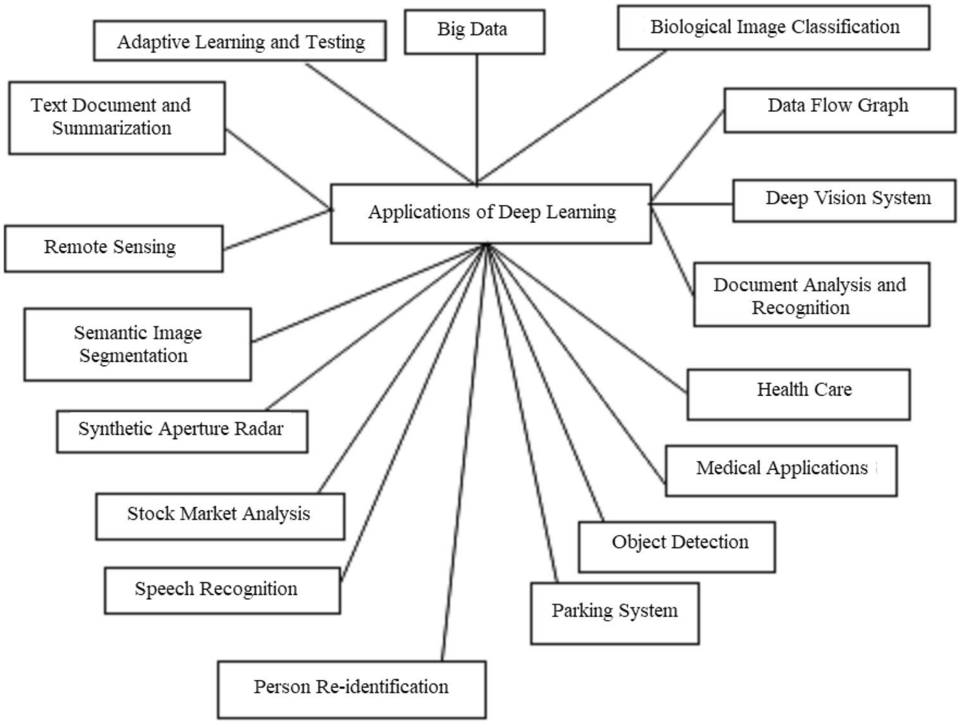


Figure 8: Applications of deep learning

The majority of these applications involves the usage and the analysis of images. This is not a case: within images there are a lot of simple features that could be combined together to create more and more complex ones in order to explain what there is the input. This is the case of **Text document and Summarization, Medical Application, Object detection, Semantic image segmentation** … Nowadays, just because neural networks seem to be a great way to analyze images to accomplish a certain task, they have become a fundamental part of modern computer vision.

1.4.4 Computer Vision & Deep Learning

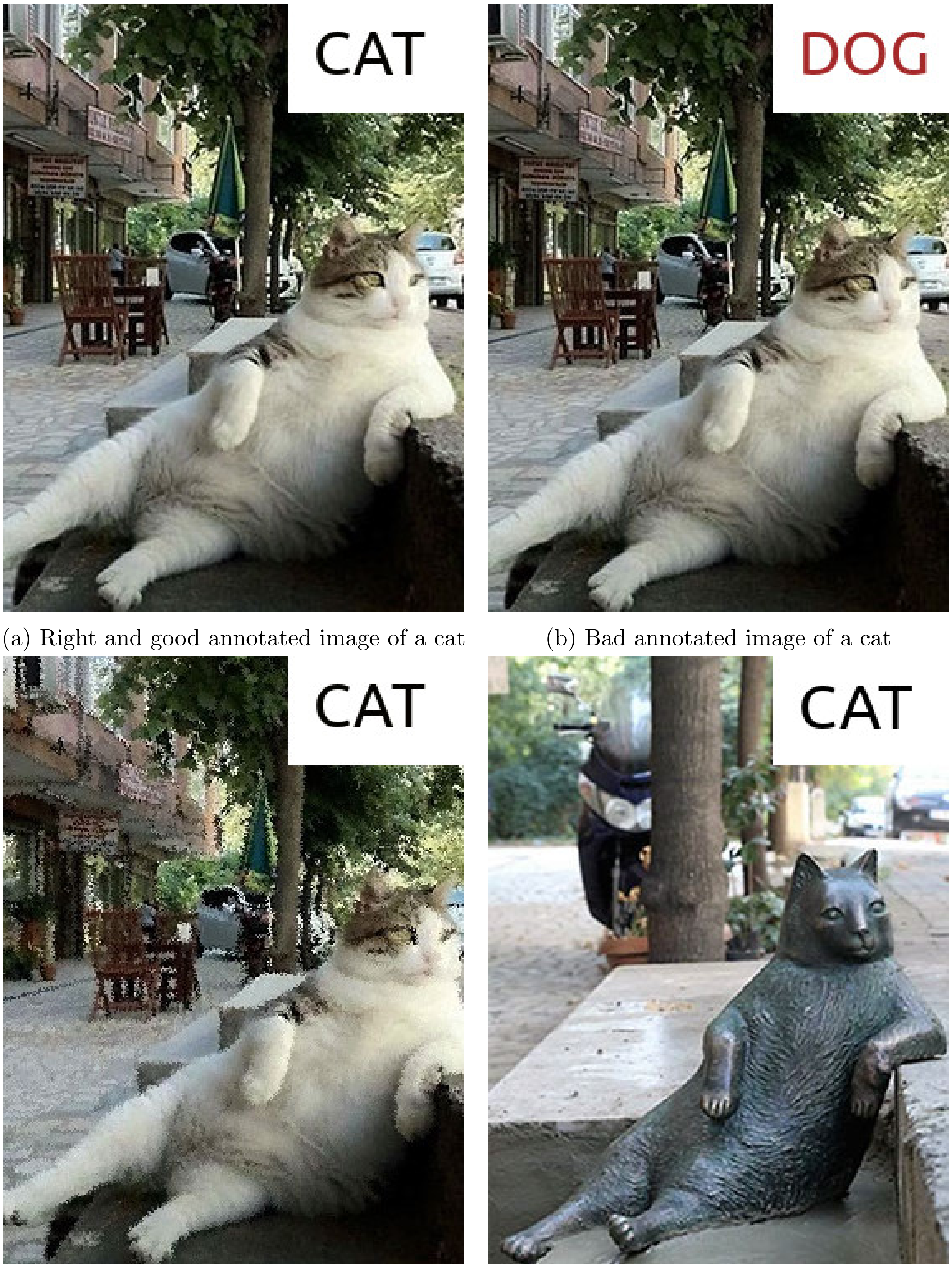
We can say that Computer Vision (CV) is not only Deep Learning (DL) and Deep Learning is not only Computer Vision....however, they are for sure in love. This seems reasonable, since CV engineers were able to achieve greater accuracy in tasks such as image classification, semantic segmentation, object detection and Simultaneous Localization and Mapping (SLAM). Moreover, since neural networks used in DL are trained rather than programmed, applications using this approach often require less expert analysis and fine-tuning and exploit the tremendous amount of video data available in today's systems. DL also provides superior flexibility because Convolutional Neural Networks (CNN) models and frameworks can be re-trained using a custom dataset for any use case, contrary to CV algorithms, which tend to be more domain-specific [24]. CNN are a particular kind of DL network that make use of kernels (also known as filters), to detect features (e.g. edges) throughout an image. A kernel is just a matrix of values, called weights, which are trained to detect specific features. This means that CNN also replace classical CV feature-extraction algorithms (e.g. edge-detection). From the previous description, seems that DL is the only technology we need to reach good results in Computer Vision, for their flexibility and goodness. However, nothing is obtained from nothing and we have to pay some costs to use deep learning algorithms:

- We need a huge amount of data, which has to be also vary varied to make the DL model learn the right pattern to carry out a specific task;
- The learning process could be very costly in terms of time and hardware resources;
- Despite DL inference can be considered fast, usually classical CV algorithms are faster.
- Even DL inference has an higher resource-consuming respect to classical CV.

Indeed, classical CV algorithms are yet used to carry out tasks like color thresholding and pixel counting. Moreover, DL models can be see as a black box: is practical impossible modify single parameters to achieve better result in a specific task. On other hand, CV algorithms are fully editable and transparent.
In conclusion, DL is becoming more and more popular in the field of CV but classical algorithms can certainly still make their contribution.

1.5 Computer Vision & Deep learning methods to augment and processing images

We already stated in Section 1.4.2 that DL models need a huge and well varied amount of data to carry out a certain task. However, depending on the task,the easiness of data acquisition could change profoundly. This problem is even stronger when designing a Deep learning model to threat images. Moreover, data acquired could be unbalanced for the task we are facing and this could lead to a failure training. For example, consider to let a CNN DL model learn by seeing 90 cats and 10 dogs. From a basically statistics point of view, this model will tend to answer that there is a cat even in the photos where there is actually a dog. Additionally, the data acquired could also be corrupted, bad annotated or containing biased examples (Fig.9).



(a) Right and good annotated image of a cat

(b) Bad annotated image of a cat

(c) Corrupted image of a cat

(d) Biased image of a cat

Figure 9: Comparing different kind of image data problems

Computer Vision couldn't correct bad annotated examples but could manage corrupted images and also generate new examples to overcome the unbalancing problem. In this section, we will review some CV algorithms and techniques (used also in this work) to manage corrupted examples and partially overcome the balancing over classes.

1.5.1 Manage corrupted images

There are many techniques that try to restore original images from corrupted ones. However, I have concentrated on those that fitted better with my study-case: **Super Resolution, Denoising and Deblurring**.

Super resolution Super-resolution (SR) [25] refers to the task of restoring high resolution images from one or more low-resolution observations of the same scene (Example in Fig. 10). According to the number of input LR images, the SR can be classified into single image super-resolution (SISR) and multi-image super-resolution (MISR). Compared with MISR, SISR is much more popular because of its high efficiency. Since an HR image with high perceptual quality has more valuable details, it is widely used in many areas, such as medical imaging, satellite imaging and security imaging.

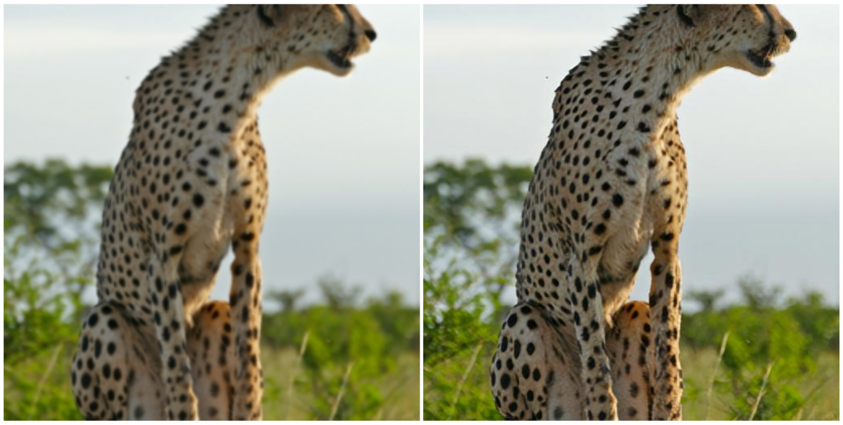


Figure 10: Super resolution example

Nowadays, a lot of Deep learning model was developed in order to achieve good results in the super resolution task. Some of theme reached very promising results [26, 27].

Denoising Image noise is a common problem that consist in a random variation of brightness or color information in images, and is usually an aspect of electronic noise.

It can be produced by the image sensor and circuitry of a scanner or digital camera. As a result, the recorded image is degraded and the recorded scene could become very confusing. The aim of image denoising is removing the noise in order to restore the original image 11. Recently, the field of noise removal has gained increasing interest in a lot of researches [28]. Both classical computer vision and deep learning techniques could reach very interesting results. However, classical computer vision often needs to know the noise model within the image in order to remove it, while deep learning could do a good job even without knowledge on the kind of noise affecting the image.

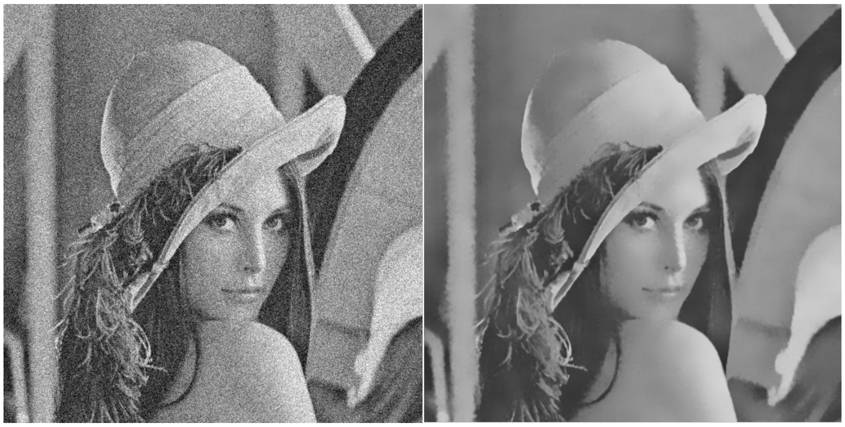


Figure 11: Denoising example

Deblurring Image blur is a common problem that occurs when recording digital images due to camera shake, long exposure time, or movement of objects. As a result, the recorded image is degraded and the recorded scene becomes unreadable. Recently, the field of blur removal has gained increasing interest in a lot of researches [29].The **deblurring problem** is known as blind deconvolution if the only available information is the blurred image and there is no knowledge about the blurring model or the Point Spread Function (PSF). In this case, the basic target of the process is to recover both the blur kernel and the deblurred (latent) image, simultaneously. An example of image deblur is in Fig.12.

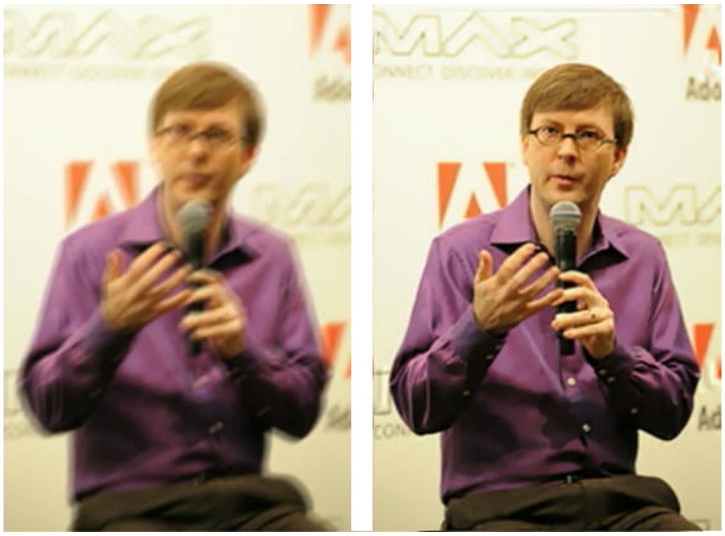
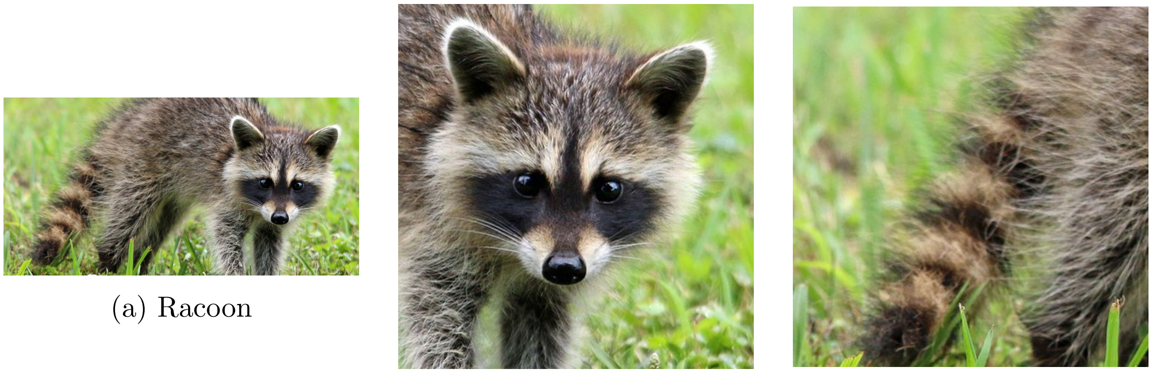


Figure 12: Deblur example

1.5.2 Generating new data

In order to overcome problem of unbalancing image datasets there are both classical and deep learning based computer vision techniques. Some of those i am going to review were used in this work.

Random Cropping Random crop is a data augmentation technique wherein we create a random subset of an original image. This helps our model generalize better because the object(s) of interest we want our models to learn are not always wholly visible in the image or the same scale in our training data. Moreover, considering only partial part of an object even if the object is always fully present in the image helps to focus on particular features (Fig. 13).



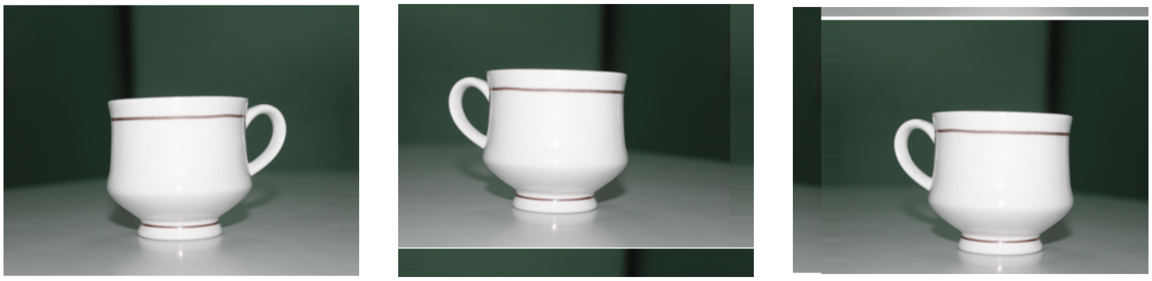
(a) Racoon

(b) Cropped racoon head

(c) Cropped racoon tail

Figure 13: Random crop example

Shifting and Translation Shifting the entire pixels of an image from one position to another position is called as shift augmentation. Normally we have two types of shift augmentation: (a) Horizontal shifting, that consist in shift all pixels of an image in horizontal direction (H) and (b) Vertical shifting, that consist in shift all pixels of an image in vertical direction (V). However, we could combine them to obtain what is called **translation** (Fig. 14).



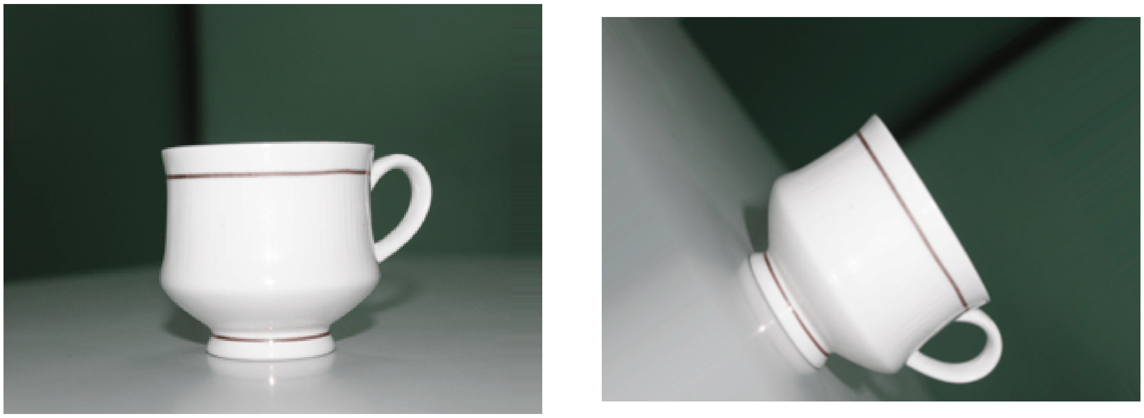
(a) Non translated image of a cup

(b) Translated image of a cup left H and upper V

(c) Translated image of a cup right H and bottom V

Figure 14: Shifting example

Rotation Rotation augmentation randomly rotate the images from 0 to 360 degrees in clock wise direction. This data augmentation technique is useful because cameras could get a rotate image of an object and this technique make the model elastic to this case (Fig. 15).

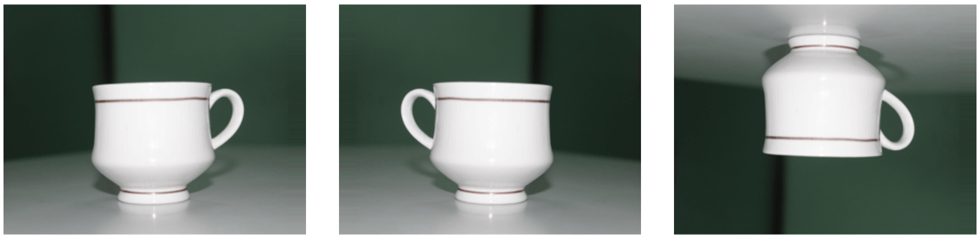


(a) Non rotated image of a cup

(b) Rotated image of a cup

Figure 15: Rotation example

Horizontal and Vertical Flipping Flipping means rotating an image in a horizontal or vertical axis. In Horizontal flip, the flipping will be on vertical axis, In Vertical flip the flipping will be on horizontal axis(Fig.16).



(a) Image of a cup without flipping

(b) Horizontal flipped image of a cup

(c) Vertical flipped image of a cup

Figure 16: Flipping example

Generative approach with DL models Deep learning models are widely used in Computer Vision to achieve tasks like recognition, segmentation, super resolution, denoising . . . however, another new task is playing an important role both in research and industrial field: the generative task. This task is carried out by models known as **Generative model**.

A generative model has a much more complex task to perform respect to the classification one. It has to understand the distribution from which the data is obtained and then use this understanding to perform the task of classification. Have learnt the distribution

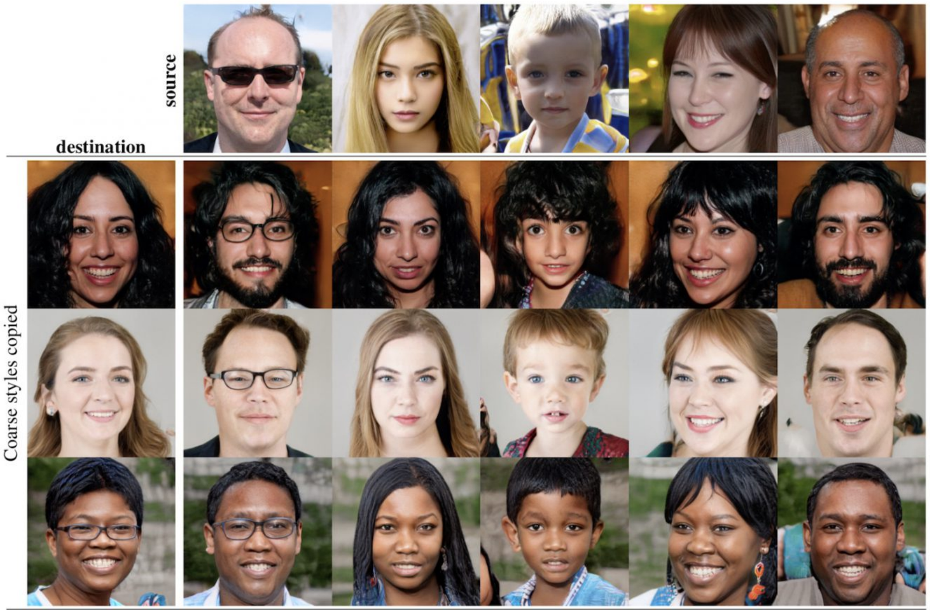


Figure 17: StyleGan generated faces sample.

they also have the capability of creating data similar to ones it received. For example. if I show a generative model a set of dog and cat images, the model should understand completely what are the features that belongs to the dog and cat class and use them to generate new images of cat and dogs. Moreover, using these features, it can compare the attributes to classify similar images as a discriminative algorithm can classify an image. The typical example of a DL generative model are GANs. An example of very powerful GAN is **StyleGAN** which is able to generate synthetic high quality face images [30]. In Fig. 17 there is a sample of faces synthetized with StyleGan.
Making these premises, it is an evidence that generative model could also be used **to generate new images to augment a dataset**, as long as he was able to understand the distribution of objects within them. This might seem like a paradox, however, this is a very used approach. For example in [31] the authors generated totally synthetic data for a binary classification problem (cancer detection). Strikingly, they showed that a decision tree classifier performed better when trained on this totally synthetic dataset than when trained on the original small dataset.

From now on, I am going to describe my thesis work, starting with the description of the IMAGO dataset and pre-processing made on it.

2 IMAGO dataset

Created and maintained at the Department for Life Quality Studies of the University of Bologna, the IMAGO dataset is a collection of **exclusive analog images** taken from familiar archives on the youthful condition in the XX century. It amounts to an important source of photos: ca 80,000 photos available, characterized by the theme "Family Album" and taken from 1845 to 2009, of which ca 23,000 were collected and labeled by master students of "Fashion Cultures and Practices" course and historians within the period [2003-2010]. More in detail, labels includes contextual information such as the shooting year, a textual description, the nationality and an unique socio-cultural context [7]. It is important to state that these images have been acquired exploiting devices for digitization (i.e. smartphones and scanners) and then loaded (exploiting a custom website) into an online database.

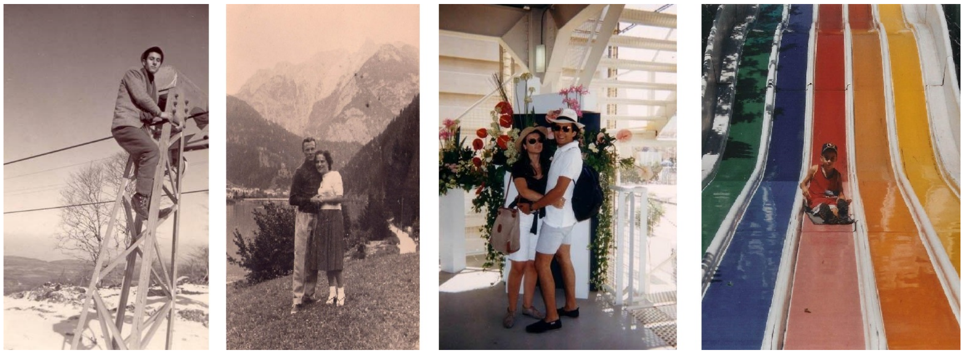


Figure 18: Sample of IMAGO analog photos among the XX century

The images actually loaded into the database are annotated with some contextual information, among which:

- **Year** in which the photo was taken;
- **Place** in which the photo was taken;
- **Macro-category of historical lecture** or **socio-cultural context**.

Years The temporal interval considered by this dataset is almost within the XX century. In particular, all the photos within were taken between 1840 and 2009.

Places Images in this dataset were taken in a lot of places (almost 70 countries),however, the greatest part of it have been snapped in Italy.

Socio-cultural context As mentioned, each picture is annotated with one or more values of socio-cultural context. This set of values were previously defined by historical researchers [32] and is defined as follows:

- **Job**: This class is the attempt to identify “who are” and “where are” the workers during the period considered. The images belonging to this class are mostly characterized by people standing in workplaces and wearing work clothes;

- **Free-time**: This group was born not only to investigate the forms and ways of experiencing the time of “no work” but to reconstruct, wherever possible, generational and gender differences, placing choices in the foreground individual in parallel to those of groups. For these reasons images within this class are very generic and may contain elements typical of other classes;

- **Holidays**: In close connection with the Free-time (this can be considered a subset of it), the investigation into the history of holidays characterize this class. It can be considered Cultural “forge”, in which people experience new aspects of their own identity, in social relationships and in the interaction with nature;

- **Motorization & Music**: These categories are closely related to the Free-time category, however, it has been kept separated because of presence of symbolical objects. Symbolical objects such as cars, motorcycles as well as musical instruments represent not only an socio-cultural context, but also the evolution of certain social customs;

- **Fashion & Customs**: Considering, in general, the history of clothing as mirror of the articulated intertwining of socio-economic, political, cultural and customs phenomena, is what orientates the cataloging in the Fashion and Costume categories. Also these classes are characterized by the presence of symbolical objects. These are mainly everyday clothes like suits, trousers, skirts, coats.

- **Affectivity & Friendship & Family**: The objective of these groups is to investigate how the affectivity and friendship relations between wives and husbands, parents and children and friends were evolved during the twentieth century. Photographs in these categories are mostly characterized by the presence of many people (e.g., friends, families, colleagues).

- **Rites & Marriages**: Through the choice of rites young people can share, accept, reject family rules and customs or affirm their own beliefs. Closely related to the Rites category to highlight the Marriage one (which can be considered a subset of it). In this group the crucial role of marriage and the profoundly changes of its meaning in the course of the twentieth century are highlighted.

- **School**: This class describes the history of the school, which is central in our society. The ruling classes identified the school as channel of education and transmission of dominant models in the process of modernization of society. Pictures in this class are often characterized by symbolical objects (e.g., school desk, blackboard) or group of students;
- **Politics**: Last but non least, the Politics category. The main aim of the latter is to focus on behaviours, body expressions and feelings of belonging to different forms of political and civil militancy. Despite the appearances, pictures within this group are closely related to the one belong to Job and School classes since the structure of the pictures is the same but the uniform is political influenced.

2.1 IMAGO dataset: structure and distributions

We have to state that, because of naming conflicts between real image filepaths and database filepaths, only 16k images over the original 23k are available for the analysis. Moreover, to manage the **unbalancing problem** some pre-processing on classes have been done.
In the following sections the Imago dataset structures and distributions will be described.

2.1.1 Imago dataset structure

All the metadata linked to the images collected in the IMAGO database are inside a single .csv file generated by me changing columns and row values via script to improve the usability of this data in the perspective of deep learning models training.
Follow a sample extracted from this .csv file:

filename	**semantic_context_classification**	**year**
a.jpg	politics	1962
rome.jpg	friendship	1984
grandma.jpg	family	1935

Table 1: Sample from .csv Imago descriptor file

2.1.2 IMAGO dataset distributions

Basing on values in the previously described .csv descriptor file, i have generated some histograms to visualize the distribution of IMAGO photos over years and categories (Fig.19).

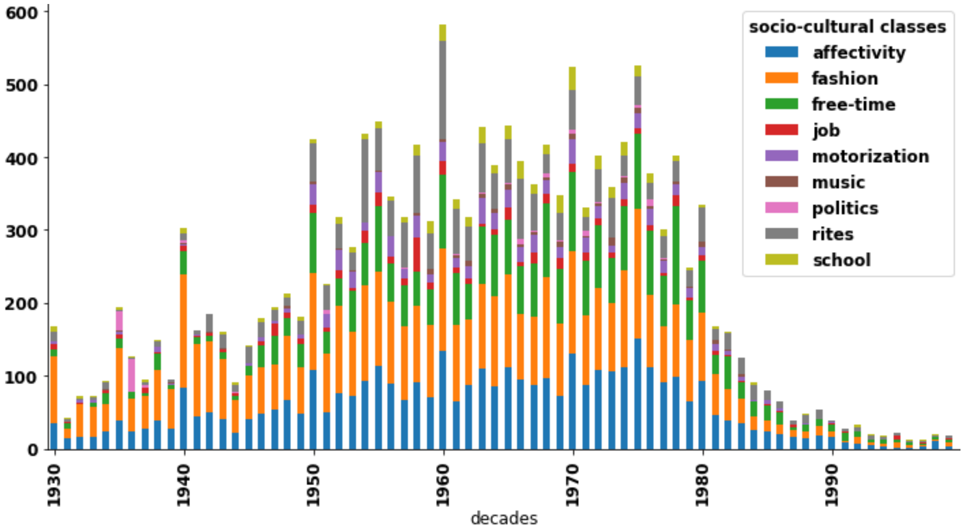


Figure 19: IMAGO distribution over years [1930-1999] and categories.

Year distribution As stated in Section 2, pictures within IMAGO cover a time-span that starts in 1840 and ends in 2009. However, to overcome the **unbalancing problem**, I have decided to consider only a subset of this time interval. In particular, I have considered photos that were taken between 1930 and 1999. The IMAGO distribution over years is reported in Fig. 19.
It is evident that, the years distribution is more concentrated in the middle decades of the XX century and, even slashing the time interval, the unbalancing is still very present. In particular, this distribution is a left-skewed normal distribution. This is reasonable, because was asked to the students, that have collected the pictures, to collect only picture with the theme "Family Album".

Semantic class distribution As stated in Section 2, pictures within IMAGO is annotated with one or more values of socio-cultural context within the set [**Politics, Job, Free-time, Holidays, Motorization, Music, Affectivity, Friendship, Fashion, Customs, Family, School, Marriages, Rites**]. From the stackbar in Fig.19, is obvious that there is a huge unbalancing between class number of examples. However this seems quite logical, because classes **affectivity, fashion and free-time** are the most general ones. Since different categories contains quite similar images, I decide to reduce the number of classes, both to reduce the difficulty of the problem from an artificial intelligence point of view and to increase the significance from historical point of view. This reduction will be described in 2.4.

2.2 A new way of interpreting IMAGO: IMAGO-FACES

In Section 3.1 I will describe a lot of works related to the dating task that exploit faces and their outlines [4, 5] to achieve quite good results. In order to create experiments comparable also with these works I have created the IMAGO-FACE dataset. To create the IMAGO-FACE dataset, was necessary an open-source implementation of YOLO-FACE [33]. However, in order to obtain more details respect to the single face image (i.e., hairstyles, earring, beard), I have enlarged the bounding box with a varying factor basing on the number of faces in the photo. This was made to extract, at least in most cases, the FACE of a single person even in a picture with 10 or 20 people. Even if some false negative were detected, the overall results were good and the generated dataset was used in the experiments. In Fig. 20 there is a sample extracted from IMAGO-FACE.

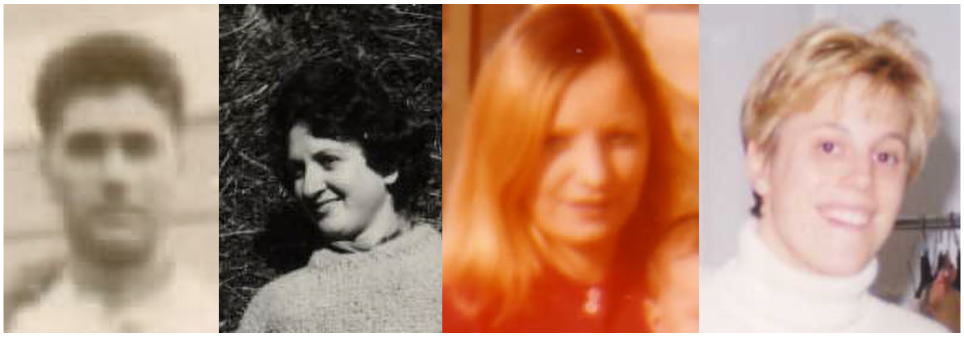


Figure 20: IMAGO-FACE samples in different decades

As this new dataset was generated from IMAGO, the unbalancing is still strongly present as you can check in Fig. 21.

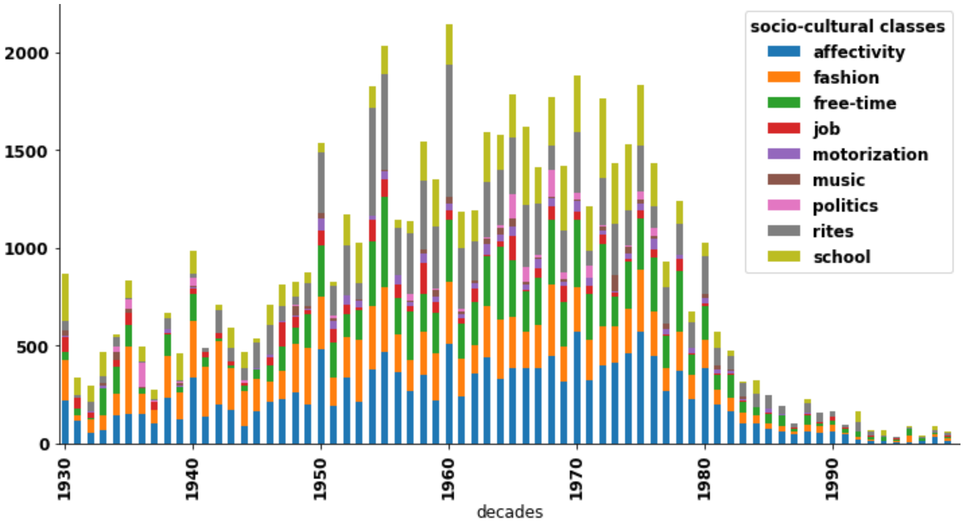


Figure 21: IMAGO-FACE distribution over years within [1930-1999] and categories.

2.3 A new way of interpreting IMAGO: IMAGO-PEOPLE

In Section 3.1 I will describe some works related to the dating task. However, none of this work have exploited the entire figure of a person. However, there are some good works in the wild that exploit people figures and, in particular, the clothes style [34]. In order to create a new kind of analysis, we have also created the IMAGO-PEOPLE dataset using an open-source implementation of YOLO [35]. As for the IMAGO-FACES dataset, in order to obtain more details respect to the single face image (i.e., hairstyles, earring, beard), I have enlarged the bounding box with a varying factor basing on the number of people in the photo. This was made to extract, at least in most cases, the FIGURE of a single person even in a picture with 10 or 20 people. Of course, people in IMAGO photos, aren't all standing, this means that we could have a variance in the aspect ratio between width and height of the regions extracted with YOLO. This could lead to some problem related to the training of deep learning models. However, using some filter on the ratio between height and width, and feeding the neural network model with images higher than wider, we could easily overcome this problem. In Fig.22 there is a sample extracted from IMAGO-PEOPLE.

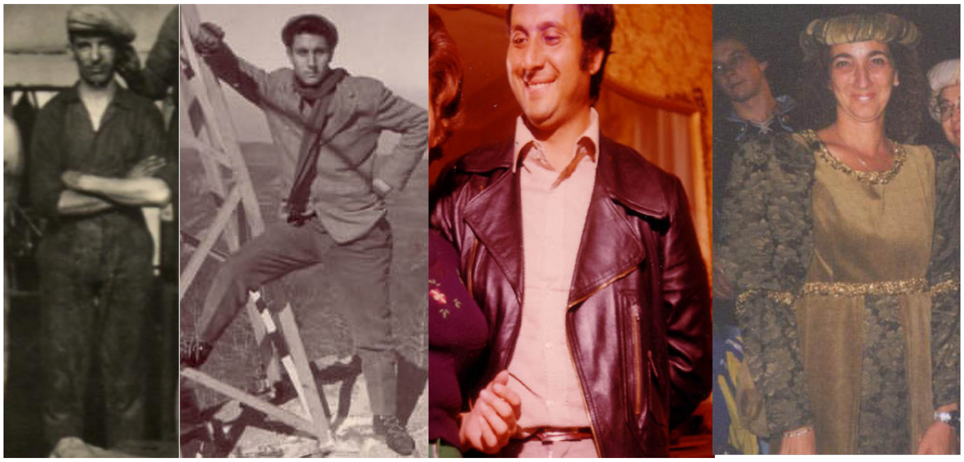


Figure 22: IMAGO-PEOPLE samples in different decades

As for the IMAGO-FACE dataset, the unbalancing is still a main phenomena (Fig.23.

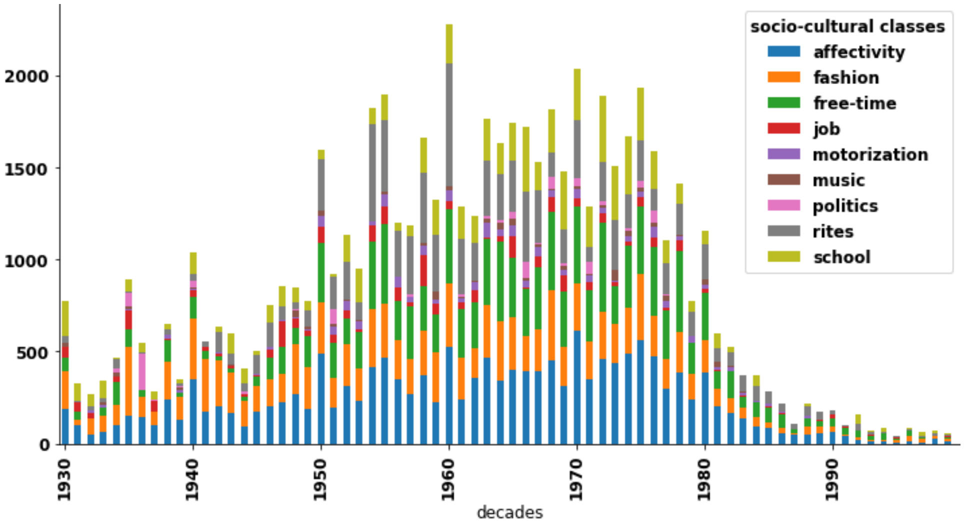


Figure 23: IMAGO-PEOPLE distribution over years [1930-1999] and categories and categories.

After creating IMAGO-FACES and IMAGO-PEOPLE, I have created a new version of the dataframe that contain all the info about an images and both its faces and people patches 2.

Fields	Record 1	Record 2
filename_original_image	a.jpg	b.jpg
dict_faces_from_original_image	{0:a_face0.jpg}	{0:b_face0.jpg, 1:b_face1.jpg}
dict_people_from_original_image	{0:a_people0.jpg}	{0:b_people0.jpg}
socio-cultural_context	politics	friendship
year	1962	1970

Table 2: Sample from IMAGO merged dataset.

2.4 Pre processing

As I stated in the introduction of this chapter, the images within IMAGO have been acquired exploiting devices for digitization (i.e. smartphones and scanners). This fact leads to add some kind of noise to images that were already noisy (because they were taken with a lot of different kind of analogical cameras) and we thought that this particular kind of noise has to be treated in some ways.
The techniques we used were contained in the **KAIR: Image Restoration Toolbox project**[36], an open source implementation of many neural network image processing based models using Pytorch as low-level framework and in the **OpenCV** library.
In particular, we have pre-process the IMAGO dataset exploiting known techniques such as:

- **Super resolution**, making use of ESRGAN [26] pre-trained model which was already contained in the KAIR toolbox. We decide to use the ESRGAN model [26], because return the best results from a details quality point of view.

- **Denoising**, in order to remove the noise that affect our images, we first try to apply the **bilateral filtering algorithm** exploit OpenCV library and then the FFDNET [37] model within the KAIR toolbox;

- **Deblurring**, making use of the Gaussian Deblurred Computer Vision algorithm within OpenCV python library. I have decide to use a simple Gaussian deblur kernel because the blur affecting some pictures within IMAGO was completely random (i.e. doesn't follow a precise model) and seems difficult to find a right kernel for each of them. Additionally, there are very few blurred images, and this could be explain by the dataset main theme: Family Album. After that, I carried out a qualitative analysis of the results directly applied to our neural network model, however the results weren't better that obtain without a deblurring action.

Each kind of pre-processing returns new processed IMAGO datasets (once for Original, faces and people patched). At this point the only way to measure which is the best kind of pre-processing, is to make a qualitative analysis. In our case, the latter consists in training different DL models (fixing train, val and test sets) and collect accuracy results. This analysis includes non-processed IMAGO datasets too. Surprisingly, the training on the non-processed IMAGO datasets returns the best results. However, we can find different reasons to explain this phenomena:

- **No processing vs Super resolution**: As stated in 1.5.1 SR, refers to the task of restoring high resolution images from one or more low-resolution observations of the same scene. This technique should have improved the quality of objects in images, mostly in those images which are very problematic (like those extracted with YOLO-FACE). However, this should also modify the main texture of image. Additionally we have to state that deep neural networks for SR are usually trained exclusively with digital images, while we are facing a problem that involves only analogical ones;

- **No processing vs Denoising**: The denoising task is defined as the search and removal of noise in order to restore the original quality of a picture. Results obtained on the qualitativ analysis exploiting ESRGAN was similar to the one obtained without any kind of denoising. Bilateral filtering algorithm return slightly better results when evaluating our model. This is strange since the noise in IMAGO pictures doesn't follow a precise model (because of the collection means) and DL model should have return images with an higher quality. However, DL models was trained on digital images and this could be also a motivation that explain this strange event;

- **No processing vs Deblurring**: Image deblur tries to overcome photos that loose quality because of shaking in cameras or movements of objects. The qualitative analysis made on the deblurred images however, doesn't return good results. THis could be explained by the fact that there are very few blurred images within this dataset, and probably the used techniques have made some features disappeared (e.g., texture, edges).

Even if **bilateral filtering** algorithm have created images that makes the learning easier for the network, we decide to avoid any kind of pre-processing, because this improvement was negligible.

Another kind of pre-processing on the dataset concern the merger between some socio-cultural classes. As stated in the introduction of Section 2 there are 14 socio-cultural classes. However, some of them, are very similar to others from a conceptual and technical point of view (e.g., rites & marriages, affectivity &friendship & family, fashion & customs, free-time & holidays). Basing on this assumption, we decide to merge some classes in a unique classes, that is, label each photo from the classes that has to be merged with a unique label. In particular, we convert classes labels with the schema described in Fig. 3. Of course label that wasn't in the latter, have been left as they were.

	Rites	→	Marriage	
	Customs	→	Fashion	
	Holidays	→	Free-time	
Friendship	→	Affectivity	←	Family

Table 3: Merge schema for socio-cultural classes.

3 The dating task

The dating task could be described as a **multi-classification problem** in which the model, given a photo, has to guess the date (usually the year) in which it was taken.

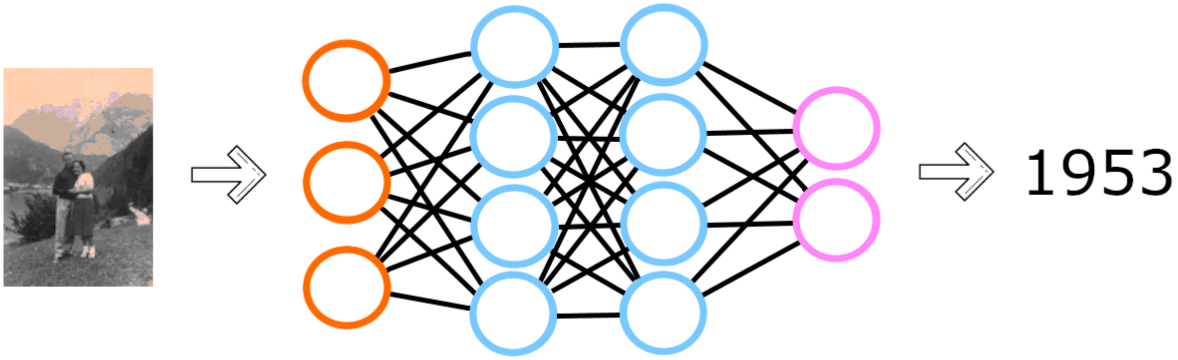


Figure 24: Example of dating

There are few but very interesting papers about dating in which the authors have faced this problem from different points of view and with different types of photos.

3.1 Related Works

In this section I am going to explore different papers related to the dating task in computer vision field.
In [3] authors exploit characteristics of the device used to acquire photographs that are considered as discriminative features for this task. They attempt to extract such characteristics from historical color photographs. They found that the acquisition device mainly effects two properties of the colors: the distribution of derivatives and the angles drawn by three consecutive pixels in the RGB space. Moreover they shown that these two color descriptors (namely color derivatives and color angles) are fundamental to obtain the state-of-the-art in the context of image dating. However, this means that this work is device dependent and this could be not good because a device built in a certain decade could be used to take some photos in another decade.
Another cool work is [5] in which authors exploit a collection of one type of widely available yet little used historical visual data (a century's worth of United States high school yearbook portraits) trough data mining techniques in order to discover the evolution in the appearance of people over time. Although the article is interesting from many points

of view, the most interesting part is that regards the analysis of the mode of the fashion aspects over decades (Figure 25).

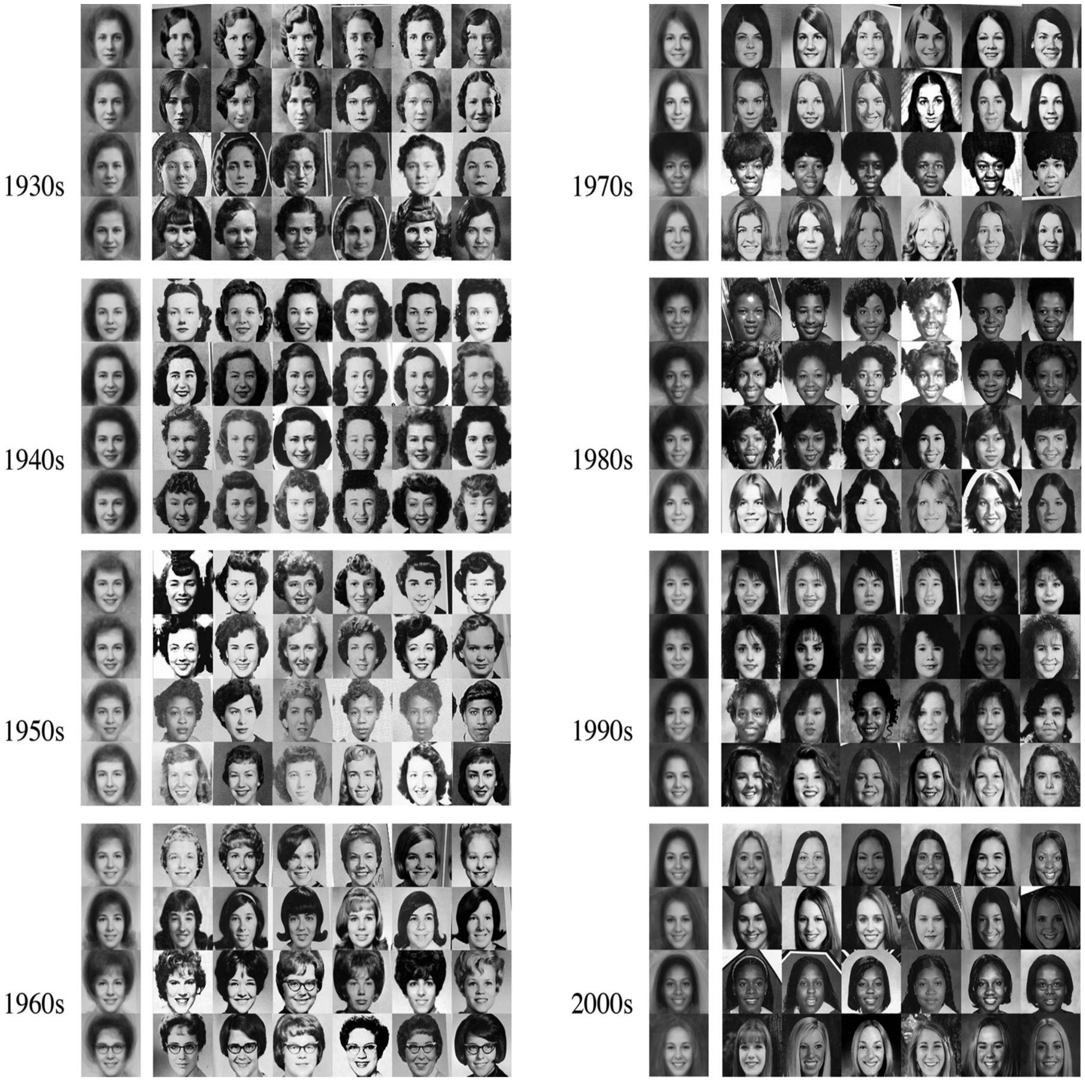
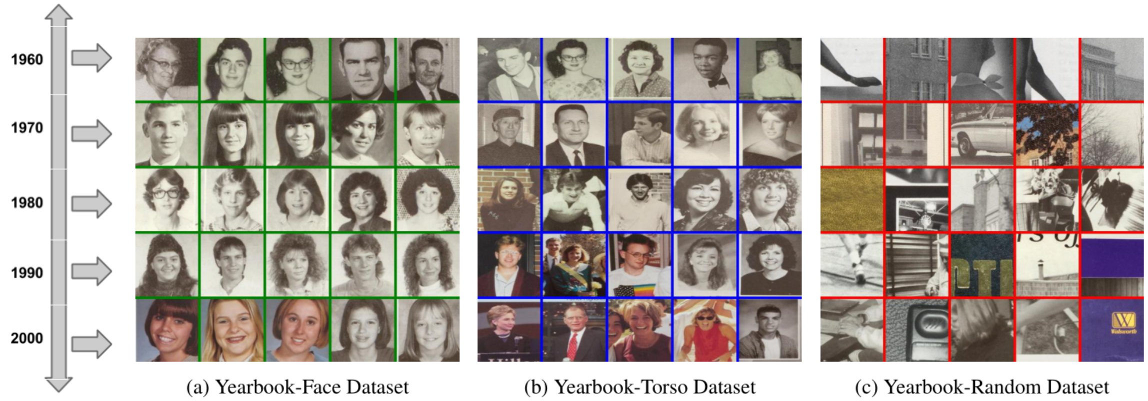


Figure 25: Means of facial female aspects between decades made in [5]

Another interesting part of the article is that related to the CNN architecture to predict the year in which an yearbook photo was taken basing on the style between decades. The results were good (almost 55 % of accuracy with a median error of 4 years) but the model was not so elastic and it takes into account only facial fashion aspect of

people.
The last paper i am going to describe is [4] by researchers from Kentucky. The authors have collected a new dataset containing images of people from 7 high school yearbooks, covering the years 1912–2014. They haven't found a complete solution to the problem of image dating,but their results show that human appearance is strongly related to time and that semantic information can be a useful cue. Moreover they proposed an approach, based on deep convolutional neural networks, to estimate when an image was captured directly from raw pixel intensities, and provided a detailed evaluation, both quantitative and qualitative, of the learned models for a variety of different settings. In particular they considered:

- Color and grayscale images;
- Face and Torso of people and Random pieces of image;



(a) Yearbook-Face Dataset (b) Yearbook-Torso Dataset (c) Yearbook-Random Dataset

Figure 26: Sample from different dataset settings in [4]

Thanks to the fact that their dataset was very large well-distributed over years, they were able to reach a very high accuracy in decades predicting using the **Yearboor-Torso Dataset** (almost 75%) in a temporal slash that start in 1950 and end in 2014. Moreover, we have to consider that, images comes from a low number of school and that, photos from certain years come only from 2 or 3 schools. This could be a great bias that could rise up accuracy! However, this paper is the one that describe a way to analyze documentary photos similar to our approach even if they have used both analogical and digital images and they take into account only faces fashion aspects and partially body styles aspects.

3.2 My approach

As described in Section 2 we have three kind of visual information available spanned over the twentieth century: original documentary photos, faces and people patches. All these three information could be used to design three different kind of DL models that consider respectively original images, faces and people patched. This decision was not easy to take because we can also design a multi-input architecture . . . but I have done this too! So, in order to carry out the dating task i have created and training 4 different kinds of DL modes:

- Single Image classifier: that takes into account the original picture;
- Single face classifier: that takes into account faces patches from the original picture;
- Single people classifier: that takes into account people patches from the original picture;
- Merged classifier: that takes into account all three kind of images.

For each model i will describe the architecture, training settings and empirical results in 3.5.

3.3 Creation of train, validation and test set

In order to train a DL model and to test its capabilities we must divide the dataset into three sets:

- The training set, which is used to train the model;
- The validation set, which is used to measure performances and compare it with other models performances;
- The test set to measure qualitative performances such as accuracy, precision, recall and F1 score.

Of course this sets must have a null intersection. This means that we have to share train examples from validation and test ones. And this must be done for the entire IMAGO dataset (that includes faces, people and original images). This was done choosing randomly 80% of records for the train and validation set and 20% for the test set **for each class**. So this means that the train set for the dating task is composed by 70% of pictures within 1934, 1956 . . . .

3.4 Pre-processing and data augmentation

As in many works of Computer vision in the field of deep learning, some techniques of data augmentation and pre-processing were applied to our data in order to reduce overfitting risks, makes the learning more stable and to increase the model performances. As described in Section 2 we decide to avoid any kind of classical image pre-processing, because of our datasets nature. However, we can for sure perform some kind of data augmentation. Despite there are a lot of techniques that we can apply, we decide to use only three data augmentations algorithms:

- Random Cropping;
- Shifting;
- Horizontal flipping;

This three were chosen because they are the only techniques that produce significant images from original ones. Indeed, random cropping and shifting are useful because they allow the model to concentrate on different kind of details within the photo while the horizontal flipping allow to see same objects but in different locations, and this could be a form or feature extraction regularization. As we know regularization adds prior knowledge to a model and this explain why data augmentations techniques are also a form of regularization. On the other hand, vertical flipping or image rotation, even if them are data augmentation techniques, could create some bias in the network (e.g., consider to apply these techniques on faces and people patches).

3.5 Experiments

In this section, i am going to review the experiments made for each different models described in Section 3.2. However, is necessary to first introduce some terms that i will use in the next section and in Section 4.6.

Fine tuning Fine-tuning, in general, means making small adjustments to a process to achieve the desired output or performance. Fine-tuning, in deep learning, involves using weights of a previous deep learning algorithm for programming another similar deep learning process. Weights are used to connect each neuron in one layer to every neuron in the next layer in the neural network. The fine-tuning process significantly decreases the time required for programming and processing a new deep learning algorithm as it already contains vital information from a pre-existing deep learning algorithm;

Loss function Loss functions are used to determine the error between the output of our algorithms and the given target value. In layman's terms, the loss function expresses how far off the mark our computed output is;

Optimizer An optimizer update the weight parameters to minimize the loss function. Loss function acts as guides to the terrain telling optimizer if it is moving in the right direction to reach the bottom of the valley, the global minimum;

Learning rate (LR) The learning rate is a tuning parameter in an optimization algorithm that determines the step size at each iteration while moving toward a minimum of a loss function. Since it influences to what extent newly acquired information overrides old information, it metaphorically represents the speed at which a machine learning model "learns".

3.5.1 Single image classifier

Architecture The single Image classifier is a DL model that takes in input an image from IMAGO dataset and try to guess the date in which this was taken (Fig. 27).

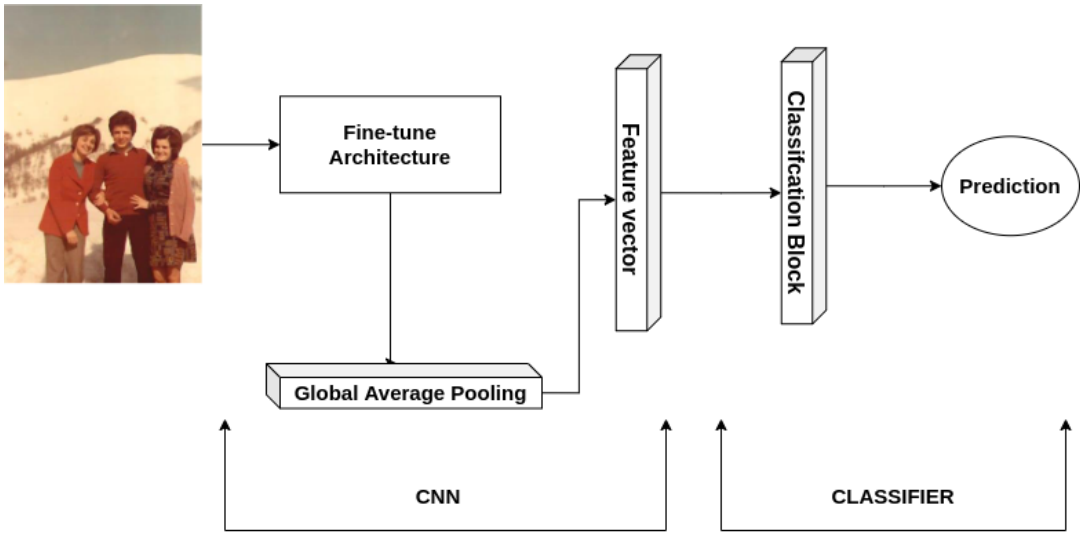


Figure 27: Single image classifier architecture.

Results As you can see, in Fig. 27 there is a token named as "Fine-tune Architecture". This is why we have fine-tuned different known deep learning architecture, exploiting

IMAGENET weights, to create what is usually call **baselines**. In Fig.4 I report all the results obtaining with different networks and different hyper-parameter configurations.

architecture	**resnet50**	**inception**	**inception_resnet**	**densenet121**
num_classes	70	70	70	70
fine tuning	partial	total	total	total
lr	0.0001	0.0001	0.0001	0.0001
epochs	40	40	40	40
optimizer	adam	adam	adam	adam
weight decay	0.0005	0.0005	0.0005	0.0005
test accuracy	11,31	10.48	9,94	10,68
dv_std error	6,76	6,89	6,89	6,74
mean error	6,01	6,2	6,33	6,13
median error	4,45	4,45	4,45	4,45

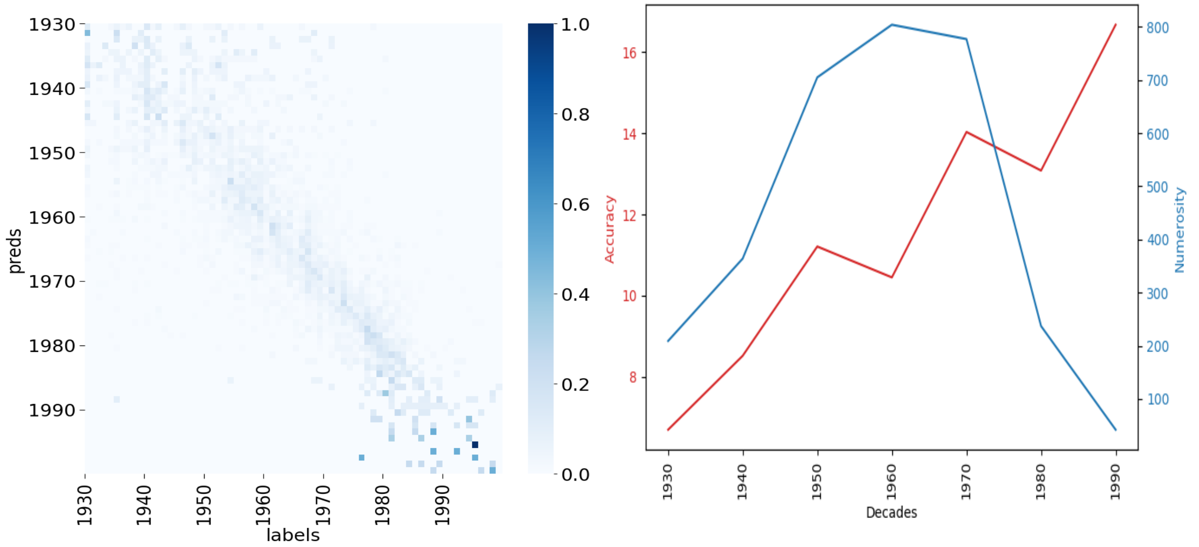
Table 4: Results for single image classifier on different DL architectures.

As we can see from these results, it seems that the accuracy changes slightly depending on the architecture and how these models were trained. However, we haven't mentioned yet the concept of **temporal window accuracy**. This metrics is used in the dating task because, in a context of single-year classification model, could be useful evaluate the accuracy of a model not only for a single year, but even for their neighbours. This fact is even more important treating documentary photos, that contains picture that characterize not only an year, but an entire decade or even an entire 20-year temporal interval. For this reason, we evaluate these models considering an increasing temporal windows that start from 0 and reach 20 with a step of ± 5 year. To do a practical example: consider a picture classifier as 1978 or 1968 while the real year is 1973; with a temporal windows of 5 or more, this classification would be considered correct!

architecture	w=0	w=5	w=10	w=20
resnet50	11,31	62,56	82,54	95,47
inception	10.48	61,38	82,82	94,74
inception_resnet	9,94	59,85	81,71	95
densenet121	10,68	60,77	82,47	95,63

Table 5: Total image baselines accuracy varying temporal window.

As we can see in 5, the accuracy increases proportionally to the increase of the variable W. From these results we can infer that, the model has understand more than is excepted, by watching only the top-1 accuracy. This phenomena is more evident, if we analyze the confusion matrix of the best model, which is considered the Resnet50 (Fig. 28a). The diagonal structure demonstrates that confusion mostly occurs between neighboring years, except for the first 20 and the last 15, indicating that the dating model can distinguish between time periods. This is a quite interesting phenomena happened in [5] too. To emphasize this fact, we can check the accuracy aggregated by decades (Fig. 28b). Comments on these plots will be discussed in 3.5.4, because they have common feature also with results obtained by those models that consider faces and people patches.



(a) Confusion matrix single image classifier.

(b) Single Image classifier accuracy by decades compared to numerosity.

In our case, this fact could be explained in two different ways:

- For what concern the confusion created within the first 20 years: From a computer science point of view, this phenomena could be addressed to the fact that the quality of the images is very low and the samples are less than other years (check Fig. 19). From a socio-cultural point of view, this great confusion could be explained by the fact that styles and situations was very similar (doing photos was not that frequent and was done only on particular occasions);
- For what concern the confusion created within the last 15 years: in our case, this phenomena could be addressed to the fact that the number of sample was very low in each of those year (check Fig. 19).

3.5.2 Single face classifier

Architecture The single Image classifier is a DL model that takes in input an image from IMAGO-FACE dataset and try to guess the date in which this was taken (Fig. 29).

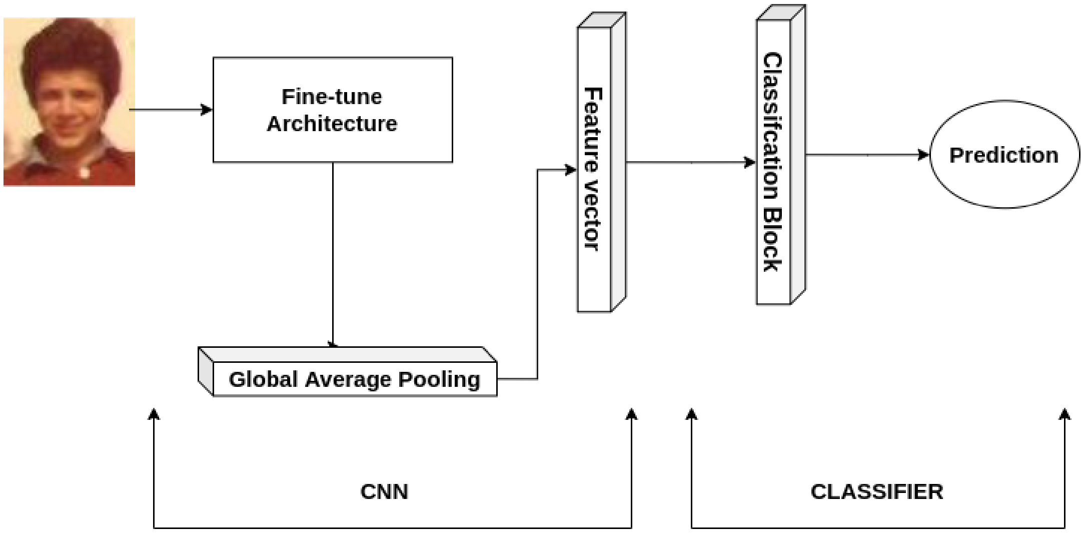
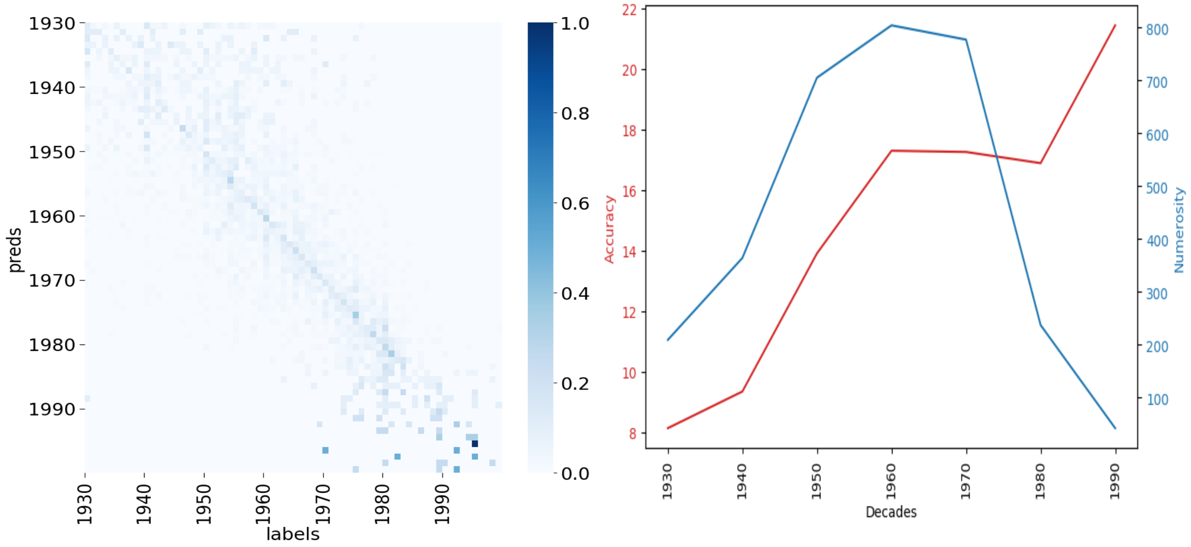


Figure 29: Single face classifier architecture.

As for the previous classifier I report results obtained on different architecture and varying the temporal window (Tab. 6, 7). We have to state that the accuracy is calculated firstly taking probability score for each face in a picture, combining them calculating the mean and then extracting the class that has the greatest probability to belong to a class (using the argmax function).

In order to explore the confusion made between classes, I have also plotted the confusion matrix of the best classifier (which is considered Resnet50) in Fig. 30a. Justifications for results in Fig. 30a are those discussed in 3.5.4.



(a) Confusion matrix single face classifier.

(b) Single face classifier accuracy by decades compared to numerosity.

architecture	resnet50	inception	inception_resnet	densenet121
num_classes	70	70	70	70
fine tuning	partial	total	total	total
lr	0.0001	0.0001	0.0001	0.0001
epochs	40	40	40	40
optimizer	adam	adam	adam	adam
weight decay	0.0005	0.0005	0.0005	0.0005
test accuracy	15,01	12,56	13,7	12,91
mean error	6.56	7.08	6.48	6,92
dv_std error	7.13	7.2	7.06	6.54
median error	4.45	4.45	4.45	4.45

Table 6: Results for single face classifier on different DL architectures.

architecture	w=0	w=5	w=10	w=20
resnet50	15,01	58,09	78,39	94,26
inception	12,56	56,95	78,46	94,33
inception_resnet	13,7	58,25	79,19	94,84
densenet121	12,91	57,81	79,7	94,87

Table 7: Results with varying temporal window.

However, these results are not the only one that have to be discussed. Indeed, a picture could contain more that one person, this means that the accuracy could be dependent on the number of faces within a photo. However, this concept has to be proved through random trials in a fixed subset of the test set. In short words, we have to extract pictures with n faces (where n is at least 2) and random sampling j faces (where j starts from 1 and goes to n), evaluating them and save the accuracy for m number of times. Greater m is, greater would be the confidence level that guarantee that goodness of the prove. I chose m=1000 and n=8 (to guarantee at least a sub-test set of 150 pictures). In this way, we are able to calculate a measure which is strictly related to **confidence interval**. Results for different baselines are listed in Fig.31.

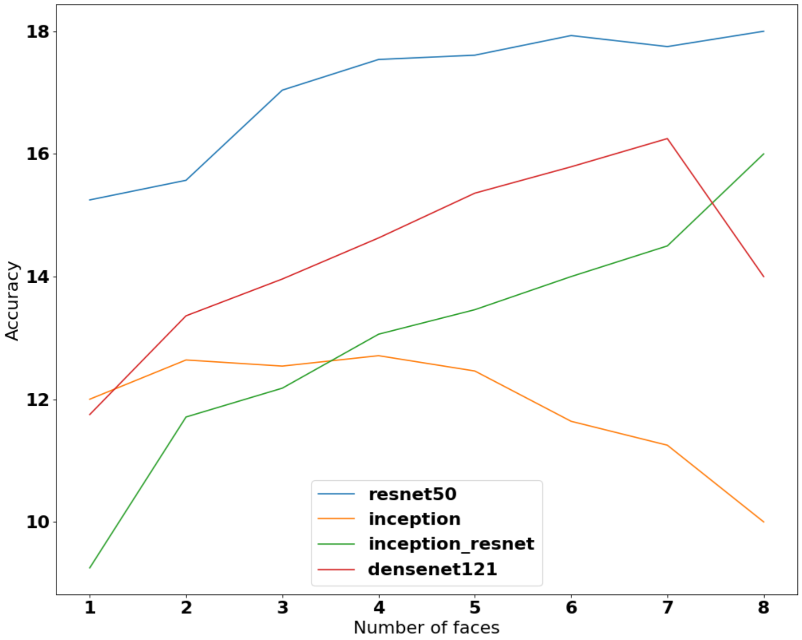


Figure 31: Accuracy varying number of faces in baselines.

3.5.3 Single person classifier

The single Image classifier is a DL model that takes in input an image from IMAGO-PEOPLE dataset and try to guess the date in which this was taken (Fig. 32).

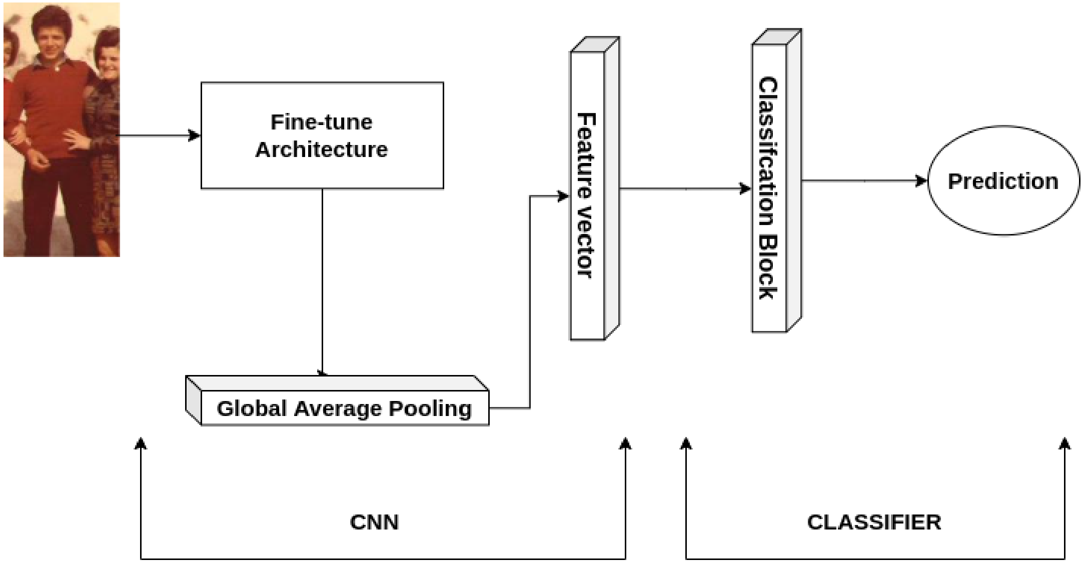
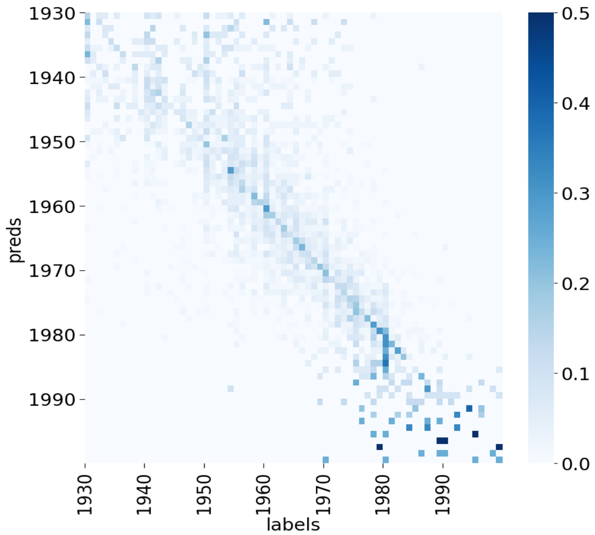
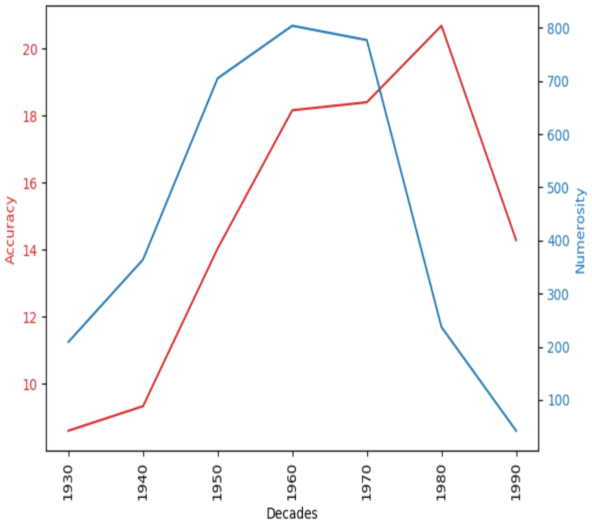


Figure 32: Single person classifier architecture.

As for the previous classifier I report results obtained on different architecture and varying the temporal window (Tab. 8, 9). As for the single face classifier, the accuracy is calculated firstly taking probability score for each person in a picture, combining them calculating the mean and then extracting the class that has the greatest probability to belong to a class (using the argmax function).
In order to explore the confusion made between classes, I have also plotted the confusion matrix of the best classifier (which is considered Resnet50) in Fig. 33a. Justifications for results in Fig. 33a are those discussed in 3.5.4.



(a) Confusion matrix single person classifier.



(b) Single person classifier accuracy by decades compared to numerosity.

architecture	resnet50	inception	inception_resnet	densenet121
num_classes	70	70	70	70
fine tuning	partial	total	total	total
lr	0.0001	0.0001	0.0001	0.0001
epochs	40	40	40	40
optimizer	adam	adam	adam	adam
weight decay	0.0005	0.0005	0.0005	0.0005
test accuracy	15,77	14,6	13,99	13,99
mean error	5,98	6,2	6,23	6,23
dv_std error	6,82	6,75	6,76	6,82
median error	4,45	4,45	4,45	4,45

Table 8: Results for single person classifier on different DL architectures.

architecture	w=0	w=5	w=10	w=20
resnet50	15,77	62,4	82,47	95
inception	14,6	60,04	81,39	95,09
inception_resnet	13,99	60,77	80,08	95,38
densenet121	13,99	59,69	81,42	95

Table 9: Results with varying temporal windows.

As for the previous classifier, these results are not the only one that have to be discussed. Indeed, a picture could contain more that one person, this means that the accuracy could be dependent on the number of person patches within a photo. However, this concept has to be proved through random trials in a subset of the test set. Results for different baselines are listed in Fig.34.

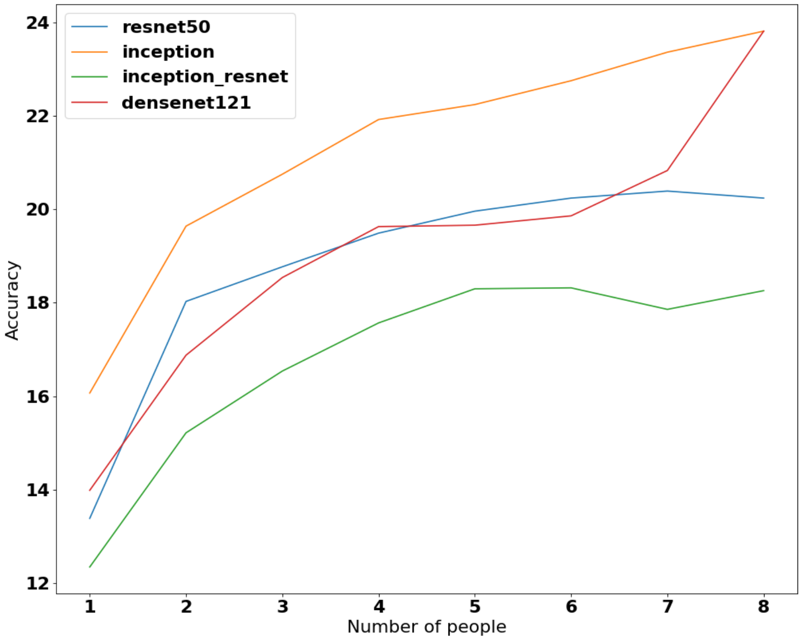


Figure 34: Accuracy varying number of people in baselines.

3.5.4 Comments on previous results

The diagonal structures in Fig. 28a, 30a and 33a, demonstrates that confusion mostly occurs between neighboring years, except for the first 20 and the last 15, indicating that the dating model can distinguish between time periods. This is a quite interesting phenomena happened in [5] too. So, even if the accuracy by year is low in all three classifier (which are, however, similar to works in 3.1, the models seems to understand which is the era we are referring to. Moreover, we can make some reflections on how the confusion is pretty higher in certain decades than others:

- For what concern the confusion created within the first 20 years: From a computer science point of view, this phenomena could be addressed to the fact that the quality of the images is very low and the samples are less than other years (check Fig. 19). From a socio-cultural point of view, this great confusion could be explained by the fact that styles and situations was very similar (doing photos was not that frequent and was done only on particular occasions);

- For what concern the confusion created within the last 15 years: in our case, this phenomena could be addressed to the fact that the number of sample was very low in each of those year (check Fig. 19).

3.5.5 An ensemble architecture

As shown in previous sections, we have reached good results building classifiers to achieve dating task exploiting different IMAGO datasets. However, I wondered if these results could be improved . . . and the answer was yes! Indeed, information within images from different IMAGO datasets have for sure a **complementarity** in identifying the date to which the picture belong. In order to exploit this complementarity I have created an ensemble architecture, that trying to guess the date taking into account: (a) the original picture (b) all faces within the original picture (c) all people within the original picture (Fig. 35).

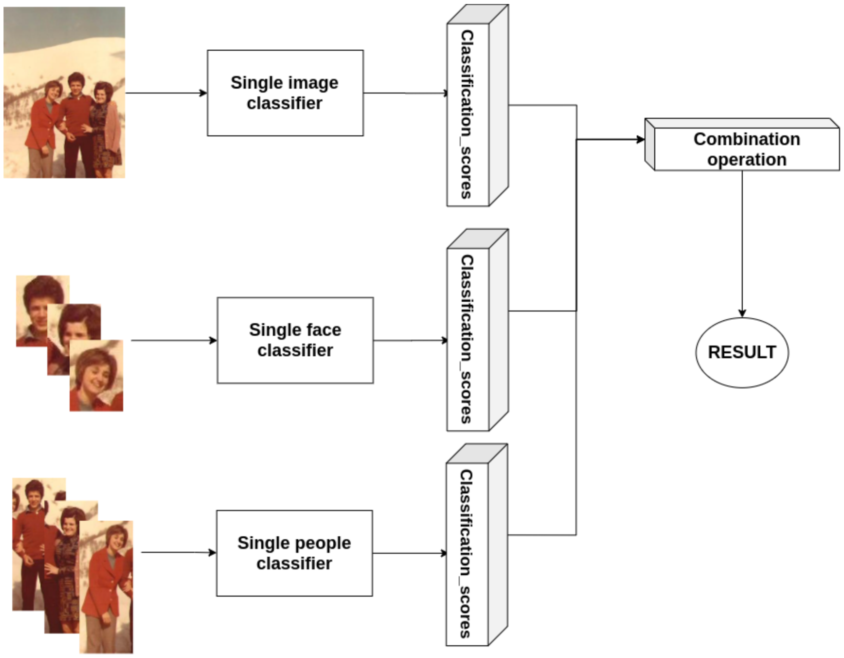


Figure 35: Ensemble classifier architecture.

As shown in Fig. 35, for each picture, I take the classification score (which are those values that indicates the probability to belong to a certain class) and combine them with some kind of **combination operation**. I have tried many combination operations, however, seems that the classical sum works better. Results for different baselines are listed in 10.

architecture	resnet50	inception	inception_resnet	densenet121
combination method	sum	sum	sum	sum
test accuracy	17,78	15,77	15,87	15,68
mean error	5,63	5,16	5,2	5,86
dv_std error	6,51	5,95	5,93	6,68
median error	4,45	4,45	2,97	4,45

Table 10: Results for ensemble classifier on different DL architectures.

The fact that the sum is the best combination operation, could be hard to accept, because of the fact that the scores taken from the **original picture model** could be less important in pictures that contains many people. Indeed, the role of the context around people within a photo could be useful to identify the age in which the photo was taken (i.e. a particular kind of car, furnishings, buildings), but we have to understand in which situations this is verified. We can prove that the importance of the context is decreasing proportionally to the increase of faces and people, by making the same **random trials** method used in previous sections, but considering both people and faces patches and inserting/removing the probability score returned by the **total image classifier**. In mathematical term we calculate, for the same test set:

- Accuracy considering combinations of increasing n faces and people , combining them with the mean;
- Accuracy considering combinations of increasing n faces and people and the fixed context value, combining them with the mean;
- Calculate the difference vector between them.

Results for different baselines are plotted in Fig. 37 and 36. In this plots, is evident that the importance of the context is generally decreasing and accuracy is increasing considering more and more faces and people.

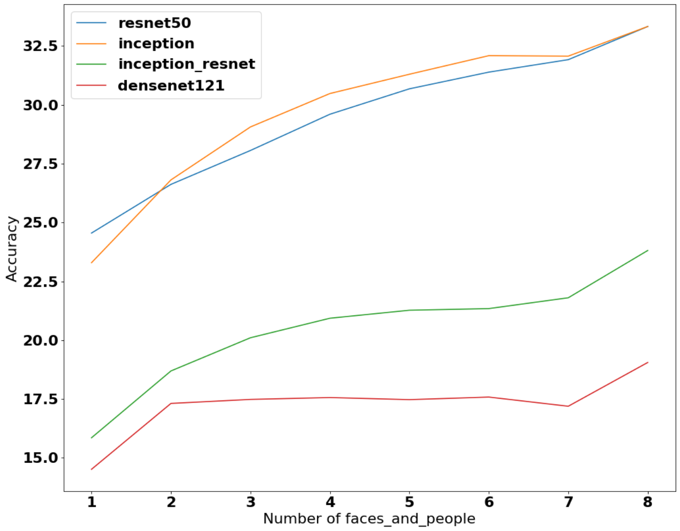


Figure 36: Increasing accuracy by number of faces & people in the picture in the ensemble model.

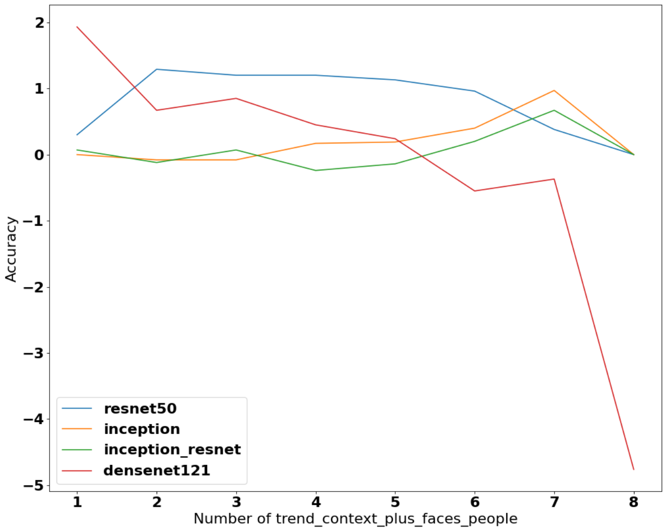


Figure 37: Contribute to accuracy given by the context in the ensemble model.

From these results, we can infer that the importance of the context is decreasing proportionally to the amount of people in a photo and that accuracy increase proportionally to the number of faces and people in the picture as shown in Fig.31 and 34. From Fig. 36 we could infer that the accuracy increase is model-dependent, however, the accuracy is considerable higher than the one considering only faces and only people (as shown in Fig.31 and 34). This probably means that the limit of the model was reached.

We can conclude this section stating that even with this ensemble classifier, accuracy increases with the concept of temporal window (Tab. 11).

architecture	w=0	w=5	w=10	w=20
resnet50	17,78	63,99	83,33	95,95
inception	15,77	63,45	83,17	95,63
inception_resnet	15,87	63,54	82,63	95,7
densenet121	15,68	63	82,98	95,51

Table 11: Increasing accuracy with temporal window for baselines ensemble architectures.

3.5.6 A merged architecture

After build the ensemble architecture, I wonder if we could get more from the complementarity of the data. So, in order to explore new possibilities, I decide to build a new merged classifier which is kinda similar to the one described in 3.5.5 but which instead combine features from different IMAGO datasets and it's trainable! This settings was made to discover if the network could learn how to combine features from a variable number of input images (number faces and people depending on the picture). The model architecture is described in Fig. 38.

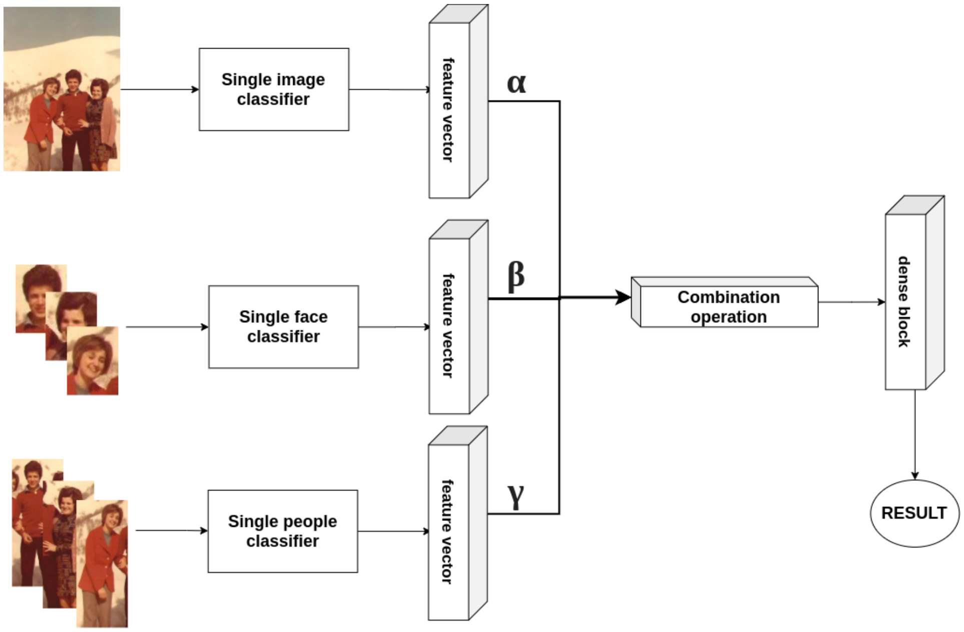


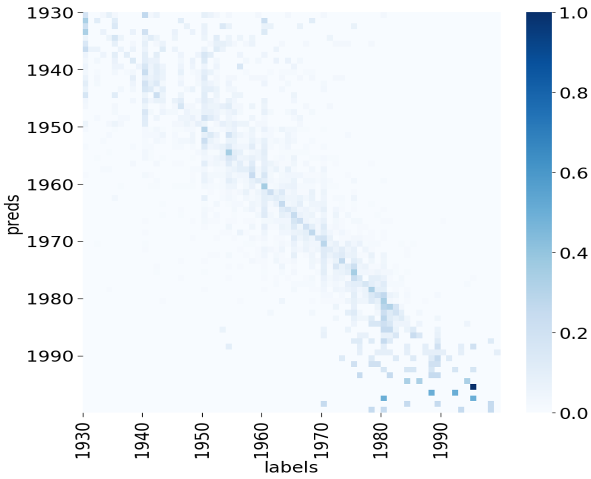
Figure 38: Merged classifier architecture.

This architecture is different from the previous one mainly because of how it combine features from different image sources. Indeed, each classifier return a feature vector (that can be the result of a mean between feature vectors in case of face and people classifier). Each of this feature vector is then multiplied by a learnable factor and then all of them are summed to produce a unique feature vector which is provided to the dense block to make the classification. So, this means that the real feature vector is a **linear combination of the feature vectors generated by single classifiers**. I trained this architecture with the same training settings of previous classifiers, except for the batch size that was increased up to 128. Results for different baselines are listed in Tab. 12.

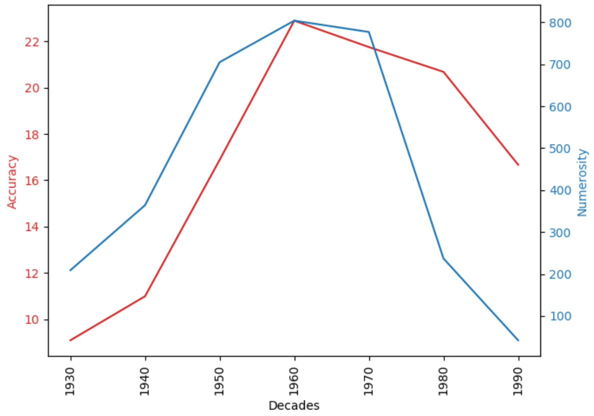
architecture	resnet50	inception	inception_resnet	densenet121
num_classes	70	70	70	70
fine tuning	partial	total	total	total
lr	0.0001	0.0001	0.0001	0.0001
epochs	40	40	40	40
optimizer	adam	adam	adam	adam
weight decay	0.0005	0.0005	0.0005	0.0005
combination method	sum	sum	sum	sum
test accuracy	18,71	17,14	16,44	16,76
mean error	5,06	5,14	5,11	5,19
dv_std error	6,01	5,92	5,85	6,01
median error	4,45	2,97	2,97	2,97

Table 12: Results for merged classifier on different DL architectures.

From these results, seems that this model learn something that previous models couldn't. This confirm once again that exist a kind of complementarity within different images from different imago datasets.
To check if there are changes in the confusion made between classes, I have also plotted the confusion matrix of the best classifier (which is considered Resnet50) in Fig. 39a. Justifications for results in Fig. 39a are the same discussed in 3.5.4.



(a) Confusion matrix merged classifier.



(b) Merged classifier accuracy by decades compared to numerosity.

Even in this model, the accuracy increases proportionally to the decades considered. Differently from the previous architecture, we can't know if the accuracy increase considering only faces and people and we are forced to consider this factor together with the context one. Baselines using different baselines as backbones are listed in Fig. 40. We can conclude this section stating that even with this merged classifier, accuracy increases with the concept of temporal window (Tab. 13).

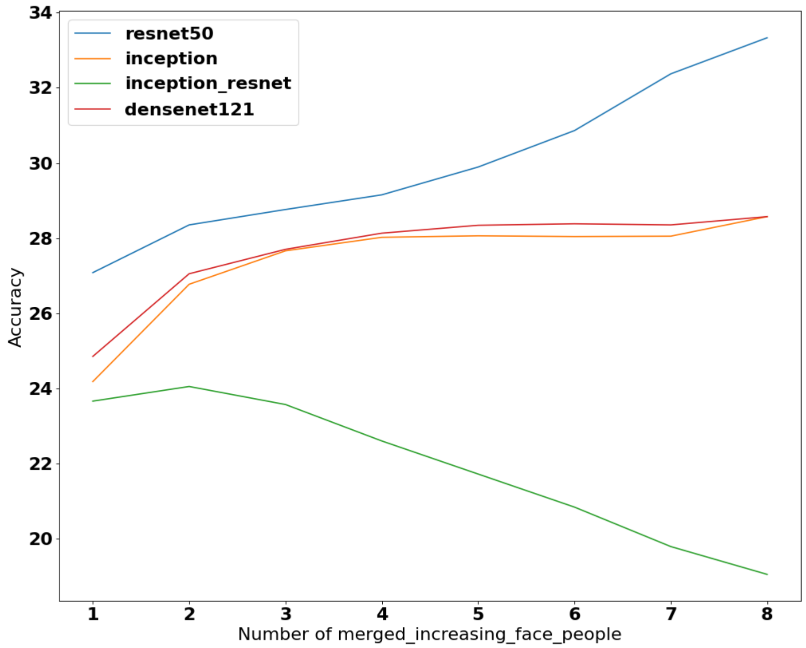


Figure 40: Increasing accuracy by number of faces & people in the picture in the merged model.

architecture	w=0	w=5	w=10	w=20
resnet50	18,71	67,59	86,17	96,97
inception	17,14	67,56	86,3	97,1
inception_resnet	16,44	67,62	86,3	97,16
densenet121	16,76	66,79	85,98	96,97

Table 13: Increasing accuracy with temporal window for baselines merged architectures.

3.5.7 What the model learn: Grad-CAM visualization

The gradient-weighted class activation mapping (Grad-CAM) is a technique to understand why a deep learning network makes its classification decisions. Grad-CAM, invented by Selvaraju and coauthors [38], uses the gradient of the classification score with respect to the convolutional features determined by the network in order to understand which parts of the image are most important for classification. This technique has also an hidden meaning: it could be used to find new pattern in images in which the classification is not obvious for human too. For example, given an image of a dog, a human is able to recognize it in less than a second...but this is not true for a cancer or to recognize an era! For this reason, I have adapted an open-source implementation of the Grad-Cam algorithm to explore what features characterize each era. In particular, i have adapted this implementation to fit the Resnet50 model, which is considered the model that has best generalized the task. An extract of the Grad-CAM features is in Fig. 41, in which I chose some pics correctly classified for each decades that contains a variable number of faces and people. From these figures, it is evident that different models concentrate on different parts of the photo, which emphasize the fact that there is a complementarity within different part of this data. Moreover, is noticeable, that there are fashion and objects characteristics that models exploits to recognize the year (i.e, hairstyles, dresses, moustache, chairs).

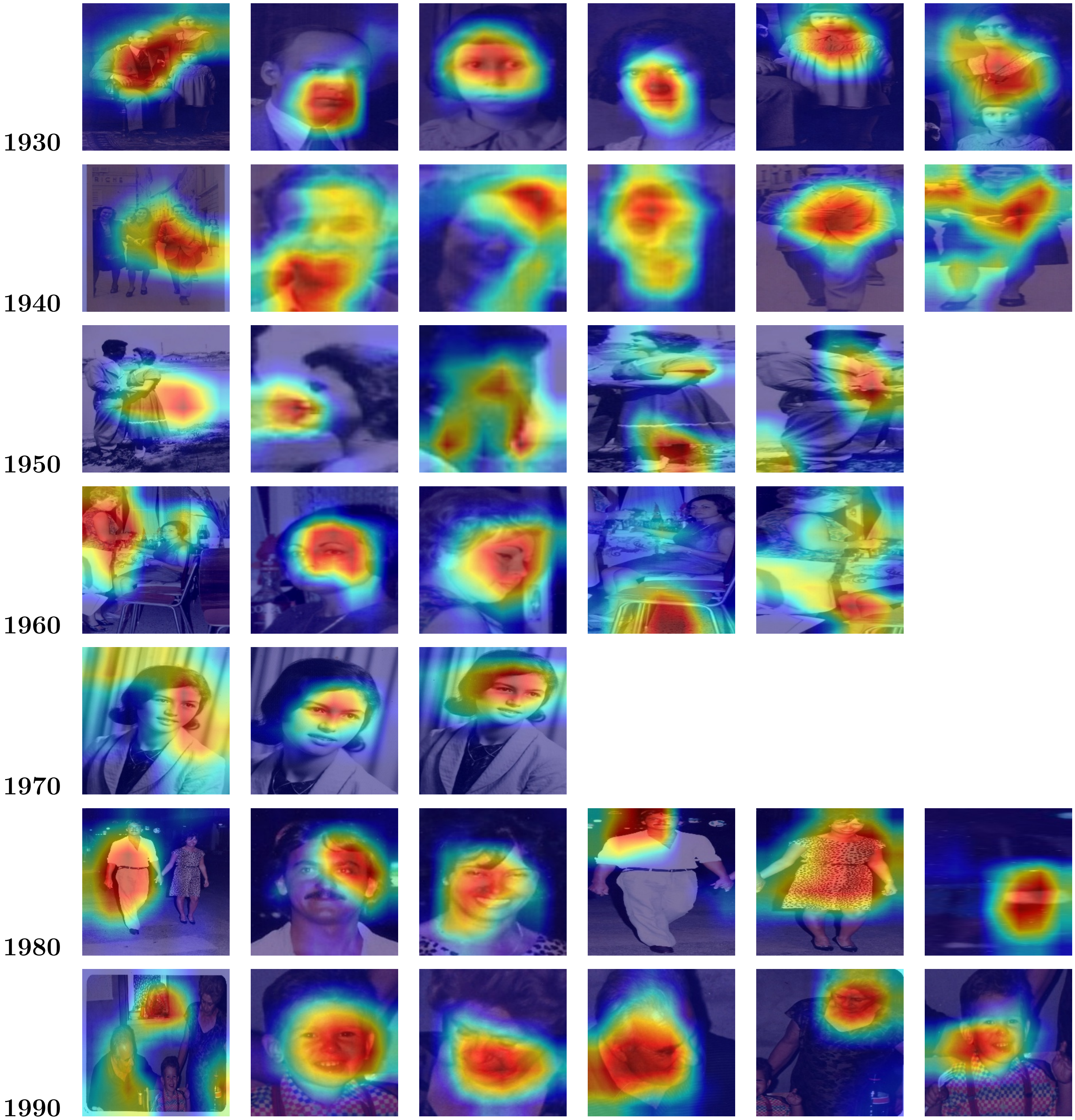


Figure 41: Gradcam analysis of different patches of an image within IMAGO spreaded over the twentieth century.

4 The socio-cultural context classification task

4.1 Task description

The socio-cultural context classification task could be described as a **multi-classification problem** in which the model, given a photo, has to guess the context (taken from the set described in Section 2) in which it was taken.

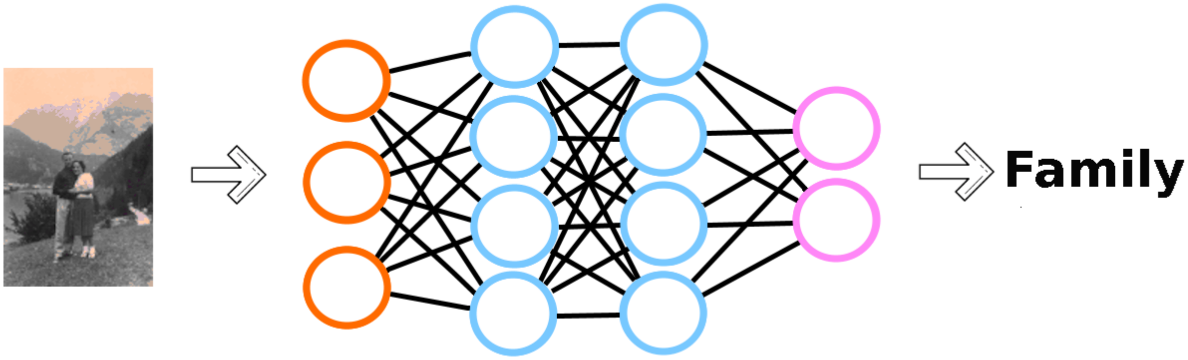


Figure 42: Example of socio-cultural context classification

There are no papers about this specific task and this means that this work introduces a new kind of task unexplored before. However, there are some works that are related to it.

4.2 Related Works

Different image classification tasks aimed to explore the presence of specific objects (e.g., animals, plants, vehicles, musical instruments). An important and widely used dataset for these tasks is ImageNet [39] that comprehend 14,197,122 digital images belonging to 1,000 classes. Nevertheless, other tasks exist in order to classify images extrapolating abstract concepts (e.g., fashion styles, painting styles).

In [34] the authors aimed to classify people fashion style analyzing documentary photographs. In particular, they implemented a CNN-based embedding architecture in order to classify the presence of fixed relevant features (e.g., sleeve length, collar presence, wearing hat, clothing pattern) in the images. Subsequently they exploited these features to classify people fashion styles in different cities making use of an unsupervised method. However, there is a limit to the granularity of styles that they obtained. For instance, their model do not learn a clean separation between eyeglasses and sunglasses, as those

were not labeled as distinct attributes, and their style embedding does not separate those two categories. In addition, their method is limited to analyzing the upper body of the person.
Another work [40] analyzed and classified the style of paintings considering the images and some historical context-based side features (time, birthplace, art movement) in order to assist the visual ones. In particular, they proposed to synthesize historical knowledge into the image label through Label Distribution Learning which is further employed to generate a proper label distribution. Multiple label distributions are finally encapsulated into a learning model which can significantly assist a CNN architecture, which extract relevant features from images, improving the classification performance.
In this section I introduce a DL based method to classify documentary photos in order to extrapolate the socio-cultural context. This task could be considered as an abstract concept task based on historical, sociological and cultural aspects. Differently from the aforementioned works, I have analyzed a dataset that comprehend only analog documentary photography within the twentieth century and implement a CNN architecture in order to classify images without adding side features. In addition, i have considered consider all the image content from different points of view and not only people face or people torso.

4.3 My approach

As described in Section 2 three kind of visual information spanned over the twentieth century are available : original documentary photos, faces and people patches. All these three information could be used to design three different kind of networks that consider respectively original images, faces and people patched. This decision was not easy to take because we can also design a multi-input architecture . . . but I have done this too!
So, in order to carry out the socio-cultural classification task i have created and training 4 different kinds of DL modes:

- Single Image classifier: that takes into account the original picture;
- Single face classifier: that takes into account faces patches from the original picture;
- Single people classifier: that takes into account people patches from the original picture;
- Merged classifier: that takes into account all three kind of images.

For each model i will describe the architecture, training settings and empirical results in 4.6.

4.4 Creation of train, validation and test set

In order to train a DL model and to test its capabilities we must divide the dataset into three sets:

- The training set, which is used to train the model;
- The validation set, which is used to measure performances and compare it with other models;
- The test set to measure qualitative performances such as accuracy, precision, recall and F1 score.

Of course this sets must have a null intersection. This means that we have to divide train examples from validation and test ones. And this must be done for the entire IMAGO dataset (that includes faces, people and original images). This was done choosing randomly 80% of records for the train and validation set and 20% for the test set **for each class**. So this means that the train set is composed by 70% of pictures in affectivity, politics ...

4.5 Pre-processing and data augmentation

As in many works of Computer vision in the field of deep learning, some techniques of data augmentation and pre-processing were applied to our data in order to reduce overfitting risks, makes the learning more stable and to increase the model performances. As described in 2 we decide to avoid any kind of classical image pre-processing, because of our datasets nature. However, we can for sure perform some kind of data augmentation. Despite there are a lot of techniques that we can apply, we decide to use only three data augmentations algorithms:

- Random Cropping;
- Shifting;
- Horizontal flipping;

This three were chosen because they are the only techniques that produce significant images from original ones. Indeed, random cropping and shifting are useful because they allow the model to concentrate on different kind of details within the photo while the horizontal flipping allow to see same objects but in different locations, and this could be a form or feature extraction regularization. As we know regularization adds prior knowledge to a model and this explain why data augmentations techniques are also a form of regularization. On the other hand, vertical flipping or image rotation, even if them are data augmentation techniques, could create some bias in the network (e.g., consider to apply these techniques on faces and people patches).

4.6 Experiments

In this section, i am going to review the experiments made for each different models described in Section 4.3. For terms syllabus check Section 3.5.

4.6.1 Single image classifier

Architecture The single Image classifier is a DL model that takes in input an image from IMAGO dataset and try to guess the context in which this was taken (Fig. 43).

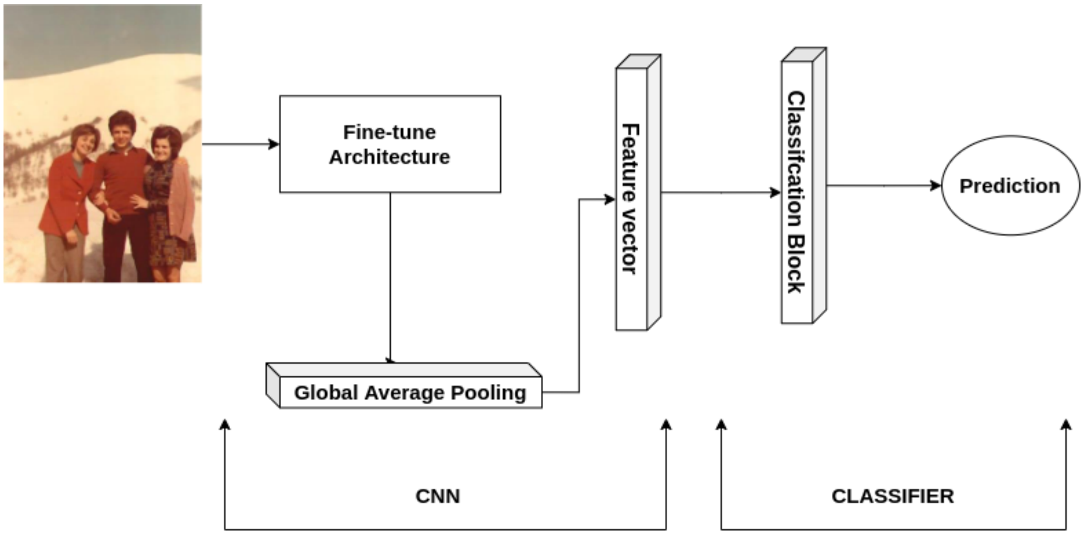


Figure 43: Single image classifier architecture.

As you can see, in Fig. 43 there is a token named as ”Fine-tune Architecture”. This is why we have fine-tuned different known deep learning architecture, exploiting IMAGENET weights, to create what is usually call **baselines**. In Fig.14 I report all the results obtaining with different networks and different hyper-parameter configurations.

architecture	resnet50	inception	densenet121	inception_resnet
tipo di dato	original	original	original	original
num_classes	9	9	9	9
fine tuning	partial	total	total	total
lr	0.0001	0.0001	0.0001	0.0001
epochs	40	40	40	40
optimizer	adam	adam	adam	adam
weight decay	0.0005	0.0005	0.0005	0.0005
test accuracy	64,35	64,08	63,7	61,8

Table 14: Single image classifier architecture baselines.

As for the **dating task** seems that the accuracy changes slightly depending on the architecture and how these were trained. However, to explore better what are the most learned and confused classes, I report a plot in which accuracy by class is listed (Fig.44) From this plot, we can also infer that classes with symbolical object and gestures, such as affectivity, motorization, politics, school and rites, are more easy to learn for this model (because they contains repeated patterns). Visual patterns learned by this model will be described in 4.6.5.

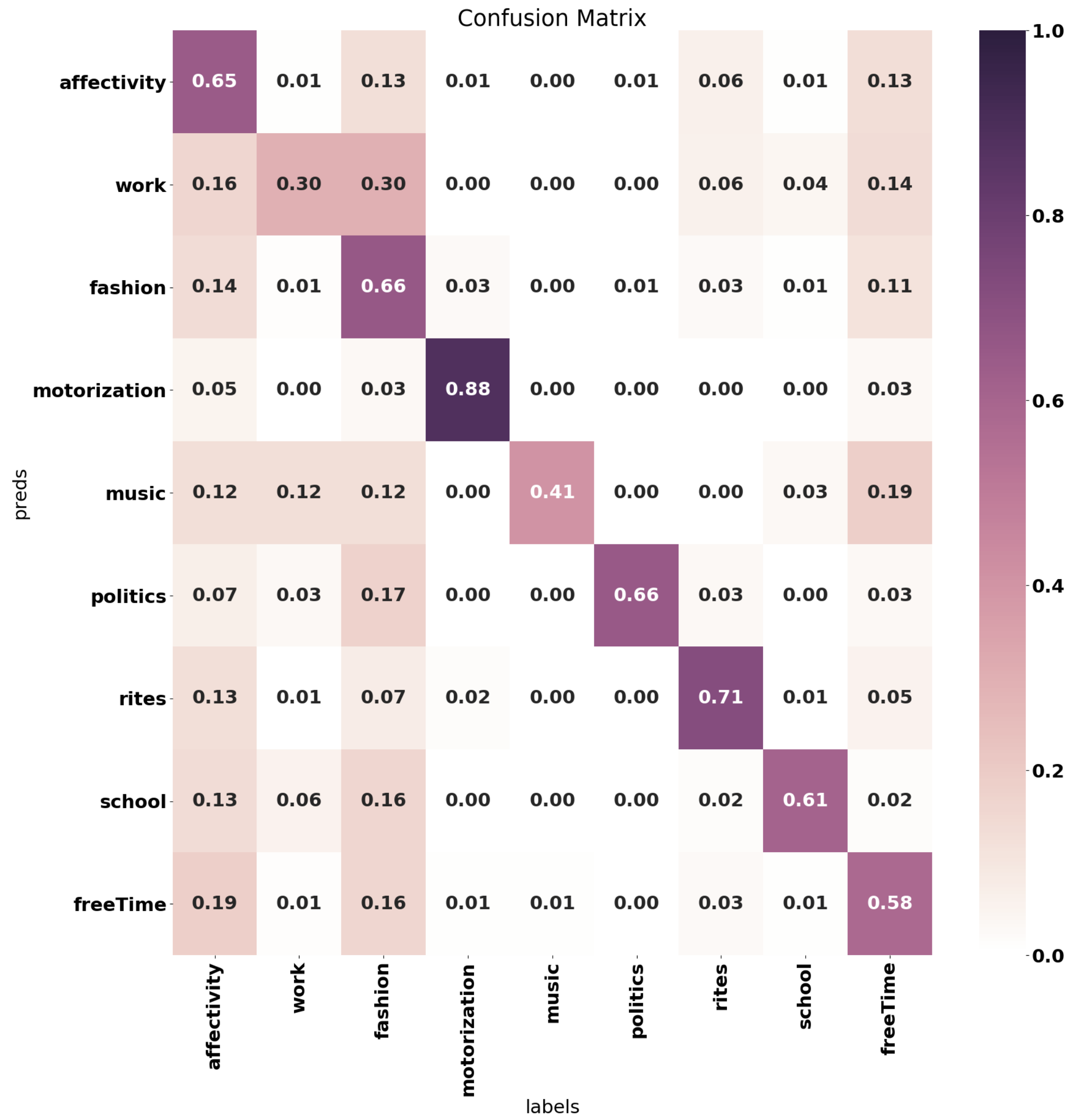


Figure 44: Confusion matrix for socio-cultural classification made by single image classifier.

4.6.2 Single face classifier

The single Image classifier is a DL model that takes in input an image from IMAGO-FACE dataset and try to guess the context in which this was taken (Fig. 45).

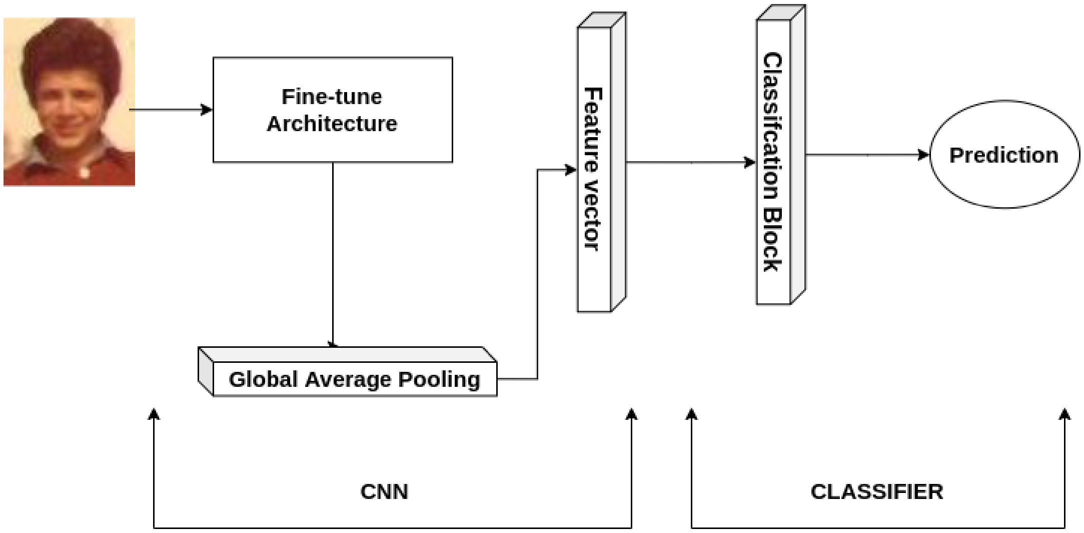


Figure 45: Single face classifier architecture.

As you can see, in Fig. 45 there is a token named as ”Fine-tune Architecture”. This is why we have fine-tuned different known deep learning architecture, exploiting IMAGENET weights, to create what is usually call **baselines**. In Fig.15 I report all the results obtaining with different networks and different hyper-parameter configurations.

architecture	resnet50	inception	densenet121	inception_resnet
tipo di dato	original	original	original	original
num_classes	9	9	9	9
fine tuning	partial	total	total	total
lr	0.0001	0.0001	0.0001	0.0001
epochs	40	40	40	40
optimizer	adam	adam	adam	adam
weight decay	0.0005	0.0005	0.0005	0.0005
test accuracy	41,3	43,37	41,38	40

Table 15: Single image classifier architecture baselines.

To explore what are the most learned classes, I report the confusion matrix in which accuracy by class is listed (Fig.46).

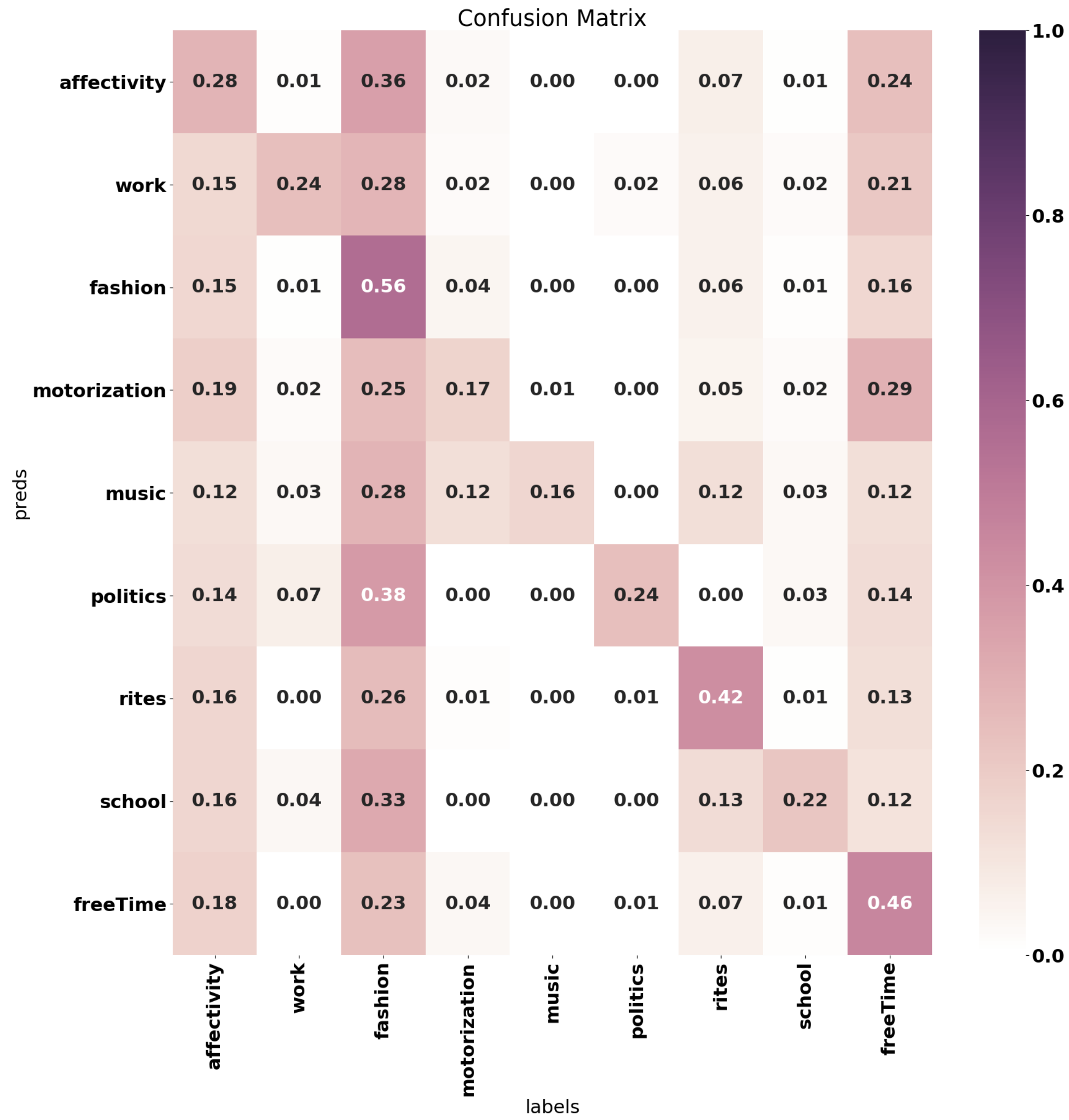


Figure 46: Single face classifier confusion matrix.

From this plot, we can infer that only some specific class had benefit of face details (i.e., freeTime,fashion) and this is reasonable, since the context of a picture is fundamental to extract abstract concepts that characterize this task. As it is considered useless, visual pattern learned by this model will be ignored.

4.6.3 Single person classifier

The single Image classifier is a DL model that takes in input an image from IMAGO-PEOPLE dataset and try to guess the context in which this was taken (Fig. 47).

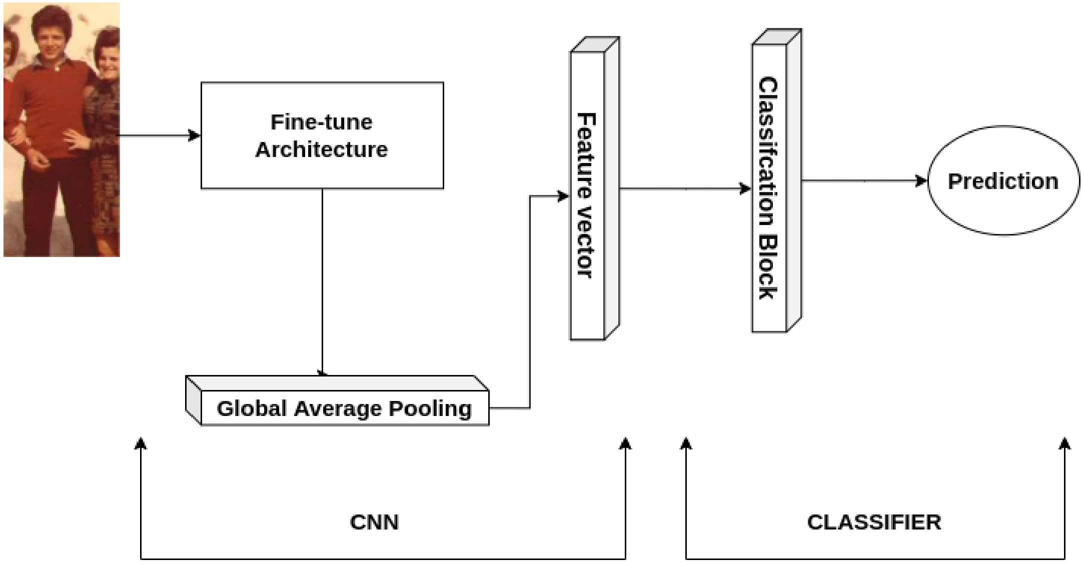


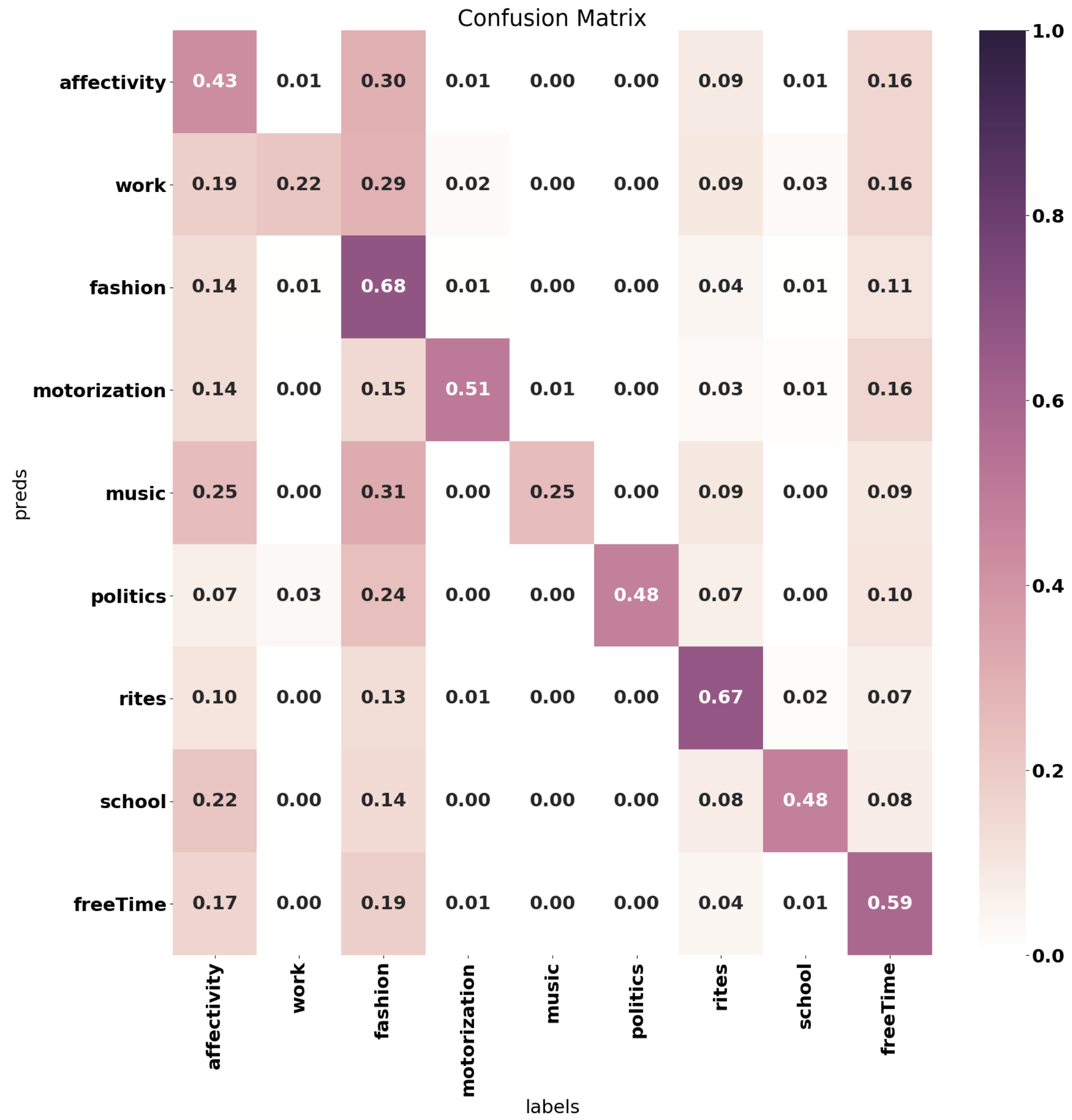
Figure 47: Single person classifier architecture.

As you can see, in Fig. 47 there is a token named as ”Fine-tune Architecture”. This is why we have fine-tuned different known deep learning architecture, exploiting IMAGENET weights, to create what is usually call **baselines**. In Fig.16 I report all the results obtaining with different networks and different hyper-parameter configurations.

architecture	resnet50	inception	densenet121	inception_resnet
tipo di dato	original	original	original	original
num_classes	9	9	9	9
fine tuning	partial	total	total	total
lr	0.0001	0.0001	0.0001	0.0001
epochs	40	40	40	40
optimizer	adam	adam	adam	adam
weight decay	0.0005	0.0005	0.0005	0.0005
test accuracy	56	57	55,63	54

Table 16: Single image classifier architecture baselines.

To explore what are the most learned classes, I report, as usual, the confusion matrix (Fig.48)



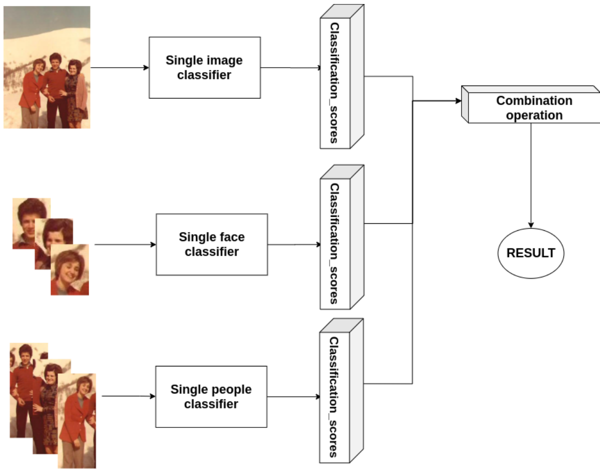
	affectivity	work	fashion	motorization	music	politics	rites	school	freeTime
affectivity	0.43	0.01	0.30	0.01	0.00	0.00	0.09	0.01	0.16
work	0.19	0.22	0.29	0.02	0.00	0.00	0.09	0.03	0.16
fashion	0.14	0.01	0.68	0.01	0.00	0.00	0.04	0.01	0.11
motorization	0.14	0.00	0.15	0.51	0.01	0.00	0.03	0.01	0.16
music	0.25	0.00	0.31	0.00	0.25	0.00	0.09	0.00	0.09
politics	0.07	0.03	0.24	0.00	0.00	0.48	0.07	0.00	0.10
rites	0.10	0.00	0.13	0.01	0.00	0.00	0.67	0.02	0.07
school	0.22	0.00	0.14	0.00	0.00	0.00	0.08	0.48	0.08
freeTime	0.17	0.00	0.19	0.01	0.00	0.00	0.04	0.01	0.59

Figure 48: Confusion matrix socio-cultural task made by single person classifier .

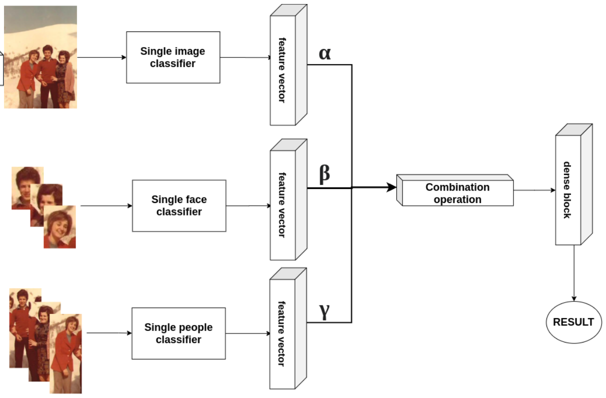
From this plot, we can infer that only some specific class had benefit on people details (i.e., fashion and rites) and this is reasonable, since the context of a picture is fundamental to extract abstract concepts that characterize this task. As it is considered useless, visual pattern learned by this model will be ignored.

4.6.4 An ensemble and a merged architecture

As shown in previous sections, we have reached good results using exclusively original images without faces and people patches. This is reasonable since faces and people segments often do not contain symbolical objects. However, some classes benefits from this fact (i.e, rites and fashion) and I wonder if this benefit could be exploited. In order to answer this question I have created an ensemble architecture, that trying to guess the date taking into account: (a) the original picture (b) all faces within the original picture (c) all people within the original picture (Fig. 49a,49b). These architecture are the same discussed in Section 3.5.6.



(a) Semantic ensemble classifier architecture.



(b) Semantic merged classifier architecture.

Unfortunately, results obtained combining different data and models doesn't improve the accuracy and, in certain classes, were lowered. For this reason we decide to not use those combined models, because it is considered a waste of resources, but using only the Single image classifier described in 4.6.1.

4.6.5 What the model learn: Grad-CAM visualization

As in 3 I have adapted an open-source implementation of the Grad-Cam algorithm to explore what features characterize each semantic class. In particular, i have adapted this implementation to fit the Resnet50 model, which is considered the model that has best generalized the task. An extract of the Grad-CAM features is in Fig. 50, in which I chose some pics correctly classified for each class. As stated in previous section, only visual patterns extracted by the model described in 4.6.1 are explored.

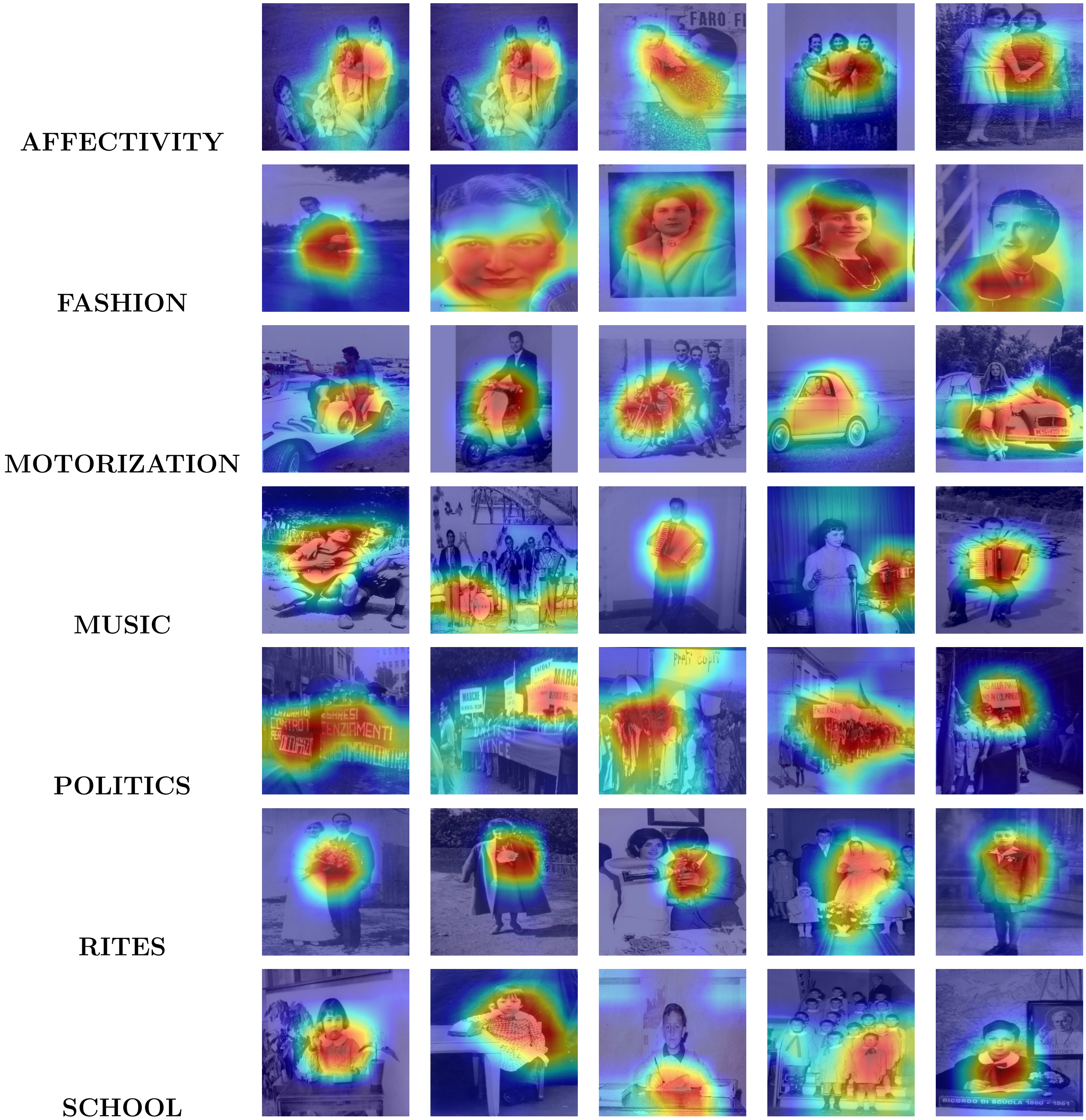


Figure 50: Gradcam analysis made with entire images from IMAGO.

It turns out that the model have learnt some symbolical abstract concept within these picture, different for each classes. In particular, the model has learnt that:

- Particular gestures are symbolical of affection between people (e.g., hugs, kisses, hold a baby, shake hands);
- There are objects like earrings, necklaces, lapels and but also particular hairstyle belong to the fashion class;
- Specific objects like earrings, necklaces, lapels and but also particular hairstyle belong to the fashion class;
- All kind of vehicles (and their parts) belong to the class motorization;
- All kind of musical instruments (and their parts) belong to the class music;
- Exist specific political objects and situations that represent a political event (i.e., manifest);
- There are symbolical rites objects and gestures that represent a rite event (i.e., white dress, flowers, pour a drink, cheers);
- Exist specific political objects and situations that represent a political event (i.e., manifest);
- There are patterns to identify people that are attending school (i.e., babies with particular dresses, a group of children dressed alike, pens);

Making those reasoning, is not surprising the fact that the model was able to categorize picture in classes motorization and music, because these are classes already defined in IMAGE-NET. However, all the other classes contains concept there are completely new and unexplored, and the results obtained are for sure a good starting line.
However, this model has some bias that lead to wrong predict the socio-cultural context of an image. As you can check in Fig. 51 there are some pictures wrong classified...but the greatest part of them, are errors related to the classification problem itself. An example of model bias is in the first figure, we can see that the model have concentrated on woman's dress and label this picture as fashion, while the real label was affectivity. Examples of data biases rely on the second and the third figure: the model labelled them as **school** and **freeTime** making a wrong classification....but is it really considered wrong?

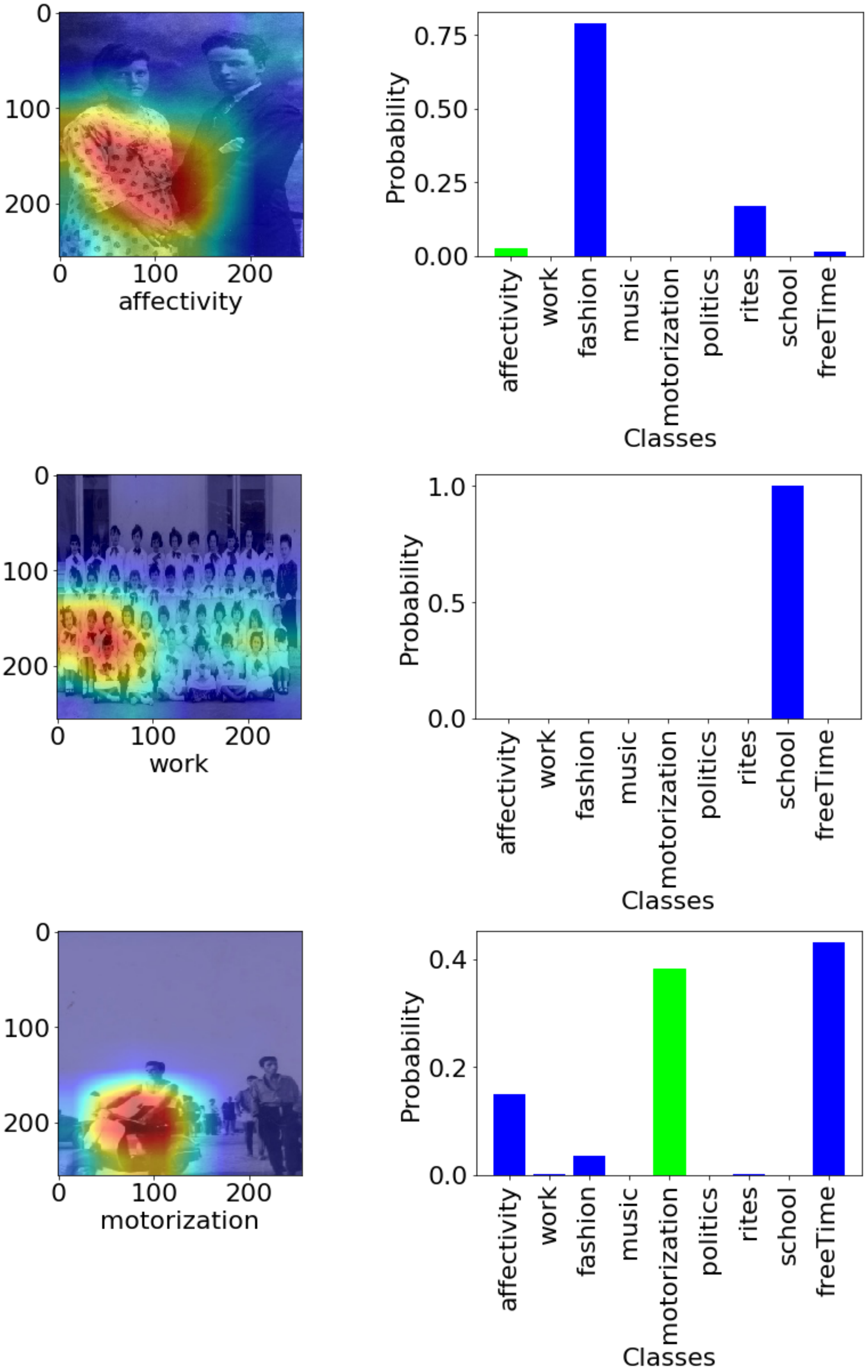


Figure 51: Gradcam analysis made with entire images from IMAGO.

5 Conclusions

5.1 Conclusions

In this work, the IMAGO data-set is introduced and explored from an historical point of view. I also described the issues related in dealing with analog images from an artificial intelligence point of view and the reason why there isn't a clear way to solve them. I have introduced a new model that exploit different patches within an historical documentary photos to identity the year in which this photo was taken. In this context, I have discovered that more faces and more people the picture has, greater is the accuracy that the a model could achieve. Moreover, i find out that there is a complementarity between the knowledge that different model learns that could be exploited to gain even greater accuracy. In addition, i have introduced a new kind of task referred as "socio-cultural context classification“ that consists in identifying the socio-cultural context within a picture. Relative to this task, it is noticeable that merge knowledge from different models is useless, since only the entire photo could give the network a general overview of what is happening, while faces and people patches couldn't. In addition, thanks to the Grad-Cam analysis, it turns out that the deep learning models were able to classify correctly **abstract concepts** (i.e., affectivity, politics, rites) focusing on **concrete symbols** (i.e, kisses and hugs for the affection class, event posters for the politics one, marriage cakes and toasts for the rites class).

5.2 Future Works

For what concern future develops for this work, i aim to test different approaches in order to improve the model accuracy and stability by:

- Erasing the conceptual classes free-Time and work, because they are negligible from an artificial intelligence point of view, reassigning photos labeled within these exploiting an unsupervised algorithm;
- Test an unsupervised approach for both dating and semantic task;
- Test a semi-supervised learning approach with balanced classes.

References

[1] B. Lavédrine, "Photography and its preservation: Continuity and changes in the digital era," 2017.

[2] M. Wevers and T. Smits, "The visual digital turn: Using neural networks to study historical images," *Digital Scholarship in the Humanities*, vol. 35, no. 1, pp. 194–207, 2020.

[3] F. Palermo, J. Hays, and A. A. Efros, "Dating historical color images," in *Computer Vision – ECCV 2012* (A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, eds.), (Berlin, Heidelberg), pp. 499–512, Springer Berlin Heidelberg, 2012.

[4] T. Salem, S. Workman, M. Zhai, and N. Jacobs, "Analyzing human appearance as a cue for dating images," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–8, 2016.

[5] S. Ginosar, K. Rakelly, S. Sachs, B. Yin, C. Lee, P. Krahenbuhl, and A. A. Efros, "A century of portraits: A visual historical record of american high school yearbooks," 2015.

[6] E. Müller, M. Springstein, and R. Ewerth, ""When was this picture taken?"–image date estimation in the wild," in *European Conference on Information Retrieval*, pp. 619–625, Springer, 2017.

[7] D. Calanca, "Italians posing between public and private. theories and practices of social heritage," *Almatourism-Journal of Tourism, Culture and Territorial Development*, vol. 2, no. 3, pp. 1–9, 2019.

[8] A. photography blog, "What is analog photography?," 2019.

[9] H. S. Becker, "Visual sociology, documentary photography, and photojournalism: It's (almost) all a matter of context," *Visual Sociology*, vol. 10, no. 1-2, pp. 5–14, 1995.

[10] R. Gregg, "The beginner's guide to documentary photography," 2019.

[11] I. Yaqoob, I. Abaker, T. Hashema, A. Gani, S. Mokhtar, E. Ahmed, N. B. Anuar, and A. V. Vasilakos, *Big data: From beginning to future.* Elsevier, 2016.

[12] R. E. Bryant, R. H. Katz, and E. D. Lazowska, *Big-Data Computing: Creating revolutionary breakthroughs in commerce, science, and society: A white paper prepared for the Computing Community Consortium committee of the Computing Research Association.* CCC-Led White Papers, 2008.

[13] Omnicoreagency, "Instagram by the numbers: Stats, demographics & fun facts," 2020.

[14] Igi-global, "Image retrieval," 2020.

[15] A. T. Adela Barriuso, "Notes on image annotation," 2020.

[16] GeeksForGeeks, "Image manipulation," 2020.

[17] M. works, "Image analysis," 2020.

[18] E. systems, "What is machine learning?," 2020.

[19] G. A. Canepa, *What You Need to Know about Machine Learning.* Packt, 2016.

[20] M. Hallard, "Introduction to semi-supervised learning and adversarial training," 2019.

[21] J. Rodriguez, "Understanding semi-supervised learning," 2017.

[22] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning.* MIT Press, 2016.

[23] D. Shaveta, K. Munish, A. M. Rohit, and K. Gulshan, "A survey of deep learning and its applications: A new paradigm to machine learning," *Springer-Link*, 2019.

[24] N. O. Mahony, S. Campbell, A. Carvalho, S. Harapanahalli, G. Velasco-Hernandez, L. Krpalkova, D. Riordan, and J. Walsh, "Deep learning vs. traditional computer vision," 2019.

[25] Sung Cheol Park, Min Kyu Park, and Moon Gi Kang, "Super-resolution image reconstruction: a technical overview," *IEEE Signal Processing Magazine*, vol. 20, no. 3, pp. 21–36, 2003.

[26] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, C. C. Loy, Y. Qiao, and X. Tang, "Esrgan: Enhanced super-resolution generative adversarial networks," 2018.

[27] K. Zhang, W. Zuo, and L. Zhang, "Deep plug-and-play super-resolution for arbitrary blur kernels," 2019.

[28] L. Fan, F. Zhang, Fan, and H. et al., "Brief review of image denoising techniques," 2019.

[29] K. A. I.M. El-Henawy, A. E. Amin and H. Adel, "A comparative study on image deblurring techniques," 2014.

[30] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," 2019.

[31] F. H. K. dos Santos Tanaka and C. Aranha, "Data augmentation using gans," 2019.

[32] P. Sorcinelli, "imago: Laboratorio di ricerca storica e di documentazione iconografica sulla condizione giovanile nel xx secolo," 2005.

[33] Thanh Nguyen, "Yolo-face," 2018.

[34] K. Matzen, K. Bala, and N. Snavely, "Streetstyle: Exploring world-wide clothing styles from millions of photos," *arXiv preprint arXiv:1706.01869*, 2017.

[35] Joseph Redmon, "YOLO: Real Time Object Detection."

[36] K. Zhang, "Kair: Pytorch toolbox for image restoration," 2020.

[37] K. Zhang, W. Zuo, and L. Zhang, "Ffdnet: Toward a fast and flexible solution for cnn based image denoising," 2017.

[38] Selvaraju, R. R., M. Cogswell, R. V. A. Das, D. Parikh, and D. Batra, "A century of portraits: A visual historical record of american high school yearbooks," 2017.

[39] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[40] J. Yang, L. Chen, L. Zhang, X. Sun, D. She, S.-P. Lu, and M.-M. Cheng, "Historical context-based style classification of painting images via label distribution learning," in *Proceedings of the 26th ACM international conference on Multimedia*, pp. 1154–1162, 2018.