VISUALFLICKR

Visual sentiment analysis of Flickr images based on MVSO

Giovanni Di Prisco

UNIVERSITÀ DEGLI STUDI
DI SALERNO

Summary

1. Introduction

The visual sentiment analysis is the attempt to infer sentiment and emotions from visual contents. The aim of this discipline is to understand what are the feelings expressed by a photo or a video, in other words, what a person can feel watching a visual content (in our project the focus is the visual sentiment analysis of photos).

Typically, this goal was achieved analysing some low-level features of the photographs, like coloration, gradient orientation exc.

In our work we follow a methodology based on a new point of view, combining the use of a neural network and an ontology. The innovation of this method is the fusion of the visual analysis and the text analysis.

Figure 1 - Two tweets with the description by their authors

In Figure 1 are shown two images shared on Twitter. To really understand the sentiments expressed by these two photos is necessary to read the descriptions and watching the pictures at the same time. One thing without another is useless to make a good sentiment prediction.

1.1 VSO and ANPs

The core of our project is MVSO (Multilingual Visual Sentiment Ontology), but it's good to make a step back and start talking about VSO (Visual Sentiment Ontology). Citing the description used on the website of VSO: "The goal of this work was to design an ontology of semantic concepts which have a link to an emotion, have a strong sentiment, and are frequently used on online platforms like Flickr or YouTube. Currently the visual sentiment ontology includes more than 3,000 concepts, with each concept being composed of an adjective and noun e.g. 'beautiful sky' or 'sad eyes'"[1].

VSO is based on a well-known psychological model: the wheel of Plutchik (Figure 2). It contains 24 emotions, and it's possible to browse the ontology according to the emotions present in the wheel.

[1] https://visual-sentiment-ontology.appspot.com/

Figure 2 - Wheel of emotions by Robert Plutchik

In the description mentioned before, it was said that the concepts of the ontology consisting of ANPs: adjective-noun pairs, but what is an adjective-noun pairs?

An adjective-noun pair is a pair formed by an adjective and a noun. The idea behind the use of the ANPs is the fact that an adjective usually has a sentimental value (adjective like ‘good’, ‘bad’ express an emotion), but the same thing can’t be said about nouns, because a noun sometimes can have a “neutral” value.

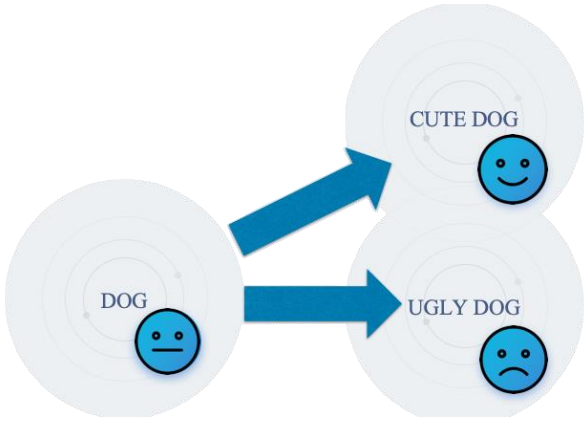

Figure 3 - Comparison by a noun and two ANPs

In figure 3 is shown an example. The noun ‘dog’ doesn’t have a sentimental value, it can’t be said anything about it, but ‘dog’, combined with an adjective (‘cute’ or ‘good’), transmit “something”. In this way all the concept contained in the ontology have a link to an emotion.

1.2 MVSO

The evolution of VSO is MVSO. The innovation of MVSO is the fact that the use of ontology is not restricted to the English language. In this the new ontology there are two more things to consider: the language and the culture. In fact, sometimes a particular noun or adjective assumes a specific meaning (and provokes specific emotions) according to the language in which is analysed. In figure 4 are shown the six languages contained in MVSO.

Figure 4 - The six languages contained in MVSO

1.3 MVSO Detector

The last thing that missing in our framework is a link between the ontology and the pictures. This missing part is constituted by the MVSO Detector. With MVSO Detector, trained on a big dataset of pictures, we can extract ANPs from an image; these ANPs are used to evaluate the sentiment score using in combination two popular and publicly available sentiment tools: *SentiStrenght* and *SentiWordnet*.

In *VisualFlickr*, using all these things, we have tried to reach the target to make a sentiment prediction of photos downloaded from the famous social network Flickr, and try to obtain several information from that prediction (like an users profiling). In the following sections the purposes and the features of the project are better explained.

2. Used technologies

Several technologies have been used in the development of *VisualFlickr*.

We needed different tools to build up our project:

- a programming language
- a development environment
- a neural network framework

As programming language our choice fell on Python. Python is very popular object-oriented programming language. We decided to adopt Python for his several uses and the great number of libraries compatible with it.

Figure 5 - Python logo

The used development environment is PyCharm. It is an integrated development environment used in computer programming, specifically for the Python language, and for this reason we have used it.

Figure 6 - PyCharm logo

As neural network, we have chosen an Inception architecture initially trained on VSO and then fine-tuned on the MVSO dataset by the team that has developed the ontology.

The neural network model is a model based on Caffe (Convolutional Architecture for Fast Feature Embedding), a deep learning framework, written in C++ with a Python interface.

Figure 7 - Caffe logo

3. Purpose of the project

In order to really understand each purpose of it is appropriate to speak about them separately.

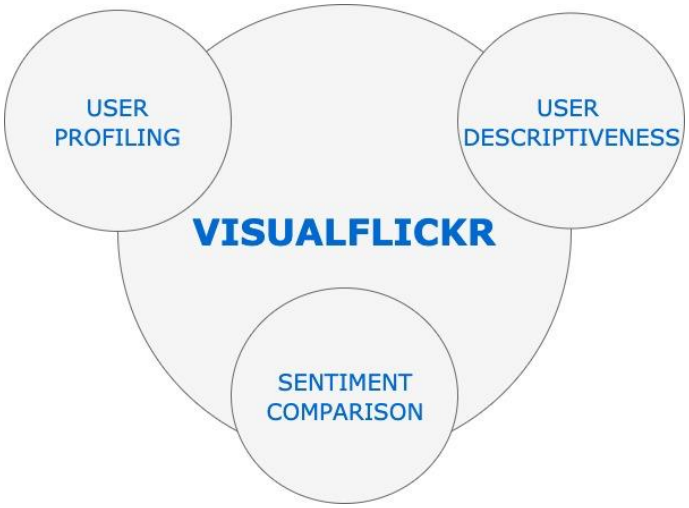

Figure 8 - In this diagram are shown the purposes of the project

In the further chapter we will speak about the pipeline of *VisualFlickr*, but it's a good thing to provide a general idea about the input and the output of our software.

VisualFlickr can take in input a picture, extract from it many ANPs, and after making some operations with them.

The most important output of our computation is the emotion and sentiment prediction of a picture using the ontology (MVSO). The main feature of our idea consists in the fact that we don't want to make only the sentiment prediction, but we also want to interpret the data we produce and try to discover something new analyzing them.

In our project we have worked with the images that the users have posted on their Flickr profile.

3.1 User profiling

The main purpose is to create a user profile based on the images that a user has posted on his social networks (in our case Flickr), producing data about emotion and sentiment inspired only by the visual recognition algorithm and averaging them for obtaining emotion and sentiment of the user.

This information should be used in many ways, for example showing to the user the right advisors according to the period he is passing through.

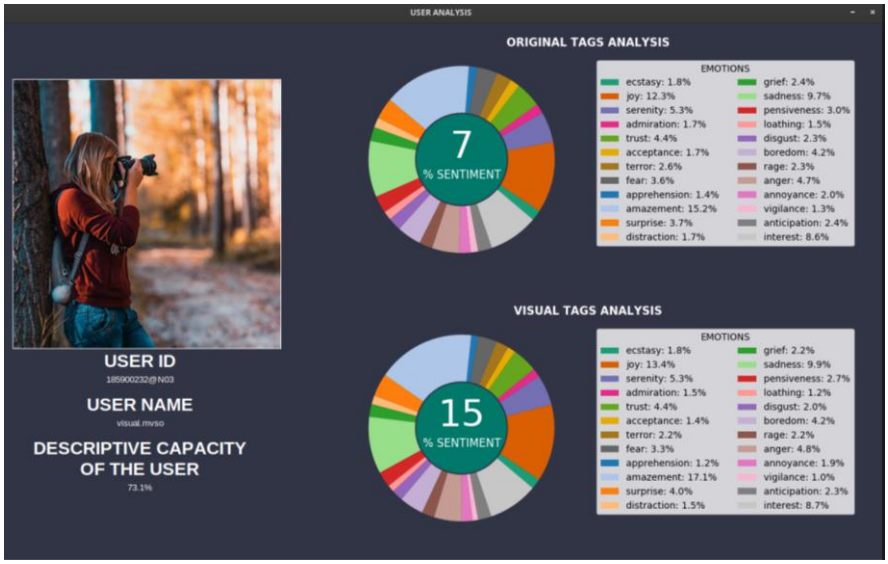

Figure 9 – Analysis of all the images uploaded by the Flickr user named "visual.mvso", showed in the summary scheme

3.2 Users descriptiveness

We start from the assumption that the neural network we have used in our project is infallible (or at least we hope).

Starting from this assumption, we can determine the ability of the user to describe the pictures that he has posted, comparing the ANPs extracted by the neural network with the ANPs present in the tags inserted by the user.

In this way we can compute a sort of score of the user regarding the precision of his description ability. This value could be used in order to understand if a sentiment and emotion estimate based on the only user tags should be considered reliable.

3.3 Emotions comparison

Another output, strictly correlated with the previous, is a comparison between the sentiment prediction based on the tags inserted by the users and the one based on the ANPs extracted by the neural network.

Thanks to this comparison, the user of *VisualFlickr* can determine how the emotions and the sentiment suggested by the owner of the images are different from the real information inspired by the images themselves.

If the results obtained thanks to the visual recognition algorithm is so far from the user tags analysis results, we can infer that the right emotion and sentiment suggested from the image is not fully or correctly deductible from the user's tags.

4. How *VisualFlickr* works

The application is provided with a command line interface in which the user can choice the language of the ontology and the command he wants to perform.

The operative pipeline of *VisualFlickr* should be divided in three parts:

- extraction of ANPs from the images, using a specific neural network
- extraction of the values of sentiment and emotions of the ANPs from the MVSO model
- elaboration of the data of interest.

User data and pictures used for the elaboration are downloaded from the famous platform Flickr (as said before) with a specific API [2].

We can use two different search mode from Flickr: the former made using the username or user identifier (and analyze picture by picture, automatically) and the latter made by tags.

The network used for the recognition of the ANPs is a pre-trained Caffe-model based on Inception.

The initialization of the visual classifier is performed using 4 files: the neural network model, its structure, the weights and the known labels (ANPs).

After an image is downloaded, the first layer of the networks carries out a pre-processing procedure downscaling the resolution. The last layer returns the indices of the row in the labels file of recognized ANPs.

Using the obtained ANPs we want to compute the emotional and sentimental values of the picture. Moreover, for every acknowledged emotion we want to associate a percentage value of the single emotion in the photo: we should see ANPs obtained from the neural network and emotions as rows and columns of a matrix, respectively, so we can obtain the mean of the single emotion performing the average on the columns.

Whatever the chosen language for the analysis, the recognizable objects compose a subset of MVSO, so we can always perform the described operation of emotion and sentiment averaging.

The consultation of the ontology is performed in unitary complexity because all the ontology, provided by its authors in a CSV file, is stored in a two-levels dictionary (ANP – emotion name – emotion value).

The next step consists in the repetition of the emotion and sentiment averaging operation but using the tags that the user has inserted to describe the photo, not the tags extracted by the network. As you can imagine, only the tags that are collected in the ontology can be interpreted to determine the associate emotions and sentiment. In particular, if *happy_dog* is an ANP of the ontology, the application is able to analyze the tags *happy_dog* and *happydog*.

[2] https://www.flickr.com/services/api/

Once the emotions and sentiment average has been calculated for each image and the emotions and sentiment average has been calculated for each associate owner-tags set, *VisualFlickr* provides to perform the last average based on the number of the images.

The mean emotions and sentiment for the selected user are shown in two pie charts: a chart based on the ANPs obtained from the neural network and the other chart based on the tags inserted by the owner of the images.

In this context the main purpose of *VisualFlickr* should be clear: building an emotional profile of the users based on the contents they share and estimate their ability to describe the images they load through tags.

To achieve this result is introduced a measure of the ability to describe of the users, made with a weighted average of the tags inserted by the user and the ANPs extracted by the network.

About this is appropriate to clarify which are the classes of the tags encountered during the process:

- **Owner tags**: tags inserted by the Flickr user attached to his photos, they can be recognized or not by the network, but usually the users use common tags in a common form (*adjective_noun* or *adjectivenoun*) with an high probability they are collected inside MVSO
- **Neural network decision**: subjects identified by the ANPs of MVSO, extracted from a single image; these ANPs are a subset of the classes available in output from the network that are in turn a subset of the ANPs available in the ontology of the chosen language.

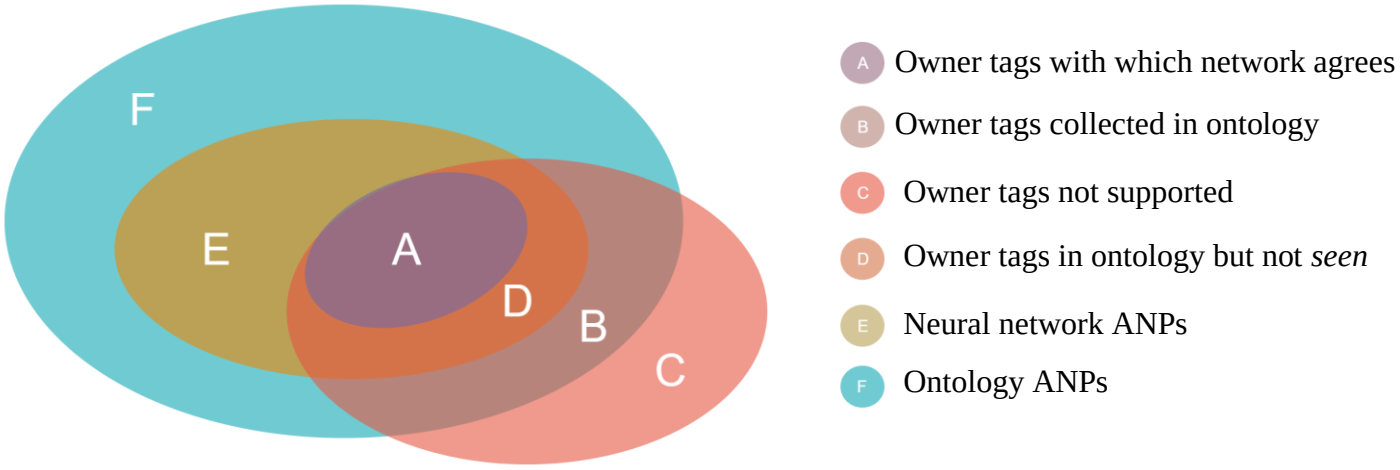

Figure 10 - Graphical representation of the owner tags analysis

The descriptiveness of the user is evaluated with this formula:

$$\frac{\alpha * A + \beta * B + \gamma * C}{A + B + C}$$

where:

- A is the intersection of the owner tags and the image tags
- B is the set of the owner tags not recognized by the network
- C is the set of the owner tags that the network will never recognize because these tags are not collected in the ontology on which the neural network is based.

The value of α, β and γ are chosen in an empirical way to give the opportune weight to the different sets, so:

- $\alpha = 1$
- $\beta = 0.5$
- $\gamma = 0.8$

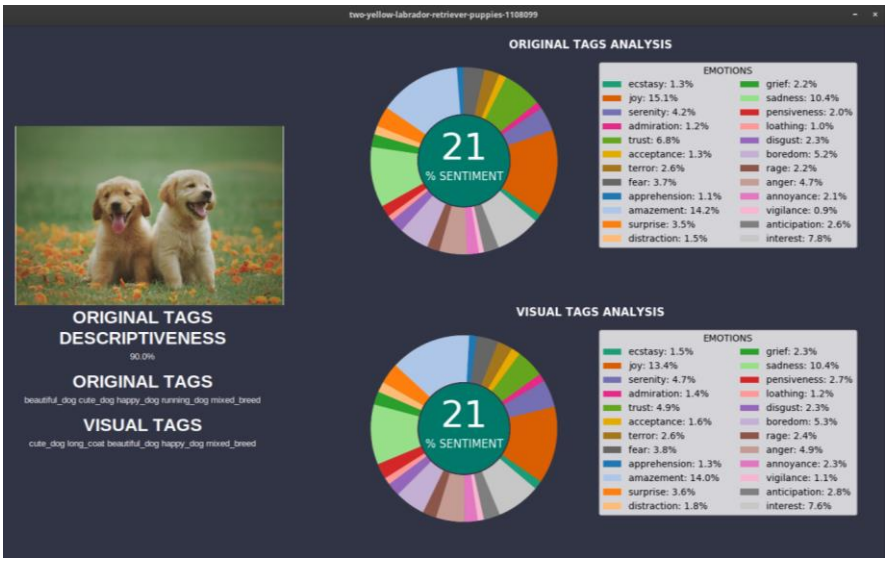

Figure 11 - Analysis of one of the images found searching the tag "happy_dog"

5. Conclusions and future development

Regarding the realized structure, it is of interest to verify the capacity of the network, trained on a finite number of ANPs, to express or not the same sentiment and the same emotions of a real user. If you want to automate the publication of social content, this could be a valid solution for the association of relevant tags. This functionality would remain bound only to the power of the model used for the association of the tags; therefore, maintaining the structure of the application developed unaltered, arbitrarily better results could be obtained, making appropriate corrections or improvements to the underlying network.

We can try to understand why determinate contents, uploaded by the users, are able to become really popular. Training ad-hoc algorithms on the emotions and sentiment transmitted by users, would allow to identify key elements of the images (whether they are objects present in them, or their own tags) in order to offer a valid alternative for the study of social networks, exploiting all **the power offered by the union of machine learning and semantic technologies**.

The opportunity may arise to find tags, images, emotions or other particular elements that have a significant weight in the world of the social networks, and in this way to create new techniques to study popular phenomena.

VisualFlickr fits very well with a lot of the objectives of the digital era. It can be used to create a successful advertising campaign, for example using *VisualFlickr* to analyze a popular TV spot and in this way understand why that advertising was so famous.

But these ones are only examples because every field in which the human emotions associated to Web content are important, *VisualFlickr* could be useful adapting its functionalities to perform different tasks. A sample could be the application to forensic field in which the detection of extreme emotions in the Web is still considered an open problem.

5.1 Future developments

In the future of *VisualFlickr*, its developers certainly see improvements in the routines dedicated to the graphical interface and the command line interface for selecting operations. In this case it is planned to integrate the selection of operations within the existing graphical interface, in a lightweight and intuitive way.

Continuous tests are envisaged on any anomalies relating to the artificial vision sector so that the application can become stable and fully functional.

Finally, some new features can be implemented:

- **Deep textual analysis** of the tags inserted by the image owner so that they can be traced back to ANPs collected in ontology through synonyms, lemmatization or extrapolation of semantics
- Calculation of the feeling and emotion of objects recognized by artificial vision in an image in a **manner dependent on the classification score** of objects obtained from the neural network; in this way it is possible to calculate also the data of the descriptive capacity of the tags inserted by the image owner in accordance with the reliability of the visual analysis;
- Currently it is possible to start the analysis algorithm through a single language chosen from among those on which neural networks have been trained (English, Italian, Spanish, French, German and Chinese): the chosen language is used both for visual analysis both for the textual analysis of the tags inserted by the owner of the images; it is planned to **extend the multilingual support** in a way that faithfully translates the Arabic-Chinese, French, English, Italian, Dutch, Persian, Polish, Russian, Spanish, German, and Turkish adjective-noun pairs into the language chosen from the six listed above; in this way the emotions and the feeling will be the products of the culture of the selected language but it will be possible to analyze a set of multilingual tags inserted by the image owner;
- The analysis of the descriptive capacity of the tags inserted to descript the images on Flickr can be considered one of the multiple extensions of the application that lends itself to accommodate many other features related to the captioning of the images.

6. Run *VisualFlickr*

If you want to run the code of *VisualFlickr* you can download all the project from our repository:

https://github.com/gdiprisco/VisualFlickr

Please, follow the instructions contained in the **README.md** file before testing the project.

7. Figures