Less is MORE: a MultimOdal system for tag REfinement

Lucia C. Passaro[1] and Alessandro Lenci[1]

CoLing Lab, Dipartimento di Filologia, Letteratura e Linguistica (FiLeLi), Università di Pisa, via Santa Maria 36, 56126 PISA, Italy
http://colinglab.humnet.unipi.it
lucia.passaro@fileli.unipi.it, alessandro.lenci@unipi.it

Abstract. With the proliferation of image-based social media, an extremely large amount of multimodal data is being produced. Very often image contents are published together with a set of user defined metadata such as tags and textual descriptions. Despite being very useful to enhance traditional image retrieval, user defined tags on social media have been proven to be noneffective to index images because they are influenced by personal experiences of the owners as well as their will of promoting the published contents. To be analyzed and indexed, multimodal data require algorithms able to jointly deal with textual and visual data. This research presents a multimodal approach to the problem of tag refinement, which consists in separating the relevant descriptors (tags) of images from noisy ones. The proposed method exploits both Natural Language Processing (NLP) and Computer Vision (CV) techniques based on deep learning to find a match between the textual information and visual content of social media posts. Textual semantic features are represented with (multilingual) word embeddings, while visual ones are obtained with image classification. The proposed system is evaluated on a manually annotated Italian dataset extracted from Instagram achieving 68% of weighted F1-score.

Keywords: Natural Language Processing · Computer Vision · Multimodal Semantics

1 Introduction

Human communication is intrinsically multimodal. Ideas can be better expressed and understood by jointly using different modalities, as proven by most advertising campaigns and social media, in which the skillful combination of images and language is able to amplify communicative intents and impact. With the evergrowing expansion of Internet-based activities in the last 10 years, an extremely large amount of multimodal data is being produced. Multimodal information processing is therefore needed in order to make sense of such large quantities of data, enabling the development of systems that jointly deal with textual and visual data. Not only the automatic interpretation of texts can be improved by exploiting additional non-verbal information such as visual contents, but the interpretation of images can also be enriched by exploiting the meaning of their surrounding text [4]. Examples of applications based on multimodal information processing are image research from textual descriptions and, vice versa, the generation of textual descriptors or captions from an image [16]. In this paper, we focus on the refinement of user defined image annotations, namely the tags provided with images when they are published on social media such as Instagram. In this platform, owners share pictures annotated with a set of tags based on personal experiences. Giannoulakis and Tsapatsoulis [14] demonstrated that only 66% of human defined tags describe the visual content of the image. This negatively affects the way we can access and use Instagram data.

Social media tags are useful to enhance traditional image retrieval technology [27,33], but they usually include a lot of noise. For instance, approximately only 20% of the Instagram hashtag datasets are appropriate to be used as training examples (i.e., image - tag pairs) for image recognition machine learning algorithms [14]. User defined tags suffer from ambiguity, carelessness and incompleteness [22], and have been proven to be highly associated with trends and events occurring in the real world, and biased toward personal perspectives [15,30]. Moreover, very often people tag objects and scenes that are not present in the visual content in order to favor image retrieval for the general audience [22]. The present work aims at separating relevant from noisy tags in order to guarantee a better indexing and retrieval of images. We approach the problem by assigning a relevance value to tags. In order to do so, we employ a mixture of Natural Language Processing (NLP) and Computer Vision (CV) techniques and resources. This allows us to take into account both textual and visual features of multimodal contents in a combined and synergic way.

This paper is structured as follows: Section 2 shows a brief overview of existing works in this field. Section 3 describes the proposed approach to tag refinement. Section 4 presents the system architecture including both the NLP and CV modules. Sections 5 and 6 report on the collection of the dataset used for the evaluation, namely a set of Italian manually annotated Instagram posts, and on the evaluation itself. Section 7 discusses the performances achieved by the current system implementation and Section 8 is left for conclusions and future research.

2 Related work

As suggested by [22], existing research on image tags focuses on three main tasks, namely *tag assignment*, *tag refinement* and *tag retrieval*. In tag assignment, given an unlabeled image, the goal consists in assigning a (fixed) number of tags related to the image content [26,34,35]. In tag refinement, given an image associated with some initial tags, the objective is to separate irrelevant tags from relevant ones

[24,37,23,38]. Finally, given a tag and a collection of images, tag retrieval focuses on retrieving relevant images with respect to the tag of interest [10,13,36].

In this work, we address the tag refinement task, and we aim at separating relevant tags from noisy ones in user-tagged Instagram images. Previous studies tackled tag refinement by considering several perspectives including the use of linguistic information only, by measuring tag relevance in terms of weighted semantic similarity between the target tag and the other ones [31] associated to the image, or by integrating textual data with the visual content [21]. A different approach exploited Principal Component Analysis [5] to build a model based on the factorization of an image-tag matrix by a low-rank decomposition with error sparsity [39]. A systematic evaluation and comparison of various tag refinement models was carried out in [22]. In order to guarantee a reliable comparison, the authors implemented several methods by exploiting the same models to process the textual and the visual content. We suggest the interested reader to refer to this survey for details on model implementation and results.

This evaluation shows that the best performing model for tag refinement is the one based on robust PCA [5] and a CNN [18] to process the image content, achieving performances between 0.57 and 0.63 depending on the size of the training set. However, the datasets employed in their evaluations present several elements that do not comply with the purpose of the present work. First, both NUS-WIDE [7] and MirFlickr [17] have been created several years ago, and since then the way to publish multimodal contents has changed considerably. While in Flickr-based datasets the content of tags is almost exclusively referential, in the last few years, and with the advent of social media such as Instagram, tags are not used only as referential descriptors, but also for several other purposes, such as for instance expressing emotions [14]. Flickr users were mostly interested in promoting their images specifically for their content. On the contrary, Instagram users tend favour interactions with other users, and are therefore inclined to produce tags that make their post accessible to the widest possible audience, regardless of the content of the image. Moreover, such datasets contain limited English dictionaries of tags that do not allow for a multilingual evaluation, which is crucial in the present work. To the best of our knowledge, no previous work addressed tag refinement by exploiting computer vision along with mulilingual distributional semantics. For these reasons, we decided to create a new evaluation set from public photos on Instagram.

3 Proposed approach

As anticipated, we approach the tag refinement task by integrating NLP and CV techniques. Our system MORE (MultimOdal Tag Refinemen) is aimed at separating relevant tags from noisy ones in user-tagged Instagram images. In particular, it exploits several resources in order to compute the tag relevance. We designed its architecture to address several issues.

First, MORE is expected to work on Italian Instagram posts. On this social network, users publish their photos accompanied by a list of tags (with a maximum of 30 per photo). For Italian users, these tags are often both in Italian and English, with the result of increasing the number of tags and adding noise and redundancy. In Computational Linguistics, the last years have witnessed the growing use of *word embeddings*, that are dense distributional vectors typically built with neural language models to represent the meaning of words [20]. Usually, these embeddings are monolingual, but given the nature of Instagram and the behaviour of its Italian users, we need to represent the meaning of the words not only in Italian but also in a cross-lingual way. Multilingual embedding models have been proposed in recent literature. They are able to capture words and their translations across languages in a joint embedding space [29]. This kind of embeddings is very appealing when considering applications in social media like Instagram since they allow to deal both with Italian and English tags.

The second challenge we need to address is separating relevant from noisy tags with respect to the image content as well. We define the relevance of a tag in terms of the consistency between its *denotational meaning* and the visual content of an image (e.g., objects and scenarios appearing in it). For example, the tag *boat* is relevant for an image containing a seascape with a ship. To this purpose, we use a framework to convert an image into a series of textual descriptors for all the elements contained in it. This kind of descriptors are extracted with one of the most popular Convolutional Neural Network architectures for image labeling, namely the VGG-16 Neural Network. More specifically, we used the pretrained VGGNet [32].

The third challenge is related to the prospective application context of MORE. Indeed, the system is expected to reduce the implicit noise of Instagram automatically crawled datasets. This feature is very important in order to support other industrial applications such as market surveys and sentiment or brand reputation analysis. Despite being very popular for Business Intelligence, such applications often suffer from the noise generated by user defined tags. In fact, these tags are often carefully chosen to favor image retrieval and visibilty. Therefore they require some form of preprocessing to guarantee the reliability of the aggregated results. For example, the image in Figure 1 was published with several English and Italian tags. Some describe the image content (e.g. *fiori* ('flowers'), *occhiali* ('glasses'), *orchidea* ('orchid'), *flower*), while others are clearly used to promote the image retrieval on the Instagram platform (e.g. *moda* ('fashion'), *buongiorno* ('good morning'), *Roma* ('Rome'), *outfit*).

In order to separate relevant from noisy tags, MORE exploits three main resources to compute tag relevance:

VGG-16: To establish the relevance of a tag for a given image, the pretrained VGG16 network [32] is used. In particular, we adopted the version available in Keras [6] and trained on ImageNet [9]. The ImageNet dataset [9] consists of 1.4 million images, each labelled with one of the $1,000$ different classes.

MUSE: The system compares English and Italian word vectors in a multilingual space. To build such space the Multilingual Unsupervised and Supervised Embeddings (MUSE) framework [8,19] has been used to align in a single vector space the pretrained version of the Italian and English fastText

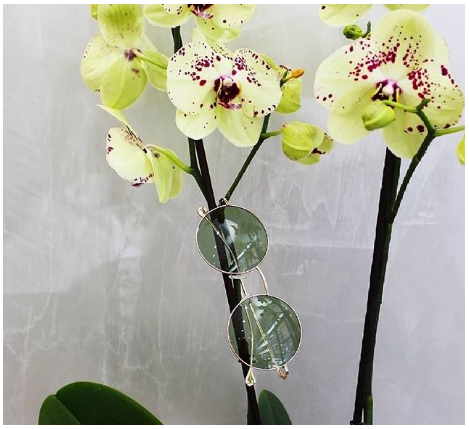


Fig. 1. A photo posted on Instagram with its Italian and English relevant and noisy tags. Original tags were: *beauty, buongiorno, colors, fiori, fiorieocchiali, flower, flowers, h52, instabeauty, instadaily, instafashion, moda, occhiali, occhialidasole, orchidea, outfit, repost, roma, sunglasses.*

Wikipedia word embeddings [1]. In particular, the Italian space is used as source space and the English one is used as target.

OMW: Multilingual WordNet synsets were used to translate the ImageNet classes and to obtain their hypernyms. Specifically, the version of Open Multilingual Wordnet (OMW) [3,2] available in NLTK [25] was used to make joint queries on the English [11] and the Italian [28] models.

4 The MORE Architecture

Given a set of Instagram posts consisting of an image and a list of tags, the goal of the MORE system is to distinguish relevant from irrelevant tags. The architecture of MORE is shown in Figure 2.

Our dataset D is formally defined as $D = \{P_1, ..., P_k\}$. Each element P_i is the pair $P = (T, p)$, where T is a list $\{t_1, ..., t_n\}$ of tags including both relevant and irrelevant ones, and p is the visual content. For each $P_i \in D$, MORE carries out the following process:

Image classification: MORE classifies p with the VGG-16 model and returns a list of $L = \{l_1, ..., l_n\}$ English labels belonging to the ImageNet $1,000$ classes. A parameter specifies the probability threshold of the output classes (by default, the system returns all non-zero values). Therefore, this step provides a list of potential labels associated with the image and referring to its content. For example, given the photo in Figure 3, the output of the classification process is as follows: CASTLE, MONASTERY, PALACE, BELL_COTE, CHURCH.

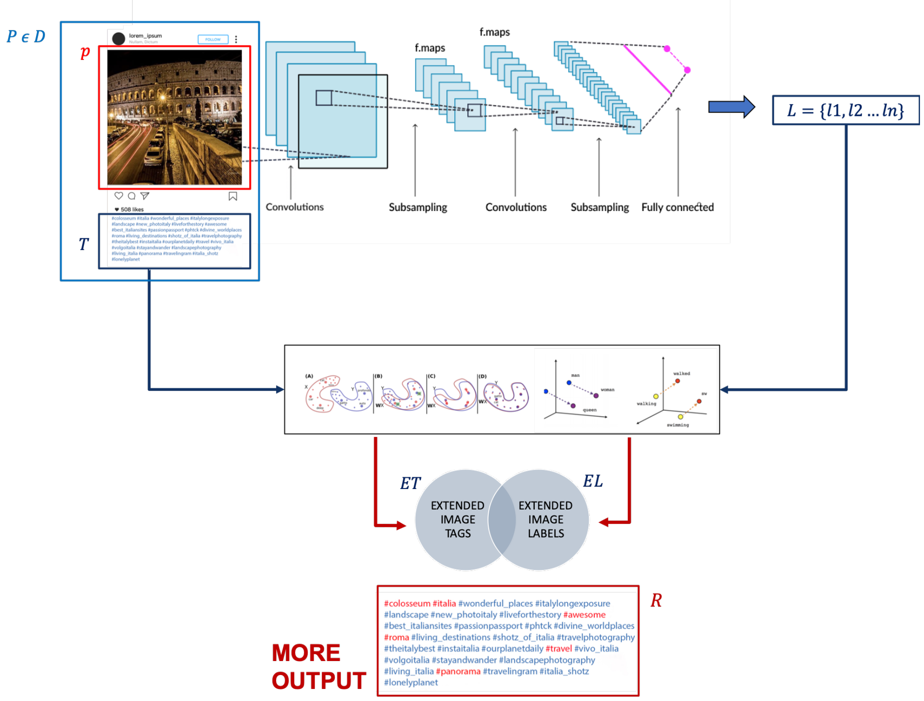

P ϵ D
p
T
Convolutions
f.maps
Subsampling
f.maps
Convolutions
Subsampling
Fully connected
L = {l1, l2 … ln}
ET
EL
EXTENDED IMAGE TAGS
EXTENDED IMAGE LABELS
R
#colosseum #italia #wonderful_places #italylongexposure
#landscape #new_photoitaly #liveforthestory #awesome
#best_italiansites #passionpassport #phtck #divine_worldplaces
#roma #living_destinations #shotz_of_italia #travelphotography
#theitalybest #instaitalia #ourplanetdaily #travel #vivo_italia
#volgoitalia #stayandwander #landscapephotography
#living_italia #panorama #travelingram #italia_shotz
#lonelyplanet
MORE OUTPUT

Fig. 2. The MORE architecture for refining the tags of an Instagram photo.

Labels translation and extension: The system translates into Italian the English labels in L by exploiting both WordNet senses (OMW) and multilingual embeddings (MUSE). For each label in L, the system retrieves from OMW all the Italian lemmas marked with the same synset of the English label. Moreover, the system extracts also the hypernyms of each English label (e.g., *cat* and *feline* from EGYPTIAN_CAT). The output of this step for each image is a list of Italian translated labels $LT = \{lt_1, ..., lt_n\}$. For the example provided in Figure 3, the list of Italian translated labels (LT) includes: CONVENTO ('monastery'), MONASTERO ('monastery'), PALAZZO ('palace'), CHIESA ('church'). The list of the extended labels EL is defined as $L \cup LT$ (i.e., the English predicted labels and their Italian translation. The EL of the previous example contains: CHURCH, MONASTERY, MONASTERO ('monastery'), CONVENTO ('monastery'), PALAZZO ('palace'), BELL_COTE, CHIESA ('church'), CASTLE, PALACE, as well as their hypernyms in Italian and English (e. g. ABITAZIONE ('dwelling'), RELIGIOUS_RESIDENCE).

Multilingual neighborhood: For each tag t belonging to T, the system carries out a query on the MUSE multilingual embeddings models to collect the top x nearest neighbors of each element both in Italian and English. In particular, we use the Italian space as source space and the English one as target

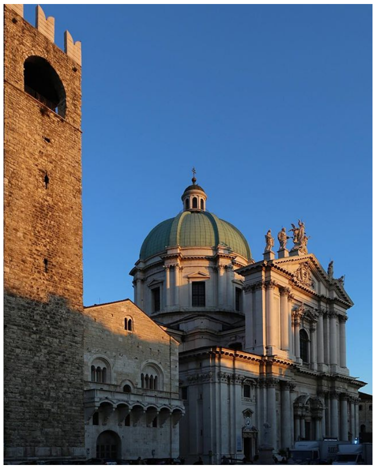


Fig. 3. An example from the Instagram dataset. Original tags for this photo were: *brescia, cattedrale, landscapephotography, monument, fotografia, pics, world, prospettiva, brixia, photo, art, city, foto, picsart, landscape, picoftheday, arte, lombardia, italia, cultura, photography, italy, architecture.*

space, in order to populate the list of extended tags (ET). By default, the parameter x is set to 5. For example, given the tag *cattedrale* ('cathedral'), its Italian nearest neighbors are *cattedrale* ('cathedral'), *procattedrale* ('pro-cathedral'), *concattedrale* ('co-cathedral'), *cathedrale* ('cathedral'), *basilica* ('basilica') while its English ones are: *cathedral*, *basilica*, *cathedra*, *cathedrals*, *church*.

Filtering: The system filters the elements in T by considering ET and EL. In particular, for each of the tags $t \in T$, if $t \in EL$ then it is added to the set R of relevant tags. Otherwise, t is added to R in three cases: (i) one of its nearest neighbors in ET belong to EL; (ii) the vector of t is similar to at least one of the vectors of EL; (iii) the vector of at least one of the neighbors of t in ET is similar to any of the vectors of EL. Similarity is computed with cosine. By default the threshold on cosine is set to 0.4. For example, the tag *cattedrale* ('cathedral') in Figure 3, was found to be similar to most of the labels (e.g., *chiesa* ('church') with a cosine similarity of 0.74). Since this is an instance of case (ii), it is added to R. In fact, even if the tag itself was not similar to any of the labels, one of its neighbors in the English space is *church*. Since this word belongs to EL, the tag would be marked as relevant anyway (case (i)). Moreover, we can consider the case of the tag *architecture*. It has a cosine similarity below the threshold for all of the elements in EL. Nonetheless, at least one of its neighbors, (i.e., *architectural*), has a cosine similarity above the specified threshold for at least one of the labels, in particular with the label *church* with a similarity of 0.42. Therefore, according to the (iii) scenario, *architecture* was added to the relevant tags R. Conversely, neither the tag *brescia* (‘Brescia’, Italian city) nor any of its neighbor are found to be similar to any of the labels (e.g., the cosine with *monastery* is 0.27 in English and 0.13 in Italian). Thus, the tag is not considered relevant by the system, despite being in fact correct. The default MORE configuration marks as relevant only the following tags among the ones coming with the image (cf. Figure 3): *architecture*, *art*, *arte* (‘art’), *cattedrale* (‘cathedral’), *cathedral*, *city*, *cultura* (‘culture’), *foto* (‘photo’), *fotografia* (‘photography’),*italia* (Italy), *italy*, *landscape*, *monument*, *photo*, *photography*, *pics*, *prospettiva* (‘perspective’), *world*.

5 Dataset description and annotation

MORE has been evaluated on a portion of the dataset collected in the context of the MUSE project, aimed at performing a multimodal analysis of both texts and images in order to improve the quality of sentiment analysis and brand reputation systems. The whole MUSE dataset has been collected between May 2018 and December 2018 by using the official Instagram API.[1]

A total of 14 hashtags were used for data collection. Seven of these were closely related to a customer company of the MUSE industrial partner while the rest are very generic and not related to any particular topic. We don’t report the list of the first group of hashtags. The list of the second group is as follows: *follow4follow, igers, followme, instago, italia, buongiorno, instaitalia.*

Overall, the dataset consists of more than $200k$ images and the effectiveness of the system has been evaluated by the company internally. However, a portion of randomly selected 50 images was manually annotated for tag relevance by 7 human annotators.

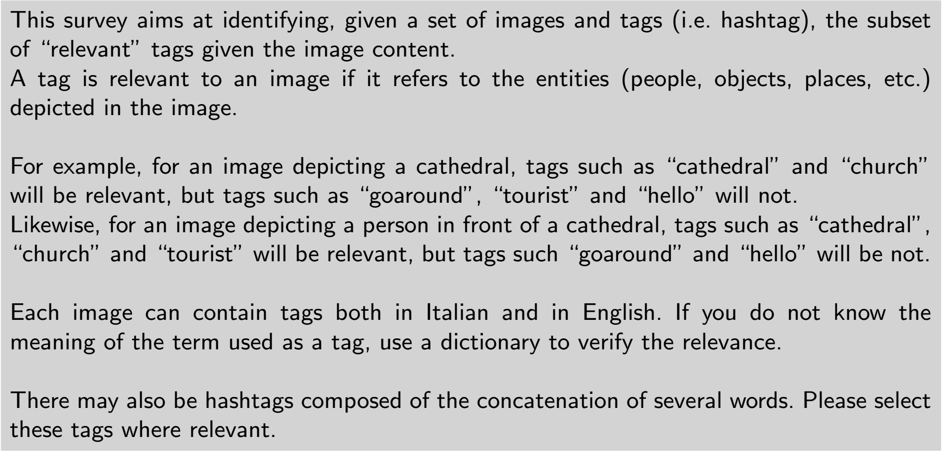
The participants to a questionnaire rated 50 images with respect to several human provided Instagram hashtags. Participants were asked to find the “relevant” hashtags with respect to the image content. The English translation of the instructions is reported in Figure 4. Each image was presented to annotators together with a multi-selection button showing the list of the original hashtags (see Figure 5).

The annotation task was quite difficult, since the annotators had to select, for each image, one, some or no tags with potential different degrees of relevance. The number of tags for each image was variable and depended on the actual tags obtained when the post was collected.

Fleiss’ kappa [12] was used to compute the inter-annotator agreement on the 1195 data points consisting of image-tag pairs. The overall agreement was of 0.42, but an ablation experiment on the raters demonstrated that a global agreement of 0.58 could be reached with 6 out of the 7 raters.

Given such agreement, we based the final decision about the relevance of each tag on the majority vote criterion (4 votes), as shown in Figure 6.

[1] After Dec. 12 2018 Instagram API changed radically to comply with new GDPR regulations `https://www.instagram.com/developer/changelog/` and the collection of the dataset was stopped.



This survey aims at identifying, given a set of images and tags (i.e. hashtag), the subset of "relevant" tags given the image content.
A tag is relevant to an image if it refers to the entities (people, objects, places, etc.) depicted in the image.

For example, for an image depicting a cathedral, tags such as "cathedral" and "church" will be relevant, but tags such as "goaround", "tourist" and "hello" will not.
Likewise, for an image depicting a person in front of a cathedral, tags such as "cathedral", "church" and "tourist" will be relevant, but tags such "goaround" and "hello" will be not.

Each image can contain tags both in Italian and in English. If you do not know the meaning of the term used as a tag, use a dictionary to verify the relevance.

There may also be hashtags composed of the concatenation of several words. Please select these tags where relevant.

Fig. 4. Annotation Instructions (translated from Italian).

6 Evaluation

In the final evaluation dataset, the average number of tags associated with each item is 23.9 ($\sigma = 9.13$). Human ratings reveal that on average 3.6 tags ($\sigma = 2.9$) were actually relevant for a given image, while 20.3 ($\sigma = 9.5$) were not. The distribution of relevant vs. noisy user-defined hashtags is in line with the findings illustrated in [14], since approximately only 18% of tags are actually relevant.

For the evaluation, each image-tag pair was considered as independent from the others. In other words, each image is represented in the final dataset by a number of data-points equal to the number of its original hashtags. Note that the set of relevant tags is always a subset of the original tags.

Since the annotators were asked to mark as correct the tags referring to the objects visible in the images, we used the output of the image classification step as a first baseline for the task. The results are reported in Table 1.

Class	P	R	F1	Support
False	**0.85**	1.00	0.92	1014
True	0.83	**0.03**	0.05	181
Macro avg	0.84	0.51	0.49	1195
Weighted avg	0.85	0.85	0.79	1195

Table 1. Baseline based on the image classification (VGG-16) output.

The baseline has a clear issue in terms of flexibility. The baseline classifier, which is simply the VGG-16 classifier trained on ImageNet, may never predict certain tags, as it is limited by the number of classes it was trained on, and may choose to give more weight to certain aspects of the image. For example, if we consider the image in Figure 1, one of the correct tags is *flower*. However, if we

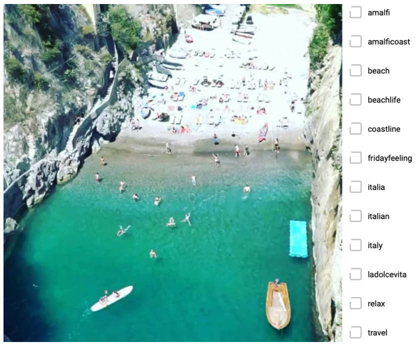


Fig. 5. An item provided during the annotation process. Users were asked to select none, one or more tags depending on their relevance with respect to the image.

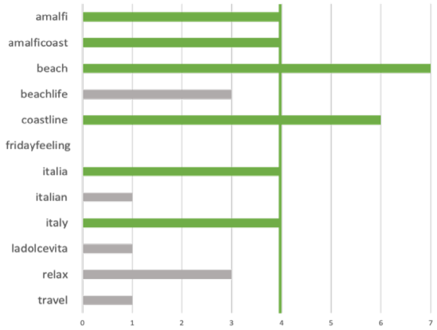


Fig. 6. Number of ratings for the hashtags associated to the image in Figure 5. A tag is associated to a given image if at least 4 annotators marked it as relevant.

look at the output of the classification step, the words *pot* and *vase* are included, while *flower* itself is not predicted. This may be possibly due to the fact that what the classifier is trained to see is actually a flower pot, but not just a flower. Therefore in this case, *flower* is actually considered as a False Negative example.

7 Results and Discussion

The overall performances of the system were assessed by comparing the model predictions against human rating. Table 2 shows the results of the evaluation in terms of Precision, Recall, F1-score and Support.

Class	P	R	F1	Support
False	**0.92**	0.61	0.73	1014
True	0.24	**0.71**	0.36	181
Macro avg	0.58	0.66	0.55	1195
Weighted avg	0.82	0.62	**0.68**	1195

Table 2. Precision (P), Recall (R) and F1-Score (F1) and Support of MORE based on a dataset of 50 Instagram images with the default configuration.

True Positives (TP) are the image-tag pairs for which both humans and the system associated the class True (relevant). At the same way, True Negatives (TN) are data points for which both the system and humans associated the class False (noisy). False Positives (FP) are predicted as relevant by the system, but rated as noisy by humans. Finally, False Negatives (FN) are the examples rated as relevant by humans but marked as noisy by the system.

Given the distribution of relevant tags in our dataset (18% of the total), most of the data-points belong to the False class, affecting the macro-averaged result. Therefore, we prefer to consider the weighted average as evaluation metric: Precision, Recall and F1-score are computed for each class, and their average is weighted by support (i.e., the number of true instances for each class).

Overall, the system reaches a weighted average F1-score of 0.68, which is distributed differently across the relevant and noisy hashtags class. Moreover, it is important to stress that in the case of irrelevant tags (the FALSE class) it is very important to maximize the Precision in order to avoid noise. On the contrary, for relevant tags (TRUE class) we are particularly interested in maximizing the Recall, in order to guarantee a satisfactory retrieval of relevant images.

In order to pursue such goals, we decided to use, along with standard metrics, an additional one consisting of the average between the Precision of the class FALSE and the Recall of the TRUE one. This metric, in fact, provides us with a useful method to evaluate if the system is able to discard noisy tags and, at the same time, to preserve the correct ones. If we look only at the Precision and Recall of the individual classes, we would not be able to capture this information. This metric, calculated with the MORE default parameters, is 0.815, outperforming the baseline (for which it was 0.44) by a wide margin, despite its weighted average F1-score was of 0.79. Even though we consider such results as promising, we performed a manual evaluation of error types to identify the most challenging cases for MORE.

One of the problems we detected consists in the recall of multiword hashtags such as *sprayart* ('spray art'), *biancoenero* ('black and white' concatenating the words "bianco e nero"), *fotodaltreno* ('photos from the train', from "foto dal treno"), fiorieocchiali ('flowers and glasses', from "fiori e occhiali"), creativemakeup ('creative makeup').

As for False Negative examples, 37 of the 53 total examples (70%) were out-of-vocabulary (OOV) in both the source and the target word space. There is surely wide room for improvement. For example, the MORE match function, could be enriched with the ability of segmenting multiword hashtags in standard lexical entries or by exploiting fastText features to extract the vectors of out-of-vocabulary words also in a multilingual space.

As for False Positive examples, we noticed that MORE is more prone to predict as relevant tags high-frequency words such as *photo, nice, look, selfie, style, pizza.* In addition, we noticed that MORE tends to mark abstract words (in Italian and English) as relevant tags. Despite being an error, we can consider it expected due to the way in which MORE has been constructed. In fact, very often the vectors of abstract words are highly associated with referential objects depicted in the photos. For example, words like *mood, enjoy, happy, verità* ('truth'), *bellezza* ('beauty'), *parole* ('words'), have been considered False by annotators because, according to the provided instructions, they "do not refer to the entities (people, objects, places, etc.) depicted in the image". Nonetheless, given the high association of such word vectors with referential objects, MORE often tags them as relevant.

We also performed further experiments to understand how the various parameters affect performances and can be exploited to reach different goals. It is clear that such parameters are very important to leverage the behavior of the application. For example, a higher cosine similarity threshold eventually discards many tags. This is useful if the final goal is to refine tags as accurately as possible. On the other hand, this setting will negatively affect the recall of relevant hashtags. The same consideration goes for increasing the nearest neighbors of tags and labels. We performed a parameter tuning experiment in which we evaluated MORE with a cosine similarity threshold of 0.3, 0.4 and 0.5. Similarly, we assessed the system by changing the number of nearest neighbors. In this case, the evaluation was performed with 3, 5 and 10 nearest neighbors. In order to choose the best performing configuration, we selected the one maximizing the average between the Precision of the class False (not relevant) and the Recall of the True (relevant) one. For this metric, MORE performs at best 0.87. The results of this model are reported in Table 3.

By considering the standard metrics for this model, we notice that also the Weighted AVG Precision improves after the parameter tuning. Weighted AVG Recall (and thus F1-score), on the contrary, is negatively affected by the results on False, which is the majority class.

Class	P	R	F1	Support
False	**0.94**	0.55	0.70	1014
True	0.24	**0.80**	0.37	181
Macro avg	0.59	0.68	0.53	1195
Weighted avg	0.83	0.59	**0.65**	1195

Table 3. Precision (P), Recall (R) and F1-Score (F1) and Support of MORE based on a dataset of 50 Instagram images after the parameter tuning (Evaluation performed with cosine similarity threshold set ti 0.3 and nearest neighbors set to 10).

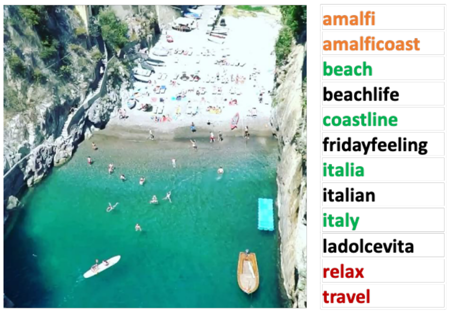


Fig. 7. MORE predictions against human annotation with the default configuration (cosine sim. threshold: 0.4; nearest neighbors: 5).

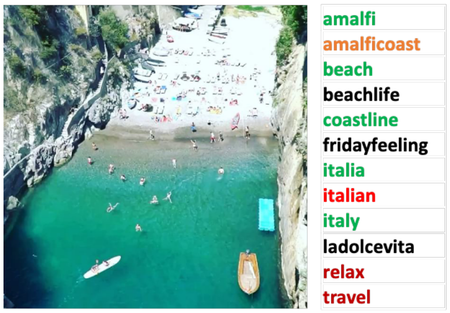


Fig. 8. MORE predictions against human annotation after parameter tuning (cosine sim. threshold: 0.3; nearest neighbors: 10).

In order to further improve the overall performances, a viable option could be to leverage the frequency and the popularity of the tags on Instagram. Such tags, in fact, represent a lot of false positives (e.g. high frequency words), and decreasing their number could increase also the Weighted AVG F1-score.

Figures 7 and 8 show the predictions of both the default and the tuned configuration compared to human annotation. Green tags are true positive examples, red tags are false positives (the system considered them as relevant while humans did not), orange tags are the false negative. We can see that abstract words are false positives in both the models. This means that the thresholds are not able to properly mitigate such phenomenon because words like beach, travel and relax are strongly associated one another (e.g. to remove the tags *relax* and *travel* the system requires at least a threshold of cosine of 0.5, with the effect of removing also true positive tags).

8 Conclusions

In this paper, we presented the MultimOdal Tag Refinement (MORE), a system aimed at improving image descriptors by exploiting NLP and CV techniques. The system starts from Instagram user defined image annotation, and merges visual and textual information to find a match between the tags provided with an image and its semantic visual content. Textual features have been extracted from text based tags by exploiting the (multilingual) word embeddings. Visual features have been gathered by exploiting image classification. The system has been evaluated on an Italian manually annotated dataset achieving 68% of performances in terms of weighted F1-score.

The results of MORE are promising, but there are still wide margins of improvement for several key aspects of the system, including: (i) the construction of multilingual embeddings; (ii) the management of multilingual hashtags; (iii) the refinement and extension of the evaluation process; (iv) the distribution of the manually annotated dataset. In the near future, our efforts will be focused towards these directions. As for (i), we aim at training embeddings on a mixture of social media and general purpose corpora. This combination is expected to positively affect (ii), as it would enable the collection of reliable vectors for multiword hashtags, thus reducing the number of OOV words during filtering. As for (iii), we plan to fine-tune all the modules including the neural network for image classification, and to study the contribution of each module of the architecture to the classification. Finally, for (iv), we plan to extend the manually annotated dataset to improve the evaluation and to make it available for research purposes, in accordance to the Instagram Privacy constraints.

Acknowledgments

This research has been supported by the Project MUltimodal Semantic Extraction (MUSE), in collaboration with Bnova s.r.l., funded by Regione Toscana with the grant POR FSE 2014-2020 Asse A.

References

1. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606 (2016)
2. Bond, F., Foster, R.: Linking and extending an open multilingual wordnet. Sofia (2013)
3. Bond, F., Paik, K.: A survey of wordnets and their licenses. In: Proceedings of the 6th Global WordNet Conference (GWC 2012). Matsue (2012), 64–71
4. Bruni, E., Tran, G.B., Baroni, M.: Distributional semantics from text and images. In: Proceedings of the GEMS 2011 workshop on geometrical models of natural language semantics. pp. 22–32 (2011)
5. Candès, E.J., Li, X., Ma, Y., Wright, J.: Robust principal component analysis? Journal of the ACM (JACM) **58**(3), 11 (2011)
6. Chollet, F.: Deep learning with python (2017)
7. Chua, T.S., Tang, J., Hong, R., Li, H., Luo, Z., Zheng, Y.: Nus-wide: a real-world web image database from national university of singapore. In: Proceedings of the ACM international conference on image and video retrieval. pp. 1–9 (2009)
8. Conneau, A., Lample, G., Ranzato, M., Denoyer, L., Jégou, H.: Word translation without parallel data. arXiv preprint arXiv:1710.04087 (2017)
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: CVPR09 (2009)
10. Duan, L., Li, W., Tsang, I.W.H., Xu, D.: Improving web image search by bag-based reranking. IEEE Transactions on Image Processing **20**(11), 3280–3290 (2011)
11. Fellbaum, C. (ed.): WordNet An Electronic Lexical Database. The MIT Press (1998)
12. Fleiss, J.: Measuring nominal scale agreement among many raters. Psychological bulletin **76**(5), 378—382 (November 1971)
13. Gao, Y., Wang, M., Zha, Z.J., Shen, J., Li, X., Wu, X.: Visual-textual joint relevance learning for tag-based social image search. IEEE Transactions on Image Processing **22**(1), 363–376 (2012)
14. Giannoulakis, S., Tsapatsoulis, N.: Evaluating the descriptive power of instagram hashtags. Journal of Innovation in Digital Ecosystems **3**(2), 114–129 (2016)
15. Golder, S., Huberman, B.: Usage patterns of collaborative tagging systems. J. Information Science **32**, 198–208 (04 2006)
16. Gong, Y., Ke, Q., Isard, M., Lazebnik, S.: A multi-view embedding space for modeling internet images, tags, and their semantics. International journal of computer vision **106**(2), 210–233 (2014)
17. Huiskes, M.J., Thomee, B., Lew, M.S.: New trends and ideas in visual concept detection: the mir flickr retrieval evaluation initiative. In: Proceedings of the international conference on Multimedia information retrieval. pp. 527–536 (2010)
18. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012)
19. Lample, G., Conneau, A., Ranzato, M., Denoyer, L., Jégou, H.: Word translation without parallel data. In: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings (2018)
20. Lenci, A.: Distributional Models of Word Meaning. Annual review of Linguistics **4**, 151–171 (2018)

21. Li, X., Snoek, C.G., Worring, M.: Learning social tag relevance by neighbor voting. IEEE Transactions on Multimedia **11**(7), 1310–1322 (2009)
22. Li, X., Uricchio, T., Ballan, L., Bertini, M., Snoek, C.G.M., Bimbo, A.D.: Socializing the semantic gap: A comparative survey on image tag assignment, refinement, and retrieval. ACM Comput. Surv. **49**(1), 14:1–14:39 (Jun 2016)
23. Li, X., Shen, B., Liu, B.D., Zhang, Y.J.: A locality sensitive low-rank model for image tag completion. IEEE Transactions on Multimedia **18**(3), 474–483 (2016)
24. Li, Z., Liu, J., Zhu, X., Liu, T., Lu, H.: Image annotation using multi-correlation probabilistic matrix factorization. In: Proceedings of the 18th ACM international conference on Multimedia. pp. 1187–1190. ACM (2010)
25. Loper, E., Bird, S.: Nltk: The natural language toolkit. In: Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1. pp. 63–70. ETMTNLP '02, Association for Computational Linguistics, Stroudsburg, PA, USA (2002)
26. Makadia, A., Pavlovic, V., Kumar, S.: Baselines for image annotation. International Journal of Computer Vision **90**(1), 88–105 (2010)
27. Nov, O., Naaman, M., Ye, C.: What drives content tagging: the case of photos on flickr. In: Proceedings of the SIGCHI conference on Human factors in computing systems. pp. 1097–1100. ACM (2008)
28. Pianta, E., Bentivogli, L., Girardi, C.: Multiwordnet: developing an aligned multilingual database. In: Proceedings of the First International Conference on Global WordNet. Mysore (India) (2002)
29. Ruder, S., Vulić, I., Søgaard, A.: A survey of cross-lingual word embedding models. arXiv preprint arXiv:1706.04902 (2017)
30. Sen, S., Lam, S.K., Rashid, A.M., Cosley, D., Frankowski, D., Osterhouse, J., Harper, F.M., Riedl, J.: Tagging, communities, vocabulary, evolution. In: Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work. pp. 181–190. ACM (2006)
31. Sigurbjörnsson, B., Van Zwol, R.: Flickr tag recommendation based on collective knowledge. In: Proceedings of the 17th international conference on World Wide Web. pp. 327–336. ACM (2008)
32. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
33. Sun, A., Bhowmick, S.S., Chong, J.A.: Social image tag recommendation by concept matching. In: Proceedings of the 19th ACM international conference on Multimedia. pp. 1181–1184. ACM (2011)
34. Tang, J., Hong, R., Yan, S., Chua, T.S., Qi, G.J., Jain, R.: Image annotation by k nn-sparse graph-based label propagation over noisily tagged web images. ACM Transactions on Intelligent Systems and Technology (TIST) **2**(2), 14 (2011)
35. Tang, J., Shu, X., Li, Z., Jiang, Y.G., Tian, Q.: Social anchor-unit graph regularized tensor completion for large-scale image retagging. IEEE transactions on pattern analysis and machine intelligence (2019)
36. Wang, Y., Zhu, L., Qian, X., Han, J.: Joint hypergraph learning for tag-based image retrieval. IEEE Transactions on Image Processing **27**(9), 4437–4451 (2018)
37. Wu, L., Jin, R., Jain, A.K.: Tag completion for image retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(3), 716–727 (2012)
38. Xu, X., He, L., Lu, H., Shimada, A., Taniguchi, R.I.: Non-linear matrix completion for social image tagging. IEEE Access **5**, 6688–6696 (2016)
39. Zhu, G., Yan, S., Ma, Y.: Image tag refinement towards low-rank, content-tag prior and error sparsity. In: Proceedings of the 18th ACM international conference on Multimedia. pp. 461–470. ACM (2010)