A proverb is worth a thousand words: Learning to associate images with proverbs

Gözde Özbal[†], **Daniele Pighin**[‡], **Carlo Strapparava**[†]
[†] FBK-Irst - Trento, Italy
[‡] Google Zurich, Switzerland
gozbalde@gmail.com, biondo@google.com, strappa@fbk.eu

Abstract

We describe a system that can associate images with English proverbs. We start from a corpus of proverbs, harvest related images from the web and use this data to train two variants of a convolutional neural network. We then collect a small set of annotations, and use these to combine the outputs of the two networks into a single prediction for each input image. We carry out feature selection experiments on a set of features derived from the images and from the predicted proverbs, and demonstrate that the metaphoricity of the proverbs plays a significant role in classification accuracy. An empirical evaluation with human raters confirms the system's ability to abstract from the raw bits in the images and to learn meaningful, non-trivial associations.

Introduction

Meaningful associations between visual information and short texts are a staple of effective and powerful communication. Instances of this form of communication can be found almost anywhere: on t-shirts, covers of books, records and magazines, social media posts, and ad campaigns, just to name a few. The empirical evidence, in agreement with our common sense and everyday experience, shows that meaningful image-text associations are very good predictors of the success of an online post (Hessel et al., 2017). To add to the value of an image, the caption must convey some information that is not already obvious. For example, consider two possible captions for the image in Figure 1. A purely descriptive caption like (a) is very accurate, but it does not add value to the image. By associating it with a proverb, a caption like (b) radically changes our perception of the image, from a collection of visual elements to an abstract representation of a familiar feeling (i.e., envy).

Recent advances in neural networks and computer vision have made it possible to generate high-quality descriptive captions such as (a) in Figure 1 automatically (Vinyals et al., 2017). Such captions are certainly remarkable from an artificial vision stand point, and very useful when it comes to organizing and accessing large databases of images. However, they do not make an image more memorable or compelling.

In this paper, we focus on the task of producing captions like (b), in which an image is associated with a memorable expression that emphasizes non trivial, suggestive aspects of the image. In particular, we leverage an existing corpus of English proverbs (Özbal et al., 2016) to learn a model that can associate any image to the most appropriate proverb in the repository. The resulting system can have many potential applications, e.g.: suggesting evocative and compelling taglines when posting an image on social media; proposing headlines for news, based on photos of an event; selecting the visual content of ad campaigns so as to evoke specific moods.

Figure 1: Different captions affect our perception of the same image: (a) "A half-barren, half-green field." (b) "The grass is always greener on the other side."

To the best of our knowledge, this is the first attempt to prove that existing models for object recognition can be successfully adapted to associate images to linguistically complex and semantically rich data such as proverbs. We demonstrate that the existing networks have enough capacity to abstract away from the mere graphical content of an image and learn original and surprising associations.

Note that we do not claim that our model can understand the language used in the proverbs. This is a complex problem per se, given the non-literal nature of most proverbs. In addition, from the point of view of our model a proverb is just a class label. Instead, we observe that by using the proverbs to retrieve related images allows the model to learn that some combinations of objects appearing in the pictures are relevant with respect to the meanings commonly attached to the proverbs, also when their meaning is far from literal.

The approach that we propose is simple and scalable, it relies on the availability of large amounts of noisy data and can be tuned using minimal supervision.

Related work

A growing body of literature, including Yamaguchi et al. (2014) and Gelli et al. (2015), has shown that image features do not contribute as much as textual features to the social popularity of multimedia content. In particular, Hessel et al. (2017) study the effect of visual and textual features on the popularity of Internet posts, and conclude that the right combination of visual and textual features plays a very important role. They also note that the cleverness of the accompanying captions can result in a very different response to pictures of very similar subjects, and make a less attractive subject more

popular than a better subject with a less remarkable caption.

Concerning the automatic captioning of images, Hall et al. (2015) propose to automatically generate natural language captions that describe the geographical context of georeferenced photos, such as "Rijksmuseum photographed at 2.15 pm at the corner of Stadhouderskade and Museumstraat near Spiegelgracht in Amsterdam, Netherlands.". Chen et al. (2015) present a large dataset consisting of groups of images observed with the same caption. The associative structure of the data is exploited to retrieve captions for query images. The retrieved captions can be further classified to select the more creative ones. Vinyals et al. (2017) present a generative model based on a deep recurrent architecture that can generate natural sentences describing an image. The model builds on recent advances in machine translation and computer vision. Szegedy et al. (2016) describe Inception-V3, a convolutional neural network that can be used to detect the main objects that appear in an image with very high accuracy.

Pertaining to the association of content with familiar expressions, (Tan et al., 2016) use neural networks to recommend quotes in writing and to make statements more compelling. They point out how computational methods can help writers select the most appropriate quote for a given context from a large repository of alternatives.

Regarding the appropriateness of proverbs as image captions, B. Mieder and Mieder (1977) analyze the reasons behind the common usage of proverbs in advertisement. Proverbs have a "familiar ring" that adds reliability, trustworthyness and a sense of timelessness to a brand or product. More recently, Qing-fang (2004) observes that proverbs are especially suitable for advertisement as they are short and concise, and they are associated with wisdom and moral guidance. To say it in the words of the author, "one proverb may say more than a thousand words".

Associating proverbs to images

In this section, we describe the architecture of a system that, given an image and a set of proverbs, decides whether the image is evocative of one of the proverbs. In particular, we use PROMETHEUS (Özbal et al., 2016) as a proverb repository, but a different set of proverbs or other types of memorable expressions (such as slogans or quotations) could be used in alternative. The resource consists of 1,054 proverbs, grouped into categories (such as "love and hate" or "fate") and annotated with metaphors at the word and sentence level. More than in other genres, such as news, fiction and essays, in proverbs metaphors can resolve a significant amount of the figurative meaning (Faycel, 2012). The richness of proverbs in terms of metaphors and their pervasiveness in all cultures makes them especially suitable for being used as evocative captions (W. Mieder, 1978).

We first use the proverbs to retrieve a large set of noisy data from the web. Then, we use this data to train two convolutional neural networks to associate proverbs to images. The two classifiers use the same architecture, but one is trained to directly associate images to proverbs, while the other builds associations between the objects that it recognizes in the images and the proverbs. Then, we use a small sample of the predictions of the two models to crowd source golden image-proverb associations. Finally, we use the noisy data and the golden labels to combine the output of the two classifiers into a unified model that decides whether it should select the proverb suggested by any of the two classifiers.

Noisy data collection

For each proverb in PROMETHEUS, we used the Flickr API to retrieve a set of candidate images. We included the full text of the proverb as part of the query string, forcing the API to only return images that mention the complete proverb in their title, description or tags. In our experiments we focus on the 98 proverbs for which we could retrieve at least 500 images.

To keep the data set reasonably balanced, we also limit the maximum number of images retrieved for each proverb to 1,000. The resulting data set consists of 83,895 images, each of which is associated with exactly one of 98 distinct proverbs. We then randomly split the data into a training (80,000 images) and a development (3,895 images) set. For the purpose of training and testing the classifiers, we used Flickr API to download 150×150 pixel versions of the images. These are obtained by cropping to a square around the main subject and then scaling to the final size, thus preventing warping or distortions of the elements of the images. As we reckon that color plays an important role with respect to the mood and perceived message of a picture, we did not convert the images to black and white.

Image classification

In this section, we describe the training of two classifiers that, given an image, predict the most likely proverb association. Both classifiers are based on Inception-V3, a convolutional neural network which has been shown to be very accurate in image classification tasks with a large number (1,000) of output classes (Szegedy et al., 2016). For each input image, the model outputs a probability distribution over all the output classes. The predicted label is the class with the highest probability density. For all our experiments, we use the Inception-V3 implementation included in the TensorFlow-Slim image classification model library[1].

Inception from scratch (I-FS) The first model is trained to establish a direct association between the visual clues present in the image and the output proverbs. It is an Inception-V3 network trained *from scratch* (I-FS) on the available training data. We use all the default settings of Slim's Inception implementation and we select the model after 669,923 iterations.

Inception fine-tuned (I-FT) We fine-tune the model starting from the Inception-V3 model[2] trained by Szegedy et al. (2016). This model was trained from the 1.2 million images of the 2012 ImageNet Large Scale Visual Recognition

[1] https://goo.gl/W5ZdQ4
[2] https://goo.gl/nrsdGG


(a)


(b)

Figure 2: Examples of reasonable predictions that differ from the noisy label. (a) Label: "Look before you leap". I-FS: "Rules are made to be broken.". (b) Label: "Beggars can't be choosers.". I-FS and I-FT: "Time and tide wait for no man".

Challenge (ILSVRC-12) (Russakovsky et al., 2015). We refer to the resulting proverb classifier as I-FT, for *Inception fine-tuned*. As the proverb classification task has a different number of output classes from ImageNet (i.e., 98 vs. 1,000), we do not restore the weights of the final layer of the network[3]. In addition, we only allow the weights of the classification layer to be updated during fine-tuning. In doing so, we expect the classifier to retain the object recognition capabilities of the internal layers of the pre-trained model and to establish meaningful association between the target proverb and the dominant objects in an image. Concerning I-FT, we select the model after 1,955,892 iterations[4].

Evaluation of I-FS and I-FT We measured the performance of the two classifiers on the 3,895 images in the development split of the noisy data. I-FT's recall is consistently higher than I-FS's (Recall@1: 0.20 vs. 0.15; Recall@5: 0.39 vs. 0.28). This is an expected result, as the inner layers of I-FT encode classification clues learned from a very large data set. While recall is relatively low for both classifiers, we should consider that each image can possibly evoke more than one proverb, whereas in our data set we only have one label for each image. Therefore, we regard these figures as very conservative lower bounds. For example, Figure 2 shows two images for which the decisions of the classifiers are quite reasonable, yet they do not agree with the noisy label.

It is also important to observe that the two classifiers learn very different models, as exemplified in Table 1. I-FS ans I-FT output a different label in the large majority of the cases (85%), and 27% of the times at least one of the two classifiers can reconstruct the correct association according to the noisy labels. In the next sections, we will explain how we leverage the different "personalities" of the two classifiers and combine them into a unified model that can predict a golden (i.e., human validated) proverb with an accuracy of 74.59%.

[3] https://goo.gl/tfHxzs

[4] We let both I-FS and I-FT learn for ≈1 week. Then, among the last 5 checkpoints, we selected the one having the smallest loss on the training data. Since there is no previous work to compare against, we are not trying to maximize accuracy at all costs. Instead, we aim to demonstrate that our pipeline produces results that are adequate for a range of user facing applications, as those mentioned in the introduction.

Statistics on development data	Count	%
Same prediction	579	14.87
Same prediction, both incorrect	226	5.80
Same predictions, both correct	353	9.06
Different predictions	3,316	85.13
Different predictions, both incorrect	2,604	66.85
I-FS correct, I-FT not correct	276	7.09
I-FT correct, I-FS not correct	436	11.19
I-FT or I-FS prediction correct	1,065	27.34

Table 1: Comparison of I-FS and I-FT. Correct and incorrect counts refer to the noisy development labels.

Gold standard collection

In the previous section, we observed that there is a number of cases in which the output of I-FS or I-FT are more suitable captions for a given image than its noisy label. To quantify this phenomenon, we set-up a crowd-sourced annotation in which we showed the raters an image and four proverbs, and asked the raters to select the most appropriate caption. To maximize the utility of the annotation, we included only the cases in which both models disagree with the noisy label. We decided to crowd-source the annotation of 500 images on the Figure-Eight platform[5].

We first included all the 226 development examples for which the two models predict the same label and the prediction is incorrect (2nd row in Table 1). We refer to these as *Type1* examples. We regard these examples as especially relevant, as we have seen before that the two models do not agree very often. Our hypothesis is that, in many such cases, the models are actually converging to a meaningful interpretation. Then, we added 274 randomly sampled images for which the predictions of the two models differ, and both predictions differ from the noisy label (*Type2*).

For Type1 examples, the raters could choose among: (1) the noisy label, (2) the proverb selected by I-FS and I-FT, and (3 and 4) two random proverbs. For Type2 examples, the raters could choose among: (1) the noisy label, (2) I-FS prediction, (3) I-FT prediction, and (4) a random proverb. In both cases, the random proverbs were selected among the 98 proverbs used to train the models. The raters were instructed to select all the relevant associations, and they also had the option to mark none of the proposed alternatives as relevant.

Due to the inherent subjectivity of the task, we decided to elicit 10 judgments for each image, for a total of 5,000 ratings. The agreement on the ratings, as reported by Figure-Eight, is 64.47%. The aggregated results of the annotation based on majority voting[6] are shown in Table 2. We can see that, overall, raters tend to prefer the decisions of I-FT over the noisy label (27.21% vs. 24.87%), and the noisy label over I-FS (20.70%). It is quite remarkable that I-FT's predictions are rated to be more accurate than the data on

[5] https://www.figure-eight.com/

[6] Even though raters could select multiple options, the majority decision has never included more than one.

	Times selected (%)		
Selected label	Overall	Type1	Type2
Noisy label	24.87	17.85	33.21
Random	3.84	3.69	4.01
None	23.37	17.54	30.29
I-FS	20.70	30.46	9.12
I-FT	27.21	30.46	23.36
I-FS or I-FT	31.39	30.46	32.48

Table 2: Results of the crowd-sourced annotation.

Label	Annotated data	Noisy data	Total
Either	99	353	452
None	312	-	312
I-FS	25	276	301
I-FT	64	436	500
Total	500	1,065	1,565

Table 3: Data distribution of the combined classifier. Note that we only annotated 500 examples out of 2,830 for which both I-FS and I-FT fail to predict the noisy label. As a consequence, 2,330 development examples are not included in this experiment.

which the model has been trained. When the two classifiers make the same decision (Type1), there is a marked preference of the raters for the predicted proverb over the noisy label (30.46% vs. 17.85%), whereas when the two classifiers do not agree (Type2) the raters generally find the noisy label preferable, even though the cases in which either I-FS or I-FT are chosen are almost the same with the noisy label (32.48% vs. 33.21%). Even though I-FS is not as accurate as I-FT to predict the noisy labels, there is a non negligible number of cases in which its decision is considered to be appropriate by the raters, and when the decisions of the two classifiers differ (Type2), I-FS selects a good option in 9.12% of the cases. There are very few cases (3.84% overall) in which a random proverb is preferred to any of the more principled alternatives, whereas there is a very significant number of cases (23.37% overall) in which none of the proposed alternatives, including the noisy label, is considered to be good.

Model combination

In this section, we describe a classifier that, given an image and the output of I-FS and I-FT, classifies the image into one of the following four classes: (a) *I*-FS, if the prediction of I-FS should be selected; (b) *I*-FT, if I-FT should be preferred instead; (c) *None*, for the cases in which neither of the two classifiers predicted an appropriate class; and (d) *Either*, if both the predictions of I-FS and I-FT are appropriate. We introduce the last class *Either* specifically to model the cases in which I-FS and I-FT output the same prediction.

Data set All the annotated examples for which the raters did not select either I-FS or I-FT predictions were mapped to the *None* class. These are all the images annotated as "Noisy label", "None" or "Random". Type1 examples where the prediction of the models was preferred by the raters were mapped to *Either*, whereas Type2 examples where I-FS or I-FT were preferred were mapped to the corresponding label. The distribution of the labels of the annotated data is summarized on the left side of Table 3. By construction, the annotated data contains only cases in which I-FS's and I-FT's predictions differ from the noisy label, and the *None* label is significantly over-represented. In order to come up with a more balanced data set, we also include the non-annotated examples in which either classifier agreed with the noisy label. If both classifiers agree with the noisy label, then we map the example to the *Either* label. If only I-FS (or I-FT) agrees, then we map the example to the I-FS (or I-FT) class. The column labeled "Noisy data" in Table 3 shows the distribution of the data added in this fashion. We regard these examples as highly accurate, given the low chance of random agreement between the noisy label and the classifiers (the output space of I-FS and I-FT consists of 98 proverbs).

Features From each example we extract 12 simple features,which we group into six sets to simplify the feature selection experiments. The set labeled "Base" (***b***) only accounts for the decisions of I-FT and I-FS. To avoid overfitting, we only include the prediction scores, and not the actual predicted classes. The set labeled "Metaphoricity" (***m***) makes use of the proverb-level metaphoricity annotations in PROMETHEUS. The metaphoricity can have one of three values: 0 (literal); 1 (slightly metaphorical); 2 (highly metaphorical). We expect proverbs which are metaphorical to be a good fit for a broader set of images. The feature set "Inception" (***i***) encodes the highest prediction score of the Inception-V3 model for the image. The intuition here is that a high prediction score, regardless of the class, means that the Inception-V3 model is confident that it can recognize a known object in the image. We use this measure as a proxy for the "concreteness" of the image, as a counterpart for the data encoded by *m*. The set "Category similarity" (***cs***) attempts to measure the compatibility between the category of the proverb (e.g., "love and hate" or "fate") and the object recognized in the picture by the Inception-V3 model. We use the DISCO (Kolb, 2009) library together with the provided English word space[7] and encode as feature the maximum cosine similarity between any synonym in the synset predicted by Inception-V3 and any content word in the predicted proverb categories. The feature set "Proverb similarity" (***ps***) is conceptually very similar, but we use the lemmas in the predicted proverb instead of its category. Finally the feature set "Difference" (***d***) encodes the difference in magnitude between the values of the feature in *b* and *m* computed for I-FS and I-FT. These features are meant to help the classifier reason more comparatively about I-FS and I-FT predictions.

Set-up To make the most of the available training data, we evaluate the combination of the two models in a leave-one-out setting, i.e., a cross-fold where the number of folds equals

[7] https://goo.gl/Rc45PW

F1	b	b+ m†‡	b+ d†	b+ ps	b+ i	b+ cs	b,m+ d†‡	b,m+ i	b,m+ ps	b,m+ cs
Macro	53.81	59.90	54.40	54.24	54.06	53.99	60.25	58.00	56.69	56.11
Micro	66.52	68.56	67.09	66.90	66.84	66.77	68.88	67.92	67.22	67.03

Table 4: Feature ablation results for the best learning algorithm. †: Significantly better than *b*. ‡: Significantly better than *b,d*. The difference between *b,m,d* and *b,m* is not significant.

the number of test examples. Please note that none of the images in the test set of the combined classifier is included in the traning of I-FS or I-FT. We compare different groupings of feature sets (always including *b*). As a learning algorithm, we use an SVM with a polynomial kernel of degree 2. We use the implementations provided by SciKit-Learn (Pedregosa et al., 2011). To compare the different feature combinations, we use McNemar's significance test (McNemar, 1947) with a 95% confidence interval ($p < 0.05$).

Results In Table 4 we report the detailed results of the feature inclusion experiments. The set of base features *b* alone achieves a micro F1 measure of 66.52. If we try to add another set of features on top of *b*, only *b,m* and *b,d* achieve a significant improvement, with *b,m* being significantly more accurate than *b,d* (68.56 vs. 67.09). If we try to add another feature set on top of *b,m*, we observe that only *b,m,d* achieves a higher accuracy (i.e., 68.88 vs. 68.56), even though the improvement is not significant. Adding any other feature set yields a negative contribution (micro F1 < 68).

As a further comparison between *b,m* and *b,d,m*, Table 5 shows the difference between the confusion matrices of the two configurations. We can observe that the error distribution of the two models is very similar, with the former being slightly more accurate on the examples labeled I-FS and *Either*, and the latter on *None* and I-FT. Interestingly, both models make very few mistakes on examples labeled *Either*, confirming that the convergence of I-FS and I-FT predictions is a strong signal of the accuracy of the predicted proverb. The error distribution also reflects the fact that I-FT, being a more accurate predictor than I-FS, is more represented in the training data. In fact, there are many more examples labeled I-FS which are predicted as I-FT than the other way round. For the same reason, the model also tends to predict I-FT when the actual label is *None*. All in all, this error analysis suggests that the best way to improve the classifier might be to introduce more data points for the classes *None* and I-FS, which are under-represented in the data (see Table 3).

From all the evidence above, we can conclude that the information about the metaphoricity of the predicted proverb provides very useful clues to the learning algorithm.[8] Contrary to our expectations, the features that account for the similarity between the objects in the pictures and the predicted proverbs (*i*, *ps* and *cs*) do not improve the classification accuracy.

Label	Predicted label: None		I-FS		I-FT		Either	
None	38	(**41**)	33	(37)	151	(148)	90	(86)
I-FS	0	(0)	**167**	(161)	134	(140)	0	(0)
I-FT	3	(1)	69	(61)	428	(**438**)	0	(0)
Either	12	(14)	0	(0)	0	(0)	**440**	(438)

Table 5: Confusion matrices for the combined model with feature groups *b,m* and *b,m.d* (in parentheses).

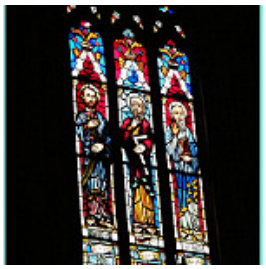
Finally, in Figure 3 we show 10 examples of system outputs (for the configuration using feature sets *b,m,d*), which we believe are quite representative of what the model has learned. Not all outputs are correct according to the golden labels, and we invite the readers to figure out which examples are correct and which are not before continuing reading (the answer is at the end of the paragraph). Looking at the outputs, we can see that in some cases (e.g., (d) and (i)) the associations are quite literal (hay, detergents). In other cases, the association is less obvious. These are the most interesting cases, in which the predictions showcase the ability of the model to abstract away from concrete objects, or to reproduce the cultural biases observed in the training data . In (a) there is a sense of frugality that is resolved to "every little helps". Concerning (b), in the training data "slow but sure" is very often associated with religious symbols, churches in particular. In (f), the model associates the flooded land with "storm" and the ships with "port". In (g), the model recognized the quietness of situation and the golden tones of the scenes. Concerning (h), a crowded school of fish evokes the association with "first come, first served". According to the golden labels, examples (a) to (e) are classified correctly, whereas the ones from (f) to (j) are incorrect. Nevertheless, for the applications that we have in mind all examples seem appropriate. This fact can be confirmed by restricting the evaluation to the examples annotated by the raters and by considering all the proverbs that have been selected by at least one human rater as good predictions. Under these conditions, the model selects an appropriate proverb in 74.59% of the cases.

Copyright and credits

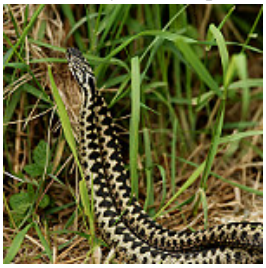
We are extremely grateful to the authors of the images included in the paper for releasing their images under a permissive licensing scheme or for explicitly allowing us to use their pictures. This section lists all the images used in the paper, including their author, licensing scheme and Flickr URL. All

[8] We have observed the same pattern also using different learning algorithms (RBF, LR), but here we omit these results due to space limitations.

(a) Every little helps.



(b) Slow but sure.



(c) Two heads are better than one.

(d) Make hay while the sun shines.

(e) Like father, like son.

(f) Any port in a storm.

(g) Silence is golden.

(h) First come, first served.

(i) Cleanliness is next to godliness.

(j) Seeing is believing.

Figure 3: Example outputs of the combined model. Five outputs differ from the corresponding noisy label. Can you tell which ones?

the listed URLs were active at the time of submission.

Figure 1. Author: Flickr user "Dano". License: CC BY 2.0[9]. Source: https://www.flickr.com/photos/mukluk/249464230.

Figure 2(a). Author: Flickr user "Gavin Clarke". License: CC BY-NC 2.0[10]. Source: https://flickr.com/photos/70824176@N00/4460439903.

Figure 2(b). Author: Jason Swain. All rights reserved. Used under permission by the author. Source: https://flickr.com/photos/24424426@N00/13058126593.

Figure 3(a). Author: Flickr User "Neil Moralee". License: CC BY-NC-ND 2.0 [11]. Source: https://flickr.com/photos/62586117@N05/21178964709.

Figure 3(b). Author: Flickr User "Cathedrals and Churches". License: CC BY 2.0[9]. Source: https://www.flickr.com/photos/eltb/7246837670/.

Figure 3(c). Author: Flickr User "Peter Trimming". License: CC BY 2.0[9]. Source: https://www.flickr.com/photos/55426027@N03/8730055756.

Figure 3(d). Author: Flickr User "Raymond Barlow". License: CC BY-NC-SA 2.0 [12]. Source: https://flickr.com/photos/62673829@N00/2631618525.

Figure 3(e). © Jay Heymans. All rights reserved. Used under permission by the author. Source: https://www.flickr.com/photos/7830239@N06/12234997804.

Figure 3(f). © Ian Huges. All rights reserved. Used under permission by the author. Source: https://flickr.com/photos/36463157@N08/3818175700.

Figure 3(g). Author: Flickr User "Geraint Rowland". License: CC BY-NC 2.0[10]. Source: https://flickr.com/photos/33909206@N04/23407737789.

Figure 3(h). Author: Flickr user "Steven Harris". License: CC BY-NC 2.0[10]. Source: https://flickr.com/photos/90288178@N00/4060998399.

Figure 3(i). © Melissa Jones. All rights reserved. Used under permission by the author. Source: https://www.flickr.com/photos/msjones166/5511643604.

Figure 3(j). Author: Flickr user "TheoJunior". License: CC BY-NC-SA 2.0[12]. Source: https://flickr.com/photos/88013568@N00/3252673888.

Conclusion and future work

In this paper, we presented a model that can associate images to proverbs. It combines two variants of a high-performance convolutional neural network in a simple voting scheme, it is easily scalable and it requires very minimal supervision. By leveraging high volumes of noisy training data, the model can learn compelling associations at surprising levels of abstraction, such as "Misery loves company." for a sweaty bunch of skaters. To our best knowledge, we are the first ones to

[9] https://creativecommons.org/licenses/by/2.0/
[10] https://creativecommons.org/licenses/by-nc/2.0/
[11] https://creativecommons.org/licenses/by-nc-nd/2.0/
[12] https://creativecommons.org/licenses/by-nc-sa/2.0/

use existing object recognition models to associate images to semantically rich, non-descriptive captions such as proverbs.

Our approach can easily be extended to cover more proverbs as well as other kinds of memorable and familiar expressions, such as slogans, citations or titles of famous works of art that have already been the focus of previous work on creative language generation (Gatti, Özbal, Guerini, Stock, & Strapparava, 2015; Özbal, Pighin, & Strapparava, 2013; Stock, Strapparava, & Valitutti, 2007). We have shown that knowledge about the metaphoricity degree of proverbs plays a significant role with respect to the classification accuracy. While PROMETHEUS already provides this information, this might not be the case for other sources of familiar expressions. On the other hand, it should be possible to automatically assess metaphoricity by leveraging recent state-of-the-art advancements in the field of metaphor detection (Özbal, Strapparava, Tekiroglu, & Pighin, 2016; Veale, Shutova, & Klebanov, 2016). In addition, we would like to generate more captivating captions, by injecting humor into the predicted proverbs through incongruity (Raskin, 1979) or other rhetorical devices. As Veale (2012) suggests, linguistic creativity can be utilized to "re-invent and re-imagine the familiar, so that everything old can be made new again".

References

Chen, J., Kuznetsova, P., Warren, D., & Choi, Y. (2015). Déja image-captions: A corpus of expressive descriptions in repetition. In *Proceedings of NAACL-HLT'15.*

Faycel, D. (2012). Food Metaphors in Tunisian Arabic Proverbs. *Rice Working Papers in Linguistics*, *3*.

Gatti, L., Özbal, G., Guerini, M., Stock, O., & Strapparava, C. (2015). Slogans are not Forever: Adapting Linguistic Expressions to the News. In *Proceedings of IJCAI'15.*

Gelli, F., Uricchio, T., Bertini, M., Del Bimbo, A., & Chang, S.-F. (2015). Image Popularity Prediction in Social Media Using Sentiment and Context Features. In *Proceedings of ICM'15.*

Hall, M. M., Jones, C. B., & Smart, P. (2015). Spatial Natural Language Generation for Location Description in Photo Captions. In *Proceedings of COSIT'15.*

Hessel, J., Lee, L., & Mimno, D. (2017). Cats and Captions vs. Creators and the Clock: Comparing Multimodal Content to Context in Predicting Relative Popularity. In *Proceedings of WWW'17.*

Kolb, P. (2009). Experiments on the Difference between Semantic Similarity and Relatedness. In K. Jokinen & E. Bick (Eds.), *Proceedings of NODALIDA'09.*

McNemar, Q. (1947, Jun 01). Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages. *Psychometrika*, *12*(2), 153–157.

Mieder, B., & Mieder, W. (1977). Tradition and Innovation: Proverbs in Advertising. *The Journal of Popular Culture*, *11*(2), 308–319.

Mieder, W. (1978). Proverbial Slogans are the Name of the Game. *Kentucky Folklore Record*, *24*(2), 49.

Özbal, G., Pighin, D., & Strapparava, C. (2013). BRAINSUP: Brainstorming Support for Creative Sentence Generation. In *Proceedings of ACL'13.*

Özbal, G., Strapparava, C., & Tekiroglu, S. S. (2016). PROMETHEUS: A Corpus of Proverbs Annotated with Metaphors. In *Proceedings of LREC'16.*

Özbal, G., Strapparava, C., Tekiroglu, S. S., & Pighin, D. (2016). Learning to Identify Metaphors from a Corpus of Proverbs. In *Proceedings of EMNLP'16.*

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Qing-fang, X. (2004). The Innovative Use of Proverbs in Advertising English. *Journal of Pla University of Foreign Languages*, *5*, 003.

Raskin, V. (1979). Semantic mechanisms of humor. In *Annual Meeting of the Berkeley Linguistics Society* (Vol. 5, pp. 325–335).

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, *115*(3), 211-252.

Stock, O., Strapparava, C., & Valitutti, A. (2007). Moving Creative Words. *Advances in Brain, Vision, and Artificial Intelligence*, 509–522.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of CVPR'16.*

Tan, J., Wan, X., & Xiao, J. (2016). A Neural Network Approach to Quote Recommendation in Writings. In *Proceedings of CIKM'16* (pp. 65–74).

Veale, T. (2012). *Exploding the Creativity Myth: The Computational Foundations of Linguistic Creativity*.

Veale, T., Shutova, E., & Klebanov, B. B. (2016). Metaphor: A computational perspective. *Synthesis Lectures on Human Language Technologies*, *9*(1), 1–160.

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2017). Show and Tell: Lessons Learned from the 2015 MSCOCO Image Captioning Challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(4), 652–663.

Yamaguchi, K., Berg, T. L., & Ortiz, L. E. (2014). Chic or Social: Visual Popularity Analysis in Online Fashion Networks. In *Proceedings of ICM'14.*