

Comparing Text-Enriched Image Representations to Improve News Image Classification

Elias Moons¹, Tinne Tuytelaars², and Marie-Francine Moens¹

¹ KU Leuven, Department of Computer Science

² KU Leuven, Department of Electrical Engineering,
elias.moons@cs.kuleuven.be, tinne.tuytelaars@esat.kuleuven.be,
sien.moens@kuleuven.be

Abstract. Images have a prominent role in the communication of news. In this paper, we classify news images into subject categories and focus on the specific case where we have only a limited amount of annotated training images. For this task, we learn text-enriched image representations or embeddings from a large dataset of news texts and their images. Once trained, the encoder will convert any image to its text-enriched representations. The text-enriched image embeddings are then used in an image classification model. We compare our model with a baseline using only standard image features in their classification. In this way we prove that we can construct better news image classifiers by using corresponding textual information during training. In our experiments we have trained classifiers with different amounts of annotated examples. We found that our text-enriched image classifier outperforms the image-only baseline model. The difference in performance is even more pronounced when the size of the annotated training dataset is smaller. We discuss the environments in which our model performs best and which parameters have an impact on the classification results.

Keywords: Multimodal classification, Neural networks, News documents, Limited training data

1 Introduction

People want access to accurate, clear and visually attractive information on their smartphones, tablets, computers, etc. Images gain in importance in the transfer of information to users and become an interesting source of information for data mining purposes. Web news companies receive a lot of images that could be used in news articles, but these are often not accompanied by text and even less by subject categories. We need technology to make sense of these images and filter out the most relevant ones. Classification of the images in subject categories is a first step in their processing, and the main topic of this paper. The within-class variability for these categories is huge and challenges traditional image

classification methods. Moreover, these categories are not mutually exclusive (e.g., an article might cover both sports and politics).

The Web is a large source of visual information, but only a small part of it is annotated with subject categories by humans. Because many classification tasks experience the problem of training with a limited number of annotated data, our aim is to build a classification model for news images that uses only a limited amount of labeled supervision.

We propose a method for learning better image representations or embeddings by exploiting naturally co-occurring text and then learn an image classifier on top using a limited amount of training data. We find on the Web a large amount of online news articles that contain both images and text, the latter in the form of the article text and/or a caption attached to an image. We incorporate this textual knowledge by training a neural encoder that learns a mapping from an image-only domain to a text-enriched image domain. The encoder is trained with a large collection of news images and co-occurring text fragments. In this way we build representations of images that are informed by text. We then use these text-enriched image representations as input of a classifier that will assign subject categories to the images and which is trained on labeled image data. Because we want to exploit the textual information co-occurring with the images, but still want to train a model that can classify images based only on image features, we have constructed our model in two parts. The first part is the neural encoder which is trained on the large image-text dataset and which at test time builds a text-enriched representation of any input image. The second part of the model acts as the actual classifier. This network takes a text-enriched image representation as input and produces a prediction for the membership of each of the subject classes for that image. This subnetwork is trained on labeled images.

We will compare the quality of the proposed model to an image classifier model that relies on pure image features, neglecting the text-enrichment of the representations. As an extra reference, we constructed a second text informed model (named ‘biview’) which is an extension of the the proposed text-enriched image model. The quality of these models will be discussed in different settings (amount of labeled/unlabeled training data, different subject categories).

2 Related work

Automated classification or categorization of Web content has a long-standing tradition with the goal to enhance the performance of a Web search [5, 13]. However, content classification is mostly restricted to the classification of textual documents where support vector machines form an established technique. The amount of visual data in the form of images and video distributed via the World Wide Web rapidly increases, hence the need for accurate image classifiers. Most of the existing models for classification of visual data rely on labeled training data (e.g., [16]). The availability of large scale benchmark datasets such as ImageNet [6] has boosted the use of data-driven methods based on deep learn-

ing. Classifiers trained on these benchmarks are used directly to classify Web images/video (e.g. [12, 14]). They can as well be used as a way to obtain a powerful image representation on top of which a final classifier can be trained [15], or as initialization of (part of) another neural network (finetuning) [21]. Such transfer learning techniques drastically reduce the amount of application-specific annotations needed compared to training models from scratch. Nevertheless, for subject categories with large within class variability, the amount of annotations required often remains well beyond what is practically viable, urging us to look into alternative methods, exploiting context or unlabeled data.

There are some works that integrate context when classifying Web visual content. Examples are the work of Guermazi et al. [10], who exploit textual descriptions, anchor texts and URLs, or Song et al. [18], who use title, description and keywords of Web videos. Yet neither of them propose methods for building joint representations or embeddings of image content and its context, as we do. Huang et al. [11] compute multimodal embeddings of social media images by exploiting image, text and links between images, but these authors focus on object classification and not on the classification into (much more diverse) subject categories. Finally, Chatbri et al. [4] propose automated MOOC video classification using transcripts and images.

We witness a recent interest in using large amounts of unlabeled or weakly labeled data to learn embeddings or representations of words and phrases in order to address the problem of lack of annotated training data in a classification task. Tang et al. [19] leverage large amounts of weakly supervised data (i.e., tweets weakly classified by emoticons) to learn sentiment embeddings of n-grams of words, which in a second step are then used as features in a classification task. Deriu et al. [7] leverage large amounts of weakly supervised data in a multilayer convolutional network in order to classify the opinions in multilingual tweets, and evidence the importance of pre-training the network using the weakly supervised data (i.e., again tweets weakly classified by emoticons). On the visual side, various unsupervised image representations based on egomotion [22], spatial [9] or temporal [20] structural information have been proposed, based on which classifiers can later be trained in a supervised manner using a limited amount of training data. In this paper we follow a similar philosophy, but instead of training text representations with the supervision of other weakly labeled texts or training visual representations with structural information, we propose to improve image representations using the weak supervision of their textual context.

3 Methods

In this paper we propose three architectures for the actual news image classification. The first architecture uses the text-enriched image representations obtained by the encoder to classify images. The second architecture builds on the idea of the previous model and combines the text-enriched representations with the original pure-image representations in the classification. The third model

is a baseline architecture that only relies on pure-image representations for the classification in subject categories. All models use some form of preprocessing discussed below.

3.1 Image preprocessing

An initial representation of the visual data is generated by a VGG-16 network, a deep convolutional network described in [17], configured with 16 weight layers. 13 of these layers are convolutional layers, all of which are configured with 3×3 receptive fields, and 3 of them are fully connected layers. In this network, the input image is sent through 5 convolutional and pooling blocks. After these blocks, the network consists of 3 fully connected layers. The used network was pre-trained on ImageNet [1] by classifying given images into 1000 different object categories of ImageNet.

A first step is to rescale all the images to the input size of the network. To obtain the initial image representations, the last layer of the pre-trained network is removed. The output of the network is now a 4096-dimensional vector. These output vectors, they are rescaled to the real $[0, 1]$ -interval. This is done to make the image and text representational vectors as similar as possible which will be easier to deal with when training the models. The image representational vectors will serve as the purely-image feature based representations for the images in the dataset and as input for the encoder that generates text-enriched representations.

3.2 Text preprocessing

A first preprocessing step is tokenization. In this phase, the text is transformed into an ordered list of words (tokens) using the NLTK word tokenizer [2]. Then, for all training documents, a set of the top 4096 most discriminating tokens (empirically determined) is calculated using a Chi-Square analysis [8]. All texts are reduced so that they only contain words from this set of key tokens, while maintaining their order. As a last step, in all texts that contain more than 20 tokens we drop all but the first 20 of them. This will only be the case for long articles surrounding an image where typically the first paragraph contains a concise summary of the whole article. Smaller texts are padded with empty tokens until they have reached 20 in total. Now all texts are of the same length, which is required by the architecture of the neural network that generates the textual representations. Similar to the preprocessing of the images, the texts are finally transformed to 4096-dimensional vectors. These initial embedding representations for the texts are produced by an LSTM-based neural network.

3.3 Text-enriched image classifier

The text-enriched model consists of two parts, which are trained separately. First an encoder is trained on a dataset of corresponding visual and textual representations. Second, an actual classifier is trained on labeled images. As a consequence

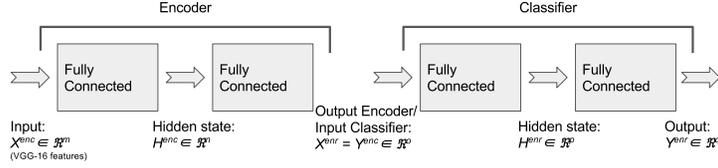


Fig. 1. Full architecture of the enriched image classifier. The network is trained in 2 phases with different datasets. At test-time, the network can be seen as 1 larger 4-layered network.

of using this 2-phased process, the data pool for the encoder can be considerably larger than the one for the classifier itself. This is interesting since, in a lot of problems, only a small portion of the available data is (qualitatively) annotated. In this model, unannotated data is still useful to train a better encoder.

Encoder The encoder has a 2-layered, symmetrical structure (visible in figure 1). The input of the encoder is a 4096-dimensional vector X^{enc} , corresponding to the initial image representation. The data passes through two fully connected layers whose transformation matrices are $W^{enc,1}$ and $W^{enc,2}$, obtaining respectively a hidden representation H^{enc} and an output Y^{enc} . All nodes in this network use a softmax activation function and the network is trained with a binary cross entropy objective function. More formally, the state of one hidden node $h_i^{enc} \in H^{enc}$ is calculated as follows (with F being a softmax activation function):

$$h_i^{enc} = \sum_{j=1}^m F(w_{i,j}^{enc,1} \cdot x_j^{enc}) \quad (1)$$

with $m = 4096$. The state of an output node $y_i^{enc} \in Y^{enc}$ is generated by:

$$y_i^{enc} = \sum_{j=1}^n F(w_{i,j}^{enc,2} \cdot h_j^{enc}) \quad (2)$$

with $n = 1024$. At training time, the encoder learns to predict the corresponding text representations for given image representations (both 4096-dimensional vectors). At test time, when the image representation of a given sample is fed into the network, the encoder produces a corresponding text-enriched image representation.

Classifier The second part of the network functions as a classifier and has a 2-layered fully-connected structure (also displayed in figure 1). The input of the classifier is an enriched image vector representation X^{enr} of an image. This vector passes through two fully connected layers with transformation matrices $W^{enr,1}$ and $W^{enr,2}$, obtaining respectively a hidden representation H^{enr} and an output Y^{enr} . This network is also trained with a binary cross entropy objective,

since the classification task is multilabel. Formally, the state of one hidden node $h_i^{enr} \in H^{enr}$ is calculated as follows:

$$h_i^{enr} = \sum_{j=1}^o F(w_{i,j}^{enr,1} \cdot x_j^{enr}) \quad (3)$$

with $o = 4096$. The state of an output node $y_i^{enr} \in Y^{enr}$ is generated by:

$$y_i^{enr} = \sum_{j=1}^p F(w_{i,j}^{enr,2} \cdot h_j^{enr}) \quad (4)$$

with $p = 256$. For a given (enriched) image (representation), the output of the network will be a vector with membership predictions for each of the subject categories. At test time, an image (representation) is subsequently sent through the encoder and then through the classifier (visible in figure 1) to obtain membership predictions for the given subject categories.

3.4 Biview image classifier

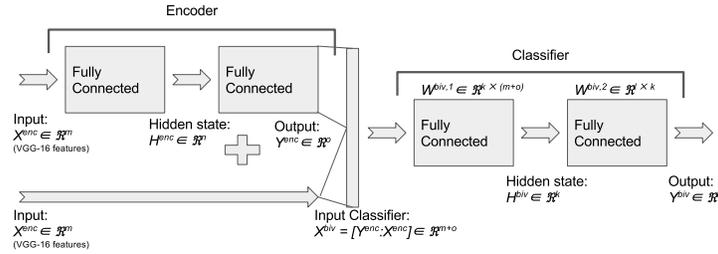


Fig. 2. Full architecture of the biview image classifier. The network is trained in 2 phases with different datasets. At test-time, the network can be seen as 1 larger 4-layered network.

This model builds on the text-enriched classifier and tries to combine the representational quality of both the original image representation and the text-enriched representation of the same image. This model consists of two important parts (see figure 2). The first part is an encoder, trained exactly the same way as in the text-enriched image model. The second part is a fully connected network which takes as input the concatenation of the enriched representation of an image as well as the original purely image-feature based representation of that image. The input of the second subnetwork is now an 8192-dimensional vector X^{biv} . This vector serves as input for the first fully connected layer of 256 nodes, with weight matrix $W^{biv,1}$ to form an internal, hidden representation H^{biv} . The output layer takes these 256-dimensional vectors to make membership predictions for each

of the subject categories, Y^{biv} , using a transformation vector $W^{biv,2}$. Similar to the text-enriched image model, this network is also trained using a binary cross entropy objective function. Formally, the output of the encoder Y^{enc} is calculated from the input X^{enc} exactly as displayed in equations 1 and 2. Let $[Y^{enc} : X^{enc}]$ denote the concatenation of these vectors Y^{enc} and X^{enc} , then the input vector X^{biv} for the second part of the biview network is constructed as follows:

$$X^{biv} = [Y^{enc} : X^{enc}]. \quad (5)$$

The state of one hidden node $h_i^{biv} \in H^{biv}$ is calculated like this:

$$h_i^{biv} = \sum_{j=1}^{m+o} F(w_{i,j}^{biv,1} \cdot x_j^{biv}) \quad (6)$$

with $m = 4096$ and $o = 4096$. The state of an output node $y_i^{biv} \in Y^{biv}$ is generated by:

$$y_i^{biv} = \sum_{j=1}^k F(w_{i,j}^{biv,2} \cdot h_j^{biv}) \quad (7)$$

with $k = 256$.

3.5 Image-only baseline

The architecture of the image-only model is identical to the classifier part of the text-enriched image model. The input is a 4096-dimensional input vector which corresponds to the purely image-feature based representation of an image. These vectors are created as explained in section 3.1. The network itself consists of 2 fully connected layers of 256 internal nodes and 7 output nodes corresponding to the predictions for the 7 categories. All these nodes make use of a softmax activation function and the network is trained using a binary cross entropy objective function. The network functions as described in equations 3 and 4.

3.6 Class assignment

For all the networks described above, class assignment is done in the same way. Whether a given instance is a member of a certain class is determined by calculating the predicted value from the classification model for a specific class and class assignment is true if this value is greater or smaller than a certain cutoff value. This cutoff value is a number between 0 and 1 and can be selected in such a way to focus the model to give higher importance to either precision (higher cutoff value) or recall (lower cutoff value).

4 Results

4.1 Data

The data used to verify the models presented in this paper comes from the Webhose [3] dataset. In this collection, in total 493432 news articles are sorted into 7

different subject categories (entertainment, finance, politics, sports, technology, travel and world news). For every category a list of JSON files is given. Each of these files corresponds to one news article on the Web. 2 parts of these files are used in the experiments in this paper: the text and a link to the main image corresponding to the article. To limit the dataset for computational efficiency, out of all the files, 10000 in total were chosen uniformly at random.³ Since all files are chosen uniformly at random, the category probability distribution of the selected dataset is similar to the overall probability distribution of the dataset, but it is not explicitly enforced when selecting the 10000 files. From all these files, the textual information was extracted and the image was downloaded (if the link was still active). The models presented in this paper require (partly annotated) data consisting of both an image and a corresponding text fragment. Because of this, if the image described in the JSON file could not be downloaded, the training sample was dropped from the dataset.

To work with these data in a neural network context, vector representations were trained both for the textual as the visual parts of the data. The key tokens were selected based on their Chi-Square statistic (see section 3.2) which was computed using the whole Webhose dataset (the text fragments from all 493432 files and not just the 10000 selected files) minus the files that are part of the test dataset, described further down, to get a more general selection of keywords over all the news texts available. If during the preprocessing of the text it became clear that a text fragment did not contain any of the 4096 keywords, this sample was dropped from the dataset. Otherwise the textual representation for that piece of text would not bear any information and would thus be useless for training purposes. This leaves a dataset of 7884 data points. Before doing any sort of training, a test set of 500 samples is drawn uniformly at random from the selected dataset. This test set is kept aside during all different tests and serves as the reference to measure the performance of the models. In our dataset, some images occur multiple times (since these images occur in multiple articles). To assure a qualitative evaluation of the models, images from the test set that occur multiple times in the whole dataset are removed. This leaves an actual test set of in total 378 samples.

The training set for the encoders contain all the samples of the whole dataset without the selected test set samples. This leaves a training set for the encoder of 7384 samples. As it is discussed below, the size of training set of the image classifier will vary in the experiments, but in all experiments 80% of the training data is used for training and 20% for validation and parameter tuning of the neural networks.

In all the experiments, unless specifically stated otherwise, a threshold cut-off value of 0.5 is used for image category prediction. Because we are dealing with a multilabel problem and with class-imbalance of the samples for the different categories, results in this paper are reported based on micro-averaged precision and recall metrics (and thus also F1 score) instead of accuracy.

³ Upon acceptance, the text and image representations will be made publicly available, as well as all data splits used in the experiments.

4.2 Results of the image classifier when trained on the full annotated dataset

The results of the image baseline, the text-enriched image classifier and the biview image classifier that are trained on the whole training dataset are listed in table 1. On the test set (described above), both models which incorporate text

Table 1. Accuracy, precision, recall and F1 scores for the 3 different image classification models trained on the full available training dataset.

Model	Acc	Prec	Rec	F1
Image	87.17%	58.69%	34.66%	43.55%
Text-enriched	87.03%	57.00%	37.56%	45.28%
Biview	86.64%	54.48%	39.95%	46.07%

in their training phase outperform the image-only baseline in terms of F1 score. The enriched-text model improves the F1 score of the image baseline with 1.73%. The biview model does 2.52% better in terms of F1 score than the baseline.

4.3 Influence of fraction of annotated data on classification results

Details of the experiment In this experiment, the quality of the different classification models is assessed with respect to the size of training data of the image classifier. For all multiples of 1000 smaller than the size of the total training set (7384), training sets are selected uniformly at random of that specific size. This experiment was repeated 5 times in total and the scores are averaged over all the experiments to obtain more qualitative results.

Results of experiment for specific sizes of annotated training dataset

In figure 3 the F1 scores for the three different image classification models are displayed. These scores are micro averaged over all different subject categories. Based on the F1 score, both classifiers that incorporate the corresponding textual information into their training phase outperform the image-only baseline classifier for all considered sizes of the training data. Out of the two proposed models however, one does not outperform the other. This probably means that the text-enriched image representation still captures (a big part of) the original pure-image information. This way it seems that the extra information that the Biview classifier has is mostly redundant.

Table 2 gives more detailed results for one relatively smaller dataset of 2000 annotated data points and one relatively larger dataset of 5000 annotated training instances. For the case of 2000 samples there is a clear outperformance in F1 score of the text-enriched and the biview model with respect to the image baseline. The biview classifier leads to an improvement of 4.08% in F1 score on the baseline, the text-enriched classifier even improves the baseline by 4.93% (p-value of 0.023, 2-tailed t-test). For the case of 5000 samples the difference in

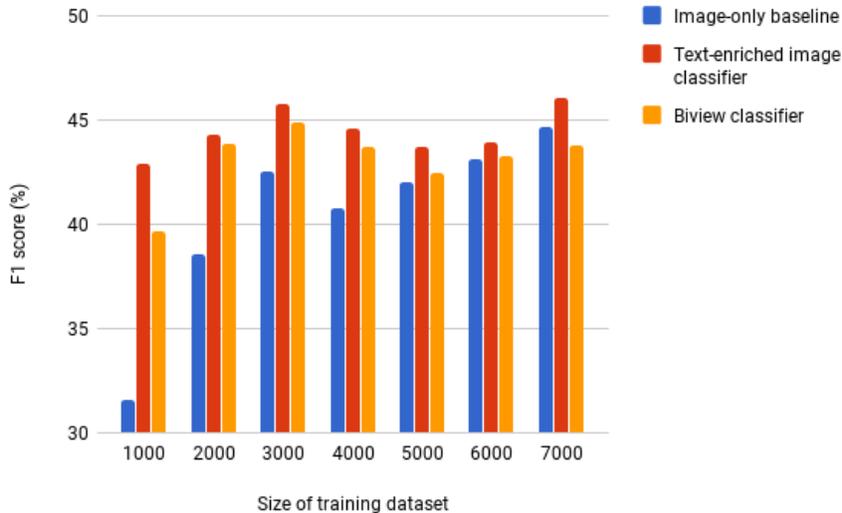


Fig. 3. F1 score of the image classifiers in function of the fraction of the annotated image dataset: Results of the image-only baseline are shown in blue, of the text-enriched image classifier in red and of the biview classifier in orange.

F1 score between the text-enriched classifiers and the pure image baseline is less than in the first case (as expected by the results displayed in figure 3) but still significant. The biview model outperforms the baseline image classifier by 1.44% in F1 score, the text-enriched model outperforms the baseline by 1.84% (p-value of 0.040, 2-tailed t-test).

Table 2. Accuracy, precision, recall and F1 scores for the image classification models trained on 2000 and 5000 annotated training instances

Model	Size	Acc	Prec	Rec	F1
Image	2000	87.18%	60.47%	29.84%	39.93%
Text-enriched	2000	87.21%	58.34%	36.46%	44.86%
Image	5000	87.41%	60.48%	34.55%	43.92%
Text-enriched	5000	87.10%	57.34%	38.10%	45.76%

Results for specific subject categories Since all our viewed results are averages over all subject categories it is also interesting to look at the results for specific categories. This is to see whether the conclusions from the averages are generally applicable or whether some specific categories influence the average.

These data (for the specific case of a training dataset of 3000 data instances⁴) is visible in table 3. These results confirm that the earlier findings are correct:

Table 3. F1 scores for each of the different subject categories for the 3 image classifiers.

Category	Image	Text-enriched	Biview
Entertainment	37.06%	47.47%	45.00%
Finance	13.24%	8.32%	11.03%
Politics	47.49%	46.09%	48.45%
Sports	70.42%	70.01%	71.32%
Technology	9.09%	34.51%	31.40%
Travel	29.85%	32.25%	35.49%
World News	12.45%	20.88%	25.31%
Macro Average	42.04%	45.12%	45.78%

the textually informed image classifiers outperform the image-only baseline. In only one of the 7 categories the image classifier performs best and this category (finance) is one that scores particularly badly compared to the other categories. For four of the other categories, the textually informed classifiers do significantly better than the image-only baseline. The difference is generally bigger when the image baseline scores are lower (more room for improvement).

5 Conclusion

In this work we have described and evaluated two image classification models, that is, the text-enriched image classifier and its variant, the biview image classifier, that incorporate textual information to obtain text-enriched representations or embeddings of image features. We have tested these models in different environments to assess their classification quality. Our experiments show that the text-enriched classifiers outperform a purely image-feature based classifier on a Web news image dataset from Webhose. The difference in performance is even more pronounced when the image classifier is trained only a small set of examples labeled with subject categories.

References

1. Imagenet dataset. <http://www.image-net.org/> (2016), accessed: 2017-10-24
2. Natural language toolkit. <http://www.nltk.org/> (2017), accessed: 2017-10-31
3. Webhose dataset. <https://webhose.io/datasets> (2017), accessed: 2017-10-24
4. Chatbri, H., McGuinness, K., Little, S., Zhou, J., Kameyama, K., Kwan, P., O’Connor, N.: Automatic mooc video classification using transcript features and convolutional neural networks. In: ACM MM - MultiEdTech Workshop (2017)

⁴ The choice for a dataset of 3000 samples is only illustrative, results are similar for different sizes of training datasets and can be provided upon acceptance.

5. Chekuri, C., Goldwasser, M.H., Raghavan, P., Upfal, E.: Web search using automatic classification. In: WWW (1997)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR, IEEE Conference on. pp. 248–255. IEEE (2009)
7. Deriu, J., Lucchi, A., Luca, V.D., Severyn, A., Müller, S., Cieliebak, M., Hoffmann, T., Jaggi, M.: Leveraging large amounts of weakly supervised data for multi-language sentiment classification. In: WWW, Perth, Australia, April 3-7, 2017. pp. 1045–1052 (2017)
8. Diaconis, P., Efron, B.: Testing for independence in a two-way table: New interpretations of the chi-square statistic. *The Annals of Statistics* 13(3), 845–874 (1985), <http://www.jstor.org/stable/2241103>
9. Doersch, C., Gupta, A., Efros, A.A.: Unsupervised visual representation learning by context prediction. In: IEEE ICCV. pp. 1422–1430 (2015)
10. Guermazi, R., Hammami, M., Hamadou, A.B.: Classification of violent web images using context based analysis. In: ACM SAC, Sierre, Switzerland, March 22-26, 2010. pp. 1768–1773 (2010)
11. Huang, F., Zhang, X., Li, Z., Mei, T., He, Y., Zhao, Z.: Learning social image embedding with deep multimodal attention networks. In: ACM MM, Mountain View, CA, USA, October 23 - 27, 2017. pp. 460–468 (2017)
12. Mei, T., Zhang, C.: Deep learning for intelligent video analysis. In: ACM MM, Mountain View, CA, USA, October 23-27, 2017. pp. 1955–1956 (2017), <http://doi.acm.org/10.1145/3123266.3130141>
13. Qi, X., Davison, B.D.: Web page classification: Features and algorithms. *ACM Comput. Surv.* 41(2), 12:1–12:31 (Feb 2009), <http://doi.acm.org/10.1145/1459352.1459357>
14. Real, E., Moore, S., Selle, A., Saxena, S., Suematsu, Y.L., Tan, J., Le, Q.V., Kurakin, A.: Large-scale evolution of image classifiers. In: ICML, Sydney, NSW, Australia, 6-11 August 2017. pp. 2902–2911 (2017), <http://proceedings.mlr.press/v70/real17a.html>
15. Sharif Razavian, A., Azizpour, H., Sullivan, J., Carlsson, S.: Cnn features off-the-shelf: an astounding baseline for recognition. In: CVPR, IEEE Conference on. pp. 806–813 (2014)
16. Simonet, V.: Classifying youtube channels: A practical system. In: WWW. pp. 1295–1304. WWW '13 Companion, ACM, New York, NY, USA (2013), <http://doi.acm.org/10.1145/2487788.2488164>
17. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. *CoRR* abs/1409.1556 (2014), <http://arxiv.org/abs/1409.1556>
18. Song, Y., Zhao, M., Yagnik, J., Wu, X.: Taxonomic classification for web-based videos. In: CVPR, IEEE Conference on. pp. 871–878. IEEE (2010)
19. Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., Qin, B.: Learning sentiment-specific word embedding for twitter sentiment classification. In: ACL, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers. pp. 1555–1565 (2014)
20. Wang, X., Gupta, A.: Unsupervised learning of visual representations using videos. In: IEEE ICCV. pp. 2794–2802 (2015)
21. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? In: NIPS. pp. 3320–3328 (2014)
22. Zhou, T., Brown, M., Snavely, N., Lowe, D.G.: Unsupervised learning of depth and ego-motion from video. *arXiv preprint arXiv:1704.07813* (2017)