Multimodal News Article Analysis*

Arnau Ramisa
Institut de Robòtica i Informàtica Industrial, IRI (UPC-CSIC), Barcelona, Catalonia, Spain
Wide Eyes Technologies, Barcelona, Catalonia, Spain
aramisa@iri.upc.edu

Abstract

The intersection of Computer Vision and Natural Language Processing has been a hot topic of research in recent years, with results that were unthinkable only a few years ago. In view of this progress, we want to highlight online news articles as a potential next step for this area of research. The rich interrelations of text, tags, images or videos, as well as a vast corpus of general knowledge are an exciting benchmark for high-capacity models such as the deep neural networks. In this paper we present a series of tasks and baseline approaches to leverage corpus such as the BreakingNews dataset.

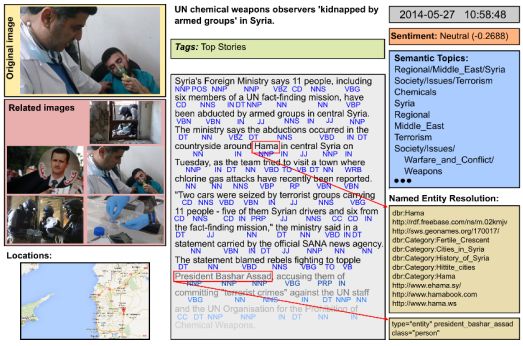

Figure 1: Example article from the BreakingNews dataset.

1 Introduction

The results presented in this paper are joint work with my co-authors Fei Yan, Francesc Moreno-Noguer and Krystian Mikolajczyk; it is as much theirs as it is mine.

Research at the intersection of traditionally separate artificial intelligence areas, such as Computer Vision and Natural Language Processing, has received a great deal of attention in recent times, with many outstanding results in tasks like automatic image captioning [Vinyals *et al.*, 2015] or image generation from simple sentences [Chang *et al.*, 2014]. These good results in suggest that other, more complex domains, may be ready to be explored.

Nowadays, online news articles have a rich ecosystem of related data coming in many modalities: text, tags, images, captions, videos, comments, mentions on social media, etc. Furthermore, news articles are often interrelated, offering complementary views of the same events. Making sense of this trove of connected data is a very challenging enterprise, both by the scale of the data itself, and by the complexity of the relations at play. Hence, it is a very attractive benchmark on which to test and develop new, powerful artificial intelligence algorithms.

In [Ramisa *et al.*, 2016] the authors present a dataset of a hundred thousand news articles from several high-ranked online newspapers and news portals published during year 2014. In addition to the article text, they include related information, like the pictures accompanying the articles and their captions, geolocation information or reader comments. See Figure 1 for statistical information on the dataset. Its size and composition make this dataset an excellent benchmark for data-hungry state-of-the-art deep learning models.

In this paper we will briefly discuss several tasks that can be undertaken using the BreakingNews dataset, plus a baseline approach to tackle some of them.

2 Tasks description

Source detection: This task consists of classifying the newspaper that originally published a given article. One approach to solve it is to model the language and writing preferences, subjects of interest, and sentiment towards certain type of news for each newspaper, but other types of data could be brought into play by also modeling the type of illustrations used, or the number and style of comments posted by the readers (where the sentiment of the audience which the newspaper caters to will be more strongly represented). The task of authorship identification is closely related.

Popularity prediction: The objective of this task is to determine the impact that an article will have after its publication, which can be a useful tool for publishers and community managers. Good indicators of the popularity of an article are the number of times it has been shared on social networks, or the number of comments made by the readers.

*This work was partly funded by the ERA-net CHISTERA project VISEN PCIN-2013-047 and the MINECO project RobInstruct TIN2014-58178-R. The authors would like to thank NVidia for the hardware donation under the GPU grant program.

Source	num. articles	avg. len. article	avg. num. images	avg. len. caption	avg. num. comments	avg. len. comment	avg. num. shares	% geolocated
Yahoo News	10,834	521 ± 338	1.00 ± 0.00	40 ± 33	126 ± 658	39 ± 71	n/a	65.2%
BBC News	17,959	380 ± 240	1.54 ± 0.82	14 ± 4	7 ± 78	48 ± 21	n/a	48.7%
The Irish Independent	4,073	555 ± 396	1.00 ± 0.00	14 ± 14	1 ± 6	17 ± 5	4 ± 20	52.3%
Sydney Morning Herald	6,025	684 ± 395	1.38 ± 0.71	14 ± 10	6 ± 37	58 ± 55	718 ± 4976	60.4%
The Telegraph	29,757	700 ± 449	1.01 ± 0.12	16 ± 8	59 ± 251	45 ± 65	355 ± 2867	59.3%
The Guardian	20,141	786 ± 527	1.18 ± 0.59	20 ± 8	180 ± 359	53 ± 64	1509 ± 7555	61.5%
The Washington Post	9,839	777 ± 477	1.10 ± 0.43	25 ± 17	98 ± 342	43 ± 50	n/a	61.3%

Table 1: BreakingNews dataset statistics. Mean and standard deviation, usually rounded to the nearest integer.

Geolocation: The goal in this case is determining the geographical location or locations where the news events take place, expressed as GPS coordinates. Both the Computer Vision [Kalogerakis *et al.*, 2009; Weyand *et al.*, 2016] and the Natural Language Processing [Serdyukov *et al.*, 2009] communities have approached this problem, and some works have combined both sources of information [Zhou and Luo, 2012] to improve the performance.

Text illustration: In this task the objective is to find the best image to illustrate a news article from a database of images. Variants of the task may include some annotations (e.g. the captions) for the images, or a small gallery of images that can be retrieved together. Many works have addressed this task, some in the news article domain [Feng and Lapata, 2010; 2013], but more commonly in datasets of images with short and accurate descriptions [Barnard *et al.*, 2003; Kovashka *et al.*, 2015; Rasiwasia *et al.*, 2007; Douze *et al.*, 2011]. In the news domain, neither the main body of the article nor the caption of the image offer a faithful description of the picture contents, adding an another layer of difficulty to the problem.

Caption generation: A very popular task in recent Computer Vision literature consists on generating short text descriptions of the contents of a picture [Chen and Zitnick, 2015; Donahue *et al.*, 2015; Karpathy and Fei-Fei, 2015; Kiros *et al.*, 2015; Vinyals *et al.*, 2015]. Yet, the task of generating captions for news articles is significantly different since, as mentioned in the previous paragraph, captions are often only indirectly related to the contents of the image. A pioneering example of automatic caption generation for news articles is [Feng and Lapata, 2008].

3 Text and image representation

In this section we will describe state-of-the-art representations for text and images that can be useful for multimodal news article analysis.

3.1 Text representations

One of the most widely used representations for text documents is the Bag-of-Words (BoW). In this representation, a document is encoded as a histogram with the frequencies with which the words in a vocabulary appear in the text. Despite its simplicity, the Bag-of-Words is a very sound baseline to compare to, as its performance is not far from that of more sophisticated approaches. One caveat of the BoW representation is that very frequent but otherwise non-informative words like "a", "the" or "this" may end up dominating the histogram of counts, thus failing to represent useful information for the subsequent tasks. To counter this undesirable effect often this very common words are excluded from the vocabulary, and ignored when building the histogram. Furthermore, the Term Frequencies (TF) can be weighted by the Inverse Document Frequencies (IDF), which boost the relevance of the most infrequent (and discriminative) words. The BoW with TF-IDF weighting is computed as follows:

$$D \in \mathbb{R}^b | D_j = t_j \log \frac{M}{c^j + 1} \quad (1)$$

where D is the b-dimensional TF-IDF Bag of Words representation, M is the total number of documents in the corpus, t_j is the term frequency of the j^{th} token (i.e. the number of times it appears in the article), and c^j is the document frequency of the token (i.e. the number of training articles where it appears).

In recent years, distributed representations for text, like word2vec, have raised a great deal of attention, due to their excellent performance in many tasks. In contrast with BoW models, these representations do not treat words as atomic units, represented as indices in a vocabulary, but as short, real-valued embeddings based on their context, in a way that preserves semantic similarity. For example, the embedding corresponding to "king" minus that of "man", and plus the one of "woman", will be close to the representation of "queen". These arithmetic properties have been shown to hold across a wide range of linguistic regularities [Mikolov *et al.*, 2013].

In order to scale to large training set sizes (in the billions or even trillions of words) these vector representations are often trained using a simple shallow model and stochastic gradient descent, where the objective to maximize are the log probabilities of words given other words in their context, where the probabilities are estimated using computationally efficient approximations to softmax, like the hierarchical softmax or noise contrastive estimation.

Certain short phrases are not combinations of other words, but have a separate identity. Think for example of the newspaper "Boston Globe": it is not a natural combination of the meanings of "Boston" and "Globe". Consequently, the expressiveness of the model is improved by treating common short phrases as individual tokens.

When it comes to longer sentences or even full documents, however, other strategies have to be considered. One very simple strategy is to do a (weighted) average of all the vector

representations in the document, which unfortunately loses the word order information, as in the case of BoW.

Alternatively, in [Le and Mikolov, 2014] the authors propose a learning framework similar to the one for words, where an embedding for the words and the documents is optimized simultaneously over the complete corpus. At test time, new documents can be encoded using gradient descent with all the parameters of the model except the document vector held fix. This framework bears some resemblance to the Fisher kernels, a image representation method with successful applications in Computer Vision.

Finally, one can frame the representation of documents as a fixed-size matrix of word embeddings, padded with zeros, and use 1-D temporal convolutions in a Convolutional Neural Network architecture (CNN), popular at the moment thanks to their extraordinary results in Speech Recognition, Computer Vision and other fields [Kim, 2014; Kalchbrenner *et al.*, 2014].

3.2 Image representation

The *de facto* standard when it comes to image representations in state-of-the-art computer vision literature are the final layers of a deep convolutional neural network trained on the ImageNet dataset [Krizhevsky *et al.*, 2012], or of a fine-tuned version of the network if some training data is available.

Convolutional Neural Networks are a type of feed-forward models that include one or more convolution layers, where responses to patterns in delimited image regions are computed by filters made up of specialized neurons. These filters are applied in each position of a partially overlapped tile over the image, which in practice results in a convolution operation that produces a collection of response maps (one for each filter). The fact that these filters are shared by every position in the tile over the image greatly reduces the number of parameters with respect to fully connected layers, resulting in many advantages, like more efficient architectures, reduced risk of overfitting and a degree of invariance to translations in the input image.

Standard CNN architectures typically project the image through a chain of convolution and pooling layers, after which a last pooling layer, or a number of fully connected layers, reduce the dimensionality further for the classification layer that outputs the predicted class probabilities. These final layers are the ones typically used as a compact representation of the image contents.

As in the case of distributed representations for text, it is possible to average (or max-pool) the vectors of several pictures to obtain a more compact embedding of the whole set.

4 Learning models

In recent literature, Deep Neural Networks models have shown impressive results. Abundant computing power, and large datasets allow these very large models to learn end-to-end the best intermediate representation for the objective task, often substantially better than the best hand-crafted representation developed for each particular case.

For multimodal article analysis, we found that the FC7 layer of a standard image classification CNN like VGG19 (after the non-linearity) produced representations of sufficient quality for the article images. Although it would also be feasible to fine-tune the vision CNN with news article images, for example predicting the tags of the articles given the pictures, in our case this did not yield any significant advantage.

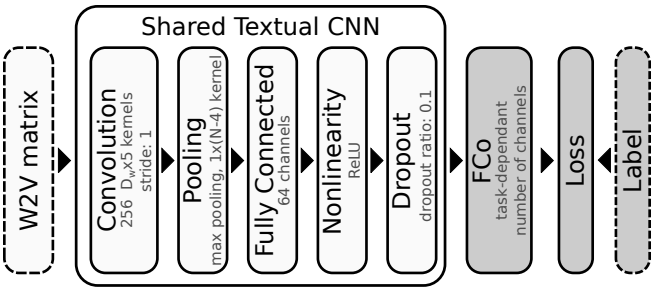

Figure 2: CNN architectures for article analysis, used in the tasks of article illustration, source detection and popularity and geolocation prediction. Dashed boxes are the CNN inputs, and solid boxes correspond to the layers. White layers are shared by all tasks, and shaded layers are task specific.

Regarding the article text representation, we obtained the best results with a small convolutional network that uses the matrices of stacked word2vec representations (zero-padded to accommodate to a fixed dimensionality) as input. The network has one convolutional layer and one pooling layer, followed by a fully connected layer, a non-linearity (ReLU proved to be the most numerically stable option) and a dropout layer to regularize the model and prevent overfitting. A graphical representation of the text-processing network can be seen in Figure 2. After this common text processing network, a layer that computes the error between the network predictions and the ground truth labels is necessary to teach the network through backpropagation. Different *objectives*, or *loss layers*, can be added to fulfill most of the different tasks in section 2. Here we will describe some of them.

Multinomial Logistic Loss: This loss maximizes the cross entropy for a one-of-k classification problem. Let $\hat{p}_i$ be a predicted probability distribution for the sample i in a batch of N training samples, and y_i the ground truth label for the same example; then this loss can be computed for a batch of training examples as:

$$E = \frac{-1}{N} \sum_{i=1}^{N} \log \hat{p}_{i,y_i} \tag{2}$$

The probability distribution $\hat{p}$ is often obtained using the softmax function on the outputs of the last fully connected layer of a network.

Canonical Correlation Analysis (CCA) loss: The CCA loss is used to find linear projections that maximize the correlation between pairs of corresponding data points in different modalities. Let Z and Y be two $d \times m$ matrices of m aligned data points, then a CCA-based loss can be defined as $L = -corr(Z, Y)$, whose gradients can be computed using a singular value decomposition of $\Gamma = \Sigma_{zz}^{-1/2} \Sigma_{zy} \Sigma_{yy}^{-1/2}$, where Σ_{zz}, Σ_{yy} and Σ_{zy} are respectively the covariances of Z and Y, and the cross covariances. Please refer to [Ramisa *et al.*, 2016] for a complete derivation of the CCA loss.

Single task vs Multitask vs Transfer Learning				
		Geolocation Median GCD	**Popularity Median ℓ_1**	**Source Bal. acc.**
Single-task		**0.90**	1.09	**80.7**
Multitask	G+P	1.17	1.68	-
	G+S	1.22	-	79.1
	P+S	-	1.94	79.1
Transfer	G→P	-	1.16	-
	G→S	-	-	80.2
	P→G	0.97	-	-
	P→S	-	-	77.6
	S→G	0.92	-	-
	S→P	-	**0.63**	-

Table 2: Comparing single-task, multitask and transfer learing. G: geolocation; P: popularity; S: source. An arrow shows the direction of transfer, for example, G→P means trained on task G and transferred to task P.

L1 distance loss: This loss is simply the L1 distance between the predicted value z and the ground truth label y, useful for regression tasks: $L = |z - y|$.

Great Circle Distance: This loss was introduced in [Ramisa *et al.*, 2016], and computes the geodesic distance between a predicted and a ground truth latitude and longitude pairs. Let the label $\mathbf{y} = [y_1, y_2]$, where $y_1 \in [-\pi/2, \pi/2]$ is the latitude and $y_2 \in [-pi, pi]$ the longitude. Then the output of the network $z \in \mathbb{R}^2$ will be the predicted latitude and longitude value, and the Great Circle Distance (GCD) loss can be computed using the spherical law of cosines approximation (and dropping the Earth radius term):

$$\mathrm{L} = \arccos(\sin y_1 \sin z_1 + \cos y_1 \cos z_1 \cos \delta) \qquad (3)$$

where $\delta = z_2 - y_2$. The derivation can be found in [Ramisa *et al.*, 2016].

Furthermore, recent work [Zhang *et al.*, 2014] has shown that sharing a common representation to fulfill several tasks at the same time, can have unforeseen advantageous effects on each of the tasks, as the internal representation of the network can share the "capacities" learned for each one of the tasks to solve the others.

Similarly, networks trained for a particular task can be quickly fine-tuned to solve another related objective with very little computational effort and training data, sometimes even surpassing the performance obtained by a network directly trained on the target task.

5 Results and conclusions

In this section we present the particular experimental details and results obtained with the BreakingNews dataset.

We split the dataset according to the standard partitions[1], which gives training, validation and tests sets of 60%, 20% and 20% respectively. The textual representations used are a TF-IDF BoW with a vocabulary of 44,665 dimensions, and a word2vec embedding of 500 dimensions computed on the training set. The images are represented as concatenated ReLU7 features from the ImageNet VGG19 and PLACES CNNs, and the model architectureis that of Figure 2 with the appropriate loss function(s).

For article illustration we measure performance primarily using the median rank (lower is better). For each test article we rank the 23,200 test images according to the projection learned by the CNN or by a CCA model. The best results are very similar with both methods: 405 for the CNN model using a W2V matrix and the concatenation of PLACES and VGG19 RelU7 features, and 397 for the CCA using the same visual representation and the BoW TF-IDF for the text.

The dataset includes a set of five *related images* for each article, that often include the actual article image. If we construct the image representation as the average of their VGG19 ReLU7 vectors, median rank goes down to 137 for CCA, which is a significant improvement. If we add the caption information to the representation on the image side, the median rank drops to 7, and if we anonimize the captions by replacing person names, locations and organizations to make them more neutral, it is still reduced by an order of magnitude to 42.

The remaining three tasks are evaluated individually, and in a multitask and transfer learning scenarios (see Table 2). For multitask, we jointly train a model with each of the three pairs of tasks. For transfer learning, we train six models using the three pairs and in both directions.

Geolocation is evaluated in terms of median Great Circle Distance between the predicted and the ground truth locations, popularity on the median l_1 distance between the predicted and real number of comments, and source prediction based on the balanced classification accuracy. The results in Table 2 show that single-task learning tends to have the best performance, and that transfer learning in general outperforms multitask learning. The fact that multitask learning does not perform as well as the other alternatives seems to indicate that forcing the lower layers of the CNN to share common representations harms the performance of both tasks, and that our tasks are not as "compatible" as those considered in [Zhang *et al.*, 2014]. CNN architectures that further decouple the tasks did not improve the results.

On the other hand, in most cases the performance of transfer learning is comparable to that of single-task learning. When transferring from "source" to "popularity" prediction, the performance is even higher than single-task "popularity" prediction. Indicating that the low and mid-level representations learned by source prediction helped the other task. For multitask and transfer learning, we set the learning rate to one tenth of the base learning rate, to help the layers change slowly and reduce the co-adaptation effect [Yosinski *et al.*, 2014].

[1] http://www.iri.upc.edu/people/aramisa/BreakingNews/index.html

References

[Barnard *et al.*, 2003] K. Barnard, P. Duygulu, D. Forsyth, N. De Freitas, D. Blei, and M. Jordan. Matching words and pictures. *The Journal of Machine Learning Research*, 3:1107–1135, 2003.

[Chang *et al.*, 2014] A. Chang, M. Savva, and C. Manning. Interactive learning of spatial knowledge for text to 3d scene generation. *Sponsor: Idibon*, page 14, 2014.

[Chen and Zitnick, 2015] X. Chen and C. Zitnick. Mind's eye: A recurrent visual representation for image caption generation. In *CVPR*, 2015.

[Donahue *et al.*, 2015] J. Donahue, L. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, . Saenko, and T.Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2015.

[Douze *et al.*, 2011] M. Douze, A. Ramisa, and C. Schmid. Combining attributes and fisher vectors for efficient image retrieval. In *CVPR*, pages 745–752, 2011.

[Feng and Lapata, 2008] Yansong Feng and Mirella Lapata. Automatic image annotation using auxiliary text information. In *Proceedings of ACL-08: HLT*, pages 272–280, Columbus, Ohio, 2008.

[Feng and Lapata, 2010] Y. Feng and M. Lapata. Topic Models for Image Annotation and Text Illustration. *Conference of the North American Chapter of the ACL: Human Language Technologies*, (June):831–839, 2010.

[Feng and Lapata, 2013] Yansong Feng and Mirella Lapata. Automatic caption generation for news images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(4):797–812, 2013.

[Kalchbrenner *et al.*, 2014] N. Kalchbrenner, E. Grefenstette, and P. Blunsom. A convolutional neural network for modelling sentences. In *ACL*, 2014.

[Kalogerakis *et al.*, 2009] E. Kalogerakis, , O. Vesselova, J. Hays, A. Efros, and A. Hertzmann. Image sequence geolocation with human travel priors. In *ICCV*, pages 253–260, 2009.

[Karpathy and Fei-Fei, 2015] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015.

[Kim, 2014] Y. Kim. Convolutional neural networks for sentence classification. In *EMNLP*, 2014.

[Kiros *et al.*, 2015] R. Kiros, R. Salakhutdinov, and R. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *TACL*, 2015.

[Kovashka *et al.*, 2015] A. Kovashka, D. Parikh, and K. Grauman. Whittlesearch: Interactive image search with relative attribute feedback. *International Journal of Computer Vision*, pages 1–26, 2015.

[Krizhevsky *et al.*, 2012] A. Krizhevsky, I. Sutskever, and G. Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*, 2012.

[Le and Mikolov, 2014] Q. Le and T. Mikolov. Distributed representations of sentences and documents. In *ICML*, 2014.

[Mikolov *et al.*, 2013] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013.

[Ramisa *et al.*, 2016] A. Ramisa, F. Yan, F. Moreno-Noguer, and K. Mikolajczyk. Breakingnews: Article annotation by image and text processing. arXiv:1603.07141 [cs.CV], 2016.

[Rasiwasia *et al.*, 2007] N. Rasiwasia, P. Moreno, and N. Vasconcelos. Bridging the gap: Query by semantic example. *IEEE Transactions on Multimedia*, 9(5):923–938, 2007.

[Serdyukov *et al.*, 2009] P. Serdyukov, V. Murdock, and R. Van Zwol. Placing flickr photos on a map. In *ACM Conference on Research and Development in Information Retrieval*, pages 484–491. ACM, 2009.

[Vinyals *et al.*, 2015] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *CVPR*, pages 3156–3164, 2015.

[Weyand *et al.*, 2016] T. Weyand, I. Kostrikov, and James Philbin. Planet - photo geolocation with convolutional neural networks. arXiv:1602.05314 [cs.CV], 2016.

[Yosinski *et al.*, 2014] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In *NIPS*, pages 3320–3328, 2014.

[Zhang *et al.*, 2014] Z. Zhang, P. Luo, C. Loy, and X. Tang. Facial landmark detection by deep multitask learning. In *ECCV*, 2014.

[Zhou and Luo, 2012] Y. Zhou and J. Luo. Geo-location inference on news articles via multimodal pLSA. *ACM International Conference on Multimedia*, page 741, 2012.