BreakingNews: Article Annotation by Image and Text Processing

Arnau Ramisa*, Fei Yan*, Francesc Moreno-Noguer, and Krystian Mikolajczyk, *Member, IEEE*

Abstract—Building upon recent Deep Neural Network architectures, current approaches lying in the intersection of Computer Vision and Natural Language Processing have achieved unprecedented breakthroughs in tasks like automatic captioning or image retrieval. Most of these learning methods, though, rely on large training sets of images associated with human annotations that specifically describe the visual content. In this paper we propose to go a step further and explore the more complex cases where textual descriptions are loosely related to the images. We focus on the particular domain of news articles in which the textual content often expresses connotative and ambiguous relations that are only suggested but not directly inferred from images. We introduce an adaptive CNN architecture that shares most of the structure for multiple tasks including source detection, article illustration and geolocation of articles. Deep Canonical Correlation Analysis is deployed for article illustration, and a new loss function based on Great Circle Distance is proposed for geolocation. Furthermore, we present BreakingNews, a novel dataset with approximately 100K news articles including images, text and captions, and enriched with heterogeneous meta-data (such as GPS coordinates and user comments). We show this dataset to be appropriate to explore all aforementioned problems, for which we provide a baseline performance using various Deep Learning architectures, and different representations of the textual and visual features. We report very promising results and bring to light several limitations of current state-of-the-art in this kind of domain, which we hope will help spur progress in the field.

Index Terms—News Dataset, Story Illustration, Geolocation, Caption Generation, Vision and Text.

1 Introduction

In recent years, there has been a growing interest in exploring the relation between images and language. Simultaneous progress in the fields of Computer Vision (CV) and Natural Language Processing (NLP) has led to impressive results in learning both image-to-text and text-to-image connections. Tasks such as automatic image captioning [8], [16], [37], [41], [75], [82], image retrieval [21], [31], [45], [48] or image generation from sentences [6], [88] have shown unprecedented results, which claimed to be similar to the performance expected from a three-year old child[1].

One of the main reasons behind the success of these approaches is the resurgence of deep learning for modeling data, which has been possible due to the development of new parallel computers and GPU architectures and due to the release of new large datasets, used to train deep models with many parameters.

The popularity of crowd sourcing tools has facilitated the proliferation of a number of these datasets combining visual and language content. Among them, the most widely known are the UIUC Pascal Sentence Dataset [66], the SBU captioned photo dataset [61], Flickr8K [31], Flickr30K [84] and MS-COCO [51]. All these datasets consist of a number of images (from 1K to 1M), each annotated by human written sentences (between 1 and 5 per image). The annotations are typically short and accurate sentences (of less than 20 words) describing the visual content of the image and the action taking place. In contrast, other and more complex types of documents, like illustrated news articles, have been barely explored.

We believe that current successes in the crossroads between NLP and Computer Vision indicate that the techniques are mature for more challenging objectives than those posed by existing datasets. The NLP community has been addressing tasks such as sentiment analysis, popularity prediction, summarization, source identification or geolocation to name a few that have been relatively little explored in Computer Vision. In this paper we propose several learning schemes. Specifically, for the source detection, article illustration and geolocation prediction, we consider an adaptive CNN architecture, that shares most of the structure for all the problems, and just requires replacing and retraining the last layers in order to tackle each particular problem. For the caption generation task, image and text representations are combined in a Long Short-Term Network (LSTM).

In order to evaluate these algorithms, we have collected *BreakingNews*, a large-scale dataset of news articles with rich meta-data. Our dataset consists of approximately 100K news articles, illustrated by one to three images and their corresponding captions. Addi-

- *Arnau Ramisa and Francesc Moreno-Noguer are with the Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Barcelona, 08028, Spain. Email: {aramisa, fmoreno}@iri.upc.edu.*
- *Fei Yan is with the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, UK. Email: f.yan@surrey.ac.uk.*
- *Krystian Mikolajczyk is with the Department of Electrical and Electronic Engineering, Imperial College London, UK. Email: k.mikolajczyk@imperial.ac.uk.*

* *The first two authors contributed equally.*

[1] http://www.ted.com/talks/fei_fei_li_how_we_re_teaching_computers_to_understand_pictures

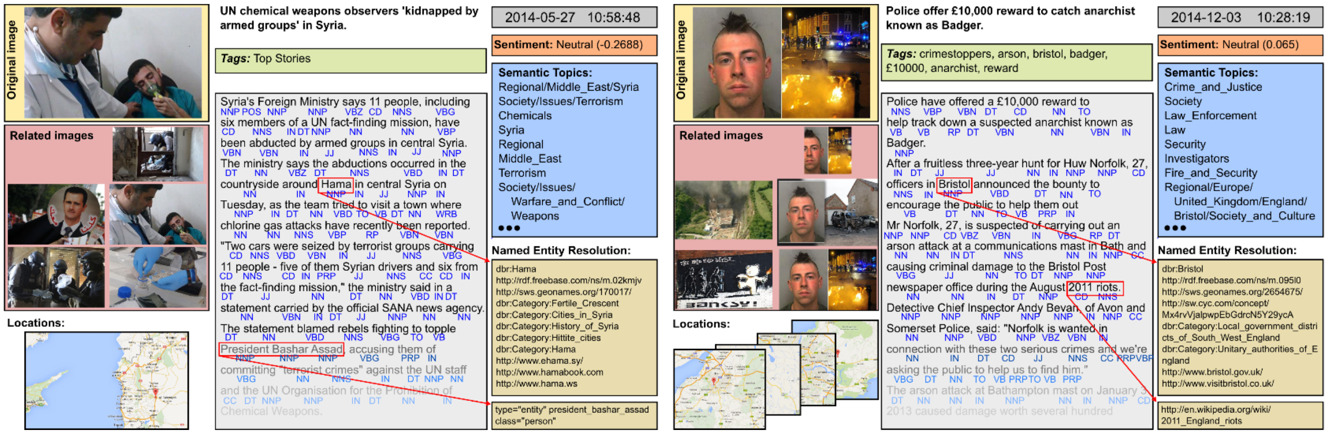

Original image
Related images
Locations:
UN chemical weapons observers 'kidnapped by armed groups' in Syria.
Tags: Top Stories
Syria's Foreign Ministry says 11 people, including
NNP POS NNP NNP VBZ CD NNS VBG
six members of a UN fact-finding mission, have
CD NNS IN DT NNP NN NN VBP
been abducted by armed groups in central Syria.
VBN VBN IN JJ NNS IN JJ NNP
The ministry says the abductions occurred in the
DT NN VBZ DT NNS VBD IN DT
countryside around Hama in central Syria on
NN IN NNP IN JJ NNP IN
Tuesday, as the team tried to visit a town where
NNP IN DT NN VBD TO VB DT NN WRB
chlorine gas attacks have recently been reported.
NN NN NNS VBP RP VBN VBN
"Two cars were seized by terrorist groups carrying
CD NNS VBD VBN IN JJ NNS VBG
11 people - five of them Syrian drivers and six from
CD NNS CD IN PRP JJ NNS CC CD IN
the fact-finding mission," the ministry said in a
DT JJ NN DT NN VBD IN DT
statement carried by the official SANA news agency.
NN VBN IN DT JJ NNP NN NN
The statement blamed rebels fighting to topple
DT NN VBD NNS VBG TO VB
President Bashar Assad, accusing them of
NNP NNP NNP VBG PRP IN
committing "terrorist crimes" against the UN staff
VBG NN NNS IN DT NNP NN
and the UN Organisation for the Prohibition of
CC DT NNP NNP IN DT NN IN
Chemical Weapons.
2014-05-27 10:58:48
Sentiment: Neutral (-0.2688)
Semantic Topics:
Regional/Middle_East/Syria
Society/Issues/Terrorism
Chemicals
Syria
Regional
Middle_East
Terrorism
Society/Issues/
Warfare_and_Conflict/
Weapons
•••
Named Entity Resolution:
dbr:Hama
http://rdf.freebase.com/ns/m.02kmjv
http://sws.geonames.org/170017/
dbr:Category:Fertile_Crescent
dbr:Category:Cities_in_Syria
dbr:Category:History_of_Syria
dbr:Category:Hittite_cities
dbr:Category:Hama
http://www.ehama.sy/
http://www.hamabook.com
http://www.hama.ws
type="entity" president_bashar_assad
class="person"
Original image
Related images
Locations:
Police offer £10,000 reward to catch anarchist known as Badger.
Tags: crimestoppers, arson, bristol, badger, £10000, anarchist, reward
Police have offered a £10,000 reward to
NNS VBP VBN DT CD NN TO
help track down a suspected anarchist known as
VB VB RP DT VBN NN VBN IN
Badger.
NNP
After a fruitless three-year hunt for Huw Norfolk, 27,
IN DT JJ JJ NN IN NNP NNP CD
officers in Bristol announced the bounty to
NNS IN NNP VBD DT NN TO
encourage the public to help them out
VB DT NN TO VB PRP IN
Mr Norfolk, 27, is suspected of carrying out an
NNP NNP CD VBZ VBN IN VBG RP DT
arson attack at a communications mast in Bath and
NN NN IN DT NNS NN IN NNP CC
causing criminal damage to the Bristol Post
VBG JJ NN TO DT NNP NNP
newspaper office during the August 2011 riots.
NN NN IN DT NNP CD NNS
Detective Chief Inspector Andy Bevan, of Avon and
NNP NNP NNP NNP NNP IN NNP CC
Somerset Police, said: "Norfolk is wanted in
NN NNP VBD NNP VBZ VBN IN
connection with these two serious crimes and we're
NN IN DT CD JJ NNS CC PRPVBP
asking the public to help us to find him."
VBG DT NN TO VB PRPTO VB PRP
The arson attack at Bathampton mast on January 3
DT NN NN IN NNP NN IN NNP CD
2013 caused damage worth several hundred
2014-12-03 10:28:19
Sentiment: Neutral (0.065)
Semantic Topics:
Crime_and_Justice
Society
Law_Enforcement
Law
Security
Investigators
Fire_and_Security
Regional/Europe/
United_Kingdom/England/
Bristol/Society_and_Culture
•••
Named Entity Resolution:
dbr:Bristol
http://rdf.freebase.com/ns/m.095l0
http://sws.geonames.org/2654675/
http://sw.cyc.com/concept/
Mx4rvVjalpwpEbGdrcN5Y29ycA
dbr:Category:Local_government_distri
cts_of_South_West_England
dbr:Category:Unitary_authorities_of_E
ngland
http://www.bristol.gov.uk/
http://www.visitbristol.co.uk/
http://en.wikipedia.org/wiki/
2011_England_riots

Fig. 1: **BreakingNews dataset**. The dataset contains a variety of news-related information including: the text of the article, captions, related images, part-of-speech tagging, GPS coordinates, semantic topics list or results of sentiment analysis, for about 100K news articles. The figure shows two sample images. All this volume of heterogeneous data makes BreakingNews an appropriate benchmark for several tasks exploring the relation between text and images.

tionally, each article is enriched with other data like related images from Google Images, tags, shallow and deep linguistic features (e.g. parts of speech, semantic topics or outcome of a sentiment analyzer), GPS latitude/longitude coordinates and reader comments. The articles cover the whole year 2014 and are collected from various reference newspapers and media agencies like BBC News, The Guardian or the Washington Post.

This dataset is an excellent benchmark for taking joint vision and language developments a step further. In contrast to existing datasets, the link between images and text in BreakingNews is not as direct, i.e., the objects, actions and attributes of the images may not explicitly appear as words in the text (see examples in Fig. 1). The visual-language connections are more subtle and learning them will require the development of new inference tools able to reason at a higher and more abstract level. Furthermore, besides tackling article illustration or image captioning tasks, the proposed dataset is intended to address new challenges, such as source/media agency detection or estimation of GPS coordinates.

For each of these tasks we benchmark several learning schemes based on state-of-the-art deep neural network architectures and different feature representations for text and images. For GPS regression and text and image correlation we propose novel loss functions, namely the Great Circle Distance and Deep Canonical Correlation Analysis. Overall results are very promising, but there is still much room for improvement, and the dataset can be easily extended with additional annotations to explore other problems such as visual question answering or text summarization. Both the baseline results we obtain and the dataset will be made publicly available, and we hope it will inspire future research in the field.

Overview: The rest of the paper is organized as follows. In Section 2 we present the different vision and language processing tasks that will be addressed in this paper, with a thorough review of the related work for each case. In Section 3 we introduce *BreakingNews*, the news article dataset. The following two sections, describe the technical details about how the visual and textual data is represented (Section 4), and which CNN architectures we have built to tackle each of the tasks (Section 5). A extensive evaluation is reported in Section 6.

2 Tasks Description and Related Work

We next describe the tasks that will be tackled in this article and the related work for each of them, as well as the existing datasets.

2.1 Source detection

This tasks deals with analyzing the content of the news articles, and detecting in which news media agency it has been originally published. It can also be used for more sociological-oriented research. As an example, we tackle the task of quantitatively assessing the intuition that different news agencies have their clear, distinctive style. This problem can be addressed by modeling the type of language, images and topical preference for each news agency, but also from correctly modeling the sentiment in each news article based on the political orientation of the agency, or other similar refinements.

Although we are not aware of other approaches explicitly tackling source identification in news articles, there exists a vast amount of related works, mostly motivated by detection of plagiarism or dealing with the authorship identification problem [44], [58], [72].

2.2 Text Illustration

This task deals with automatically retrieving the appropriate image (or a small subset) from a pool of images given a textual query of a news story.

Some approaches tackle this problem by learning classifiers to represent images through intermediate semantic concepts, that can then be easily assigned to individual keywords or to multi-attribute textual descriptions [3], [4], [45], [48], [67], [17]. Richer textual queries are allowed in [21], [31], [29], [19], [76] by mapping both image and sentences to intermediate spaces where direct comparison is possible. For all these methods, the textual input consists on keywords or short sentences at most.

There have been some efforts specifically addressing the domain of news articles. For instance [22] builds a system based on joint topic models to link text to images, and evaluates the results in the BBC News dataset [23]. In [9], it is assumed that there exist short text descriptions and tags accompanying the images, which can then be easily matched to text documents represented by means of word frequencies. Other approaches perform article illustration by exploiting Google's search engine (which also assumes text associated with each image of the database), and combine multiple queries generated from the title of the article [50], or from the narrative keywords [34]. Similarly, [38] proposes a story picturing engine, that first processes a story to detect certain keywords, and then uses these to select annotated images from a database.

Alternatively to image retrieval strategies, text illustration can be carried out from a generative perspective. There exist early approaches that extracted object arrangements from sentences to then generate computer graphic representations of static [11] and dynamic [64] scenes. [25] proposed a system to animate a human avatar based on the emotions inferred from text. And very recently, advanced sentence parsers have been used to extract objects and their relations from sentences, and then automatically render 2D [88] and 3D [6] scenes.

Finally, illustration of short texts, like chat messages or tweets, has also been investigated [36], [78].

2.3 Geolocation

This task considers the problem of geographically referencing news articles based on text and image cues. One of the pioneering works on a related topic proposed an approach for geolocating web documents by detecting and disambiguating location names in the text [15]. Also only using text information, [68] matched a predefined tagged map to the most likely GPS location of tagged Flickr photos. On the other hand, several image-based methods leverage massive datasets of geotagged images for geolocation of generic scenes on the global Earth scale [28], [40], [79], or for place recognition tasks [7]. There has been also work on this area by combining text and image information [5], [12]. For the specific domain of news articles, most existing works use only text information [50], [80]. One interesting exception is [87], which combines textual descriptors with global image representations based on GIST [59] to infer geolocation of nearly two thousand NY Times articles.

2.4 Caption Generation

Among the tasks dealing with image and text interactions, automatically generating photo captions is the one receiving most attention. Early work in this area focused in annotating images with single words [3], [70], phrases with a reduced vocabulary [21], [47] or with semantic tuples [60], [65]. The problem has also been tackled from a ranking perspective, in which given a query image, one needs to rank a set of human generated captions. Projecting images and captions to a joint representation space using Kernel or Normalized Canonical Correlation Analysis has been used for this purpose [26], [31]. The largest body of recent work, though, share a basic approach, which consists in using a Recurrent Neural Network (RNN) to learn to "translate" the image features into a sentence in English, one word at a time [8], [16], [20], [37], [41], [43], [54], [75], [82].

Yet, while the results obtained by the RNN-based approaches are outstanding, they are all focused on accurately describing the content of the picture, which is significantly different than generating a suitable caption for the illustration of a news article, in which the relation between the visual content of the image and the corresponding captions is more indirect and subtle than in typical image captioning datasets. This is precisely the main challenge posed by the BreakingNews dataset we present. There exists some previous work in automatic caption generation for images of the BBC News Dataset [23]. However, as we will discuss later, BreakingNews is about two orders of magnitude larger than BBC News Dataset, and incorporates a variety of metadata that brings the opportunity to develop new Computer Vision approaches for the analysis of multimodal and large-scale multimedia data.

2.5 Image and Text Datasets

There is no doubt that one of the pillars on which the recent advent of Deep Learning holds is the proliferation of large object classification and detection datasets like ImageNet [13], PASCAL VOC 2012 [18] and SUN [81]. Similarly, the progress on the joint processing of natural language and images largely depends on several datasets of images with human written descriptions. For instance, the UIUC Pascal Sentence Dataset [66] consists of 1,000 images each annotated by 5 relatively simple sentences, even lacking verb in a large percentage of them. The SBU photo dataset [61] consists of one million web images with one description per image. These descriptions are automatically mined and do not always describe the visual content of the image. Flickr8K [31], Flickr30K [84] and MS-COCO [51] contain five sentences for a collection of 8K, 30K and 100K images, respectively. In these datasets, the sentences specifically describe the visual content and actions occurring in the image.

Relevant work has also been done in the area of Visual Question Answering VQA [2]. DAQUAR [53], Visual Madlibs [85] and KB-VQA [77] were designed

News article tags
World news, UK news, Business, Politics, Society, Australia, Life and style, Sport, United States, World, Europe, Health, Entertainment, Australian politics, Nation, Football, Culture, Middle East and North Africa, Asia Pacific, Comment, Media, Sports, Law, Soccer, Environment, Africa, Labour, Conservatives, Russia, Blogposts, London, David Cameron, Features, Australia news, Technology, Money, Scotland, China, Food and drink, Ukraine, Education, NHS, Women, Sci/Tech, Tony Abbott, US politics, European Union, Editorial, Iraq, Economics, Barack Obama, Books, US news, Syria, Children, Music, Religion, Premier League, Film, Top Stories, Family, Scottish independence, US foreign policy, Science, Ed Miliband, Crime, Liberal Democrats, France, England, Internet, Retail industry, Israel, Television, Race issues, Labor party, Police, Economic policy, New South Wales, Gender, Victoria, Vladimir Putin, Local government, UK Independence party (Ukip)

TABLE 1: Some popular tags associated to news articles in the dataset.

Source	num. articles	avg. len. article	avg. num. images	avg. len. caption	avg. num. comments	avg. len. comment	avg. num. shares	% geo-located
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 2: BreakingNews dataset statistics. Mean and standard deviation, usually rounded to the nearest integer.

specifically for VQA and evaluating algorithms capable of answering high level questions and reasoning about image contents using external information. The goal is slightly different than for analysis of illustrated articles as it typically involves complex reasoning on the image content, however it is also similar in a sense that VQA require specific information about the image for which generic image captions are of little use.

In this paper, we focus on illustrated news articles datasets, which differ from previous image and text datasets in that the captions or the text of the article do not necessarily describe the objects and actions in the images. The relation between text and images may express connotative relations rather than specific content relations. The only dataset similar to BreakingNews is the BBC News dataset [23], which contains 3,361 articles with images, captions and body text. However, this dataset is mainly intended for caption generation tasks although it is small for training data-hungry models like LSTM recurrent neural networks, that are the state-of-the-art techniques for caption generation. Furthermore, in addition to the difference in size, the BBC News does not include geolocation information, tags, social network shares or user comments, and all articles come from a single source. In BreakingNews, we provide a two order of magnitude larger dataset, that besides images and text, contains other metadata (GPS location, user comments, semantic topics, etc) which allows exploring a much wider set of problems and exploit the power of current Deep Learning algorithms. After the initial submission of this work, another large-scale news dataset was published: the ION corpus [32]. However, in contrast with BreakingNews, it includes only the article text, images and captions.

3 BreakingNews Dataset

We have come a long way since the days when a news article consisted solely of a few paragraphs of text: online articles are illustrated by pictures and even videos, and readers can share their comments on the story in the same web-page, complementing the original document.

Studying the effects and interactions of these multiple modalities has clear interesting applications, such as easing the work of journalist by automatically suggesting pictures from a repository, or determining the best way to promote a given article in order to reach the widest readership. Yet, no benchmarks that capture this multi-modality are available for scientific research.

For these reasons, we propose a novel dataset of news articles[2] with images, captions, geolocation information and comments, which we will use to evaluate CNN and LSTM architectures on a variety of tasks. Our models are based on the most recent state-or-the-art approaches in deep learning, and thus, this dataset is intended to be a touchstone to explore the current limits of these methodologies, shown to be very effective when dealing with images associated with visually descriptive text.

3.1 Description of the Dataset

The dataset consists of approximately 100,000 articles published between the 1st of January and the 31th of December of 2014. All articles include at least one image, and cover a wide variety of topics, including sports, politics, arts, healthcare or local news. Table 1 shows some of the most popular topics. The main text of the articles was downloaded using the IJS newsfeed [73], which provides a clean stream of semantically enriched news articles in multiple languages from a pool of *RSS* feeds. We restricted the articles to those that were written in English and originated from a shortlist of highly-ranked news media agencies (see Table 2) to ensure a degree of consistency and quality. Given the geographic distribution of the news agencies, most of the dataset is

[2] The BreakingNews dataset is publicly available at http://www.iri.upc.edu/people/aramisa/BreakingNews/index.html

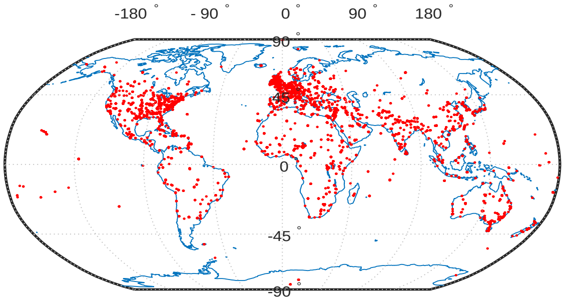


Fig. 2: Ground truth geolocations of the articles.

made of news stories in English-speaking countries in general, and the United Kingdom in particular.

For each article we downloaded the images, image captions and user comments from the original article web-page. News article images are quite different from those in existing captioned images datasets like Flickr8K [31] or MS-COCO [51]: often include close-up views of a person (46% of the pictures in BreakingNews contain faces) or complex scenes. Furthermore, news image captions use a much richer vocabulary than in existing datasets (e.g. Flickr8K has a total of 8,918 unique tokens, while eight thousand random captions from BreakingNews already have 28,028), and they rarely describe the exact contents of the picture.

We complemented the original article images with additional pictures downloaded from Google Images, using the full title of the article as search query. After manually verifying the results of a large subset of articles, we found that using the title as the query term was sufficiently precise and more stable than the full article text. The five top ranked images of sufficient size in each search were downloaded as potentially related images (the original article image usually appears among them).

Interestingly, the articles are annotated with user comments and the number of shares on different social networks (e.g. Twitter, Facebook, LinkedIn) if this information was available. Whenever possible, in addition to the full text of the comments, we recovered the thread structure, as well as the author, publication date, likes (and dislikes) and number of replies. Since there were no share or comments information available for "The Irish Independent", we searched Twitter using the full title of the articles, and collected the tweets with links to the original article, or that mentioned a name associated with the newspaper (e.g. @Independent_ie, Irish Independent, @IndoBusiness) in place of comments. We considered the collective number of re-tweets as shares of the article. Even though in this paper we do not specifically exploit this type of annotation, we believe it would be useful for methods analysing article popularity.

The IJS Newsfeed annotates the articles with geolocation information both for the news agency and for the article content. This information is primarily taken from the provided RSS summary, but sometimes it is not available and then it is inferred from the article using heuristics such as the location of the publisher, TLD country, or the story text. Fig. 2 shows a distribution of news story geolocation.

Finally, the dataset is annotated with shallow and deep linguistic features (e.g. part of speech tags, inferred semantic topics, named entity detection and resolution, sentiment analysis) with the *XLike* (http://www.xlike.org/language-processing-pipeline/) and the *Enrycher* (http://ailab.ijs.si/tools/enrycher/) NLP pipelines.

4 Representation

In this section we discuss text and image representations for news article analysis. We first present Bag-of-Words and Word2Vec text embeddings and then image representations based on deep CNNs.

4.1 Text representation

Bag-of-Words (BoW) with TF-IDF weighting: Bag-of-Words is one of the most established text representations, and their reliability and generalization capabilities have been demonstrated in countless publications. Therefore, we adopt it as a baseline for our approach. The BoW representation requires a vocabulary, which we established as unique lemmatised tokens from the training data that appear more than L times in the articles. This leads to D_b dimensional Bag-of-Words (BoW) vectors where the j^{th} dimension is given by $t^j \log \frac{M}{c^j+1}$, where t^j is the frequency of the j^{th} token (i.e. the number of times it appears in the article), c^j is document frequency of the token (i.e. the number of training articles where it appears), and M is the total number of training articles. A common practice is to truncate the BoW vector based on inverse document frequency. It has been demonstrated that performance typically improves monotonically with the number of retained dimensions.

Word2Vec: Recently, distributed representations for text, and *Word2vec* [49], [56], [57] in particular, are gaining a lot of attention in the NLP community. This representation encodes every word in a real-valued vector that preserves semantic similarity, e.g. "king" minus "man" plus "woman" will be close to "queen" in the embedding space. Using a two-layer neural network, words are modeled based on their context, defined as a window that spans both past and future words. Two methods have been proposed to learn the representations: the *Continuous bag of words* (cbow) model, where the objective is predicting a word given its context, and the *Continuous skip-gram* model, where the objective is the opposite; i.e. trying to predict the context given the word being modeled. The *negative sampling* objective function, where the target word has to be distinguished from random negative words, is used as an efficient alternative to hierarchical soft max. We investigated both cbow and skip-gram methods[3] and found that the later performed better in our applications.

[3] Code available at https://code.google.com/p/word2vec/

The trained model is used to encode each article by stacking the representation for every word and symbol in a matrix with dimension D_w times the number of tokens in the article, where D_w is the size of the embedding space. From this matrix we investigate three representations: 1) full matrix as input for the convolutional layers of a deep network; 2) mean or 3) max along the sentence dimension to construct a single D_w dimensional vector for each article. Another possible strategy would be to first apply pooling independently for each sentence, and next for all the sentence representations in a single document vector. We have however observed in early experiments that allowing the system to learn from all features and their ordering by using a convolutional layer, performs better than pooling the representations beforehand.

4.2 Image representation

Our image representations are based on pre-trained deep CNNs. We follow the state-of-the-art CNN representation developed for object recognition [69] with 19 convolutional layers (VGG19). More specifically, we compute the activations of the last ReLU layers of the VGG19 and the Places scene recognition CNN [86]. Each of these activations are 4,096 dimensional sparse vectors, and have been shown to perform well in various vision tasks. We also ℓ_2 normalize the two vectors and concatenate them to form a third image representation of size 8,192.

5 Learning

In this section we describe the learning schemes based on deep neural networks that will be used for the proposed tasks. We first discuss in Section 5.1 a CNN for article illustration, source detection, and geolocation prediction. In Section 5.2 we extend the state-of-the-art LSTM approach for caption generation.

5.1 Article analysis

CNN architecture: CNNs have recently proven successful in many NLP tasks such as sentiment analysis, machine translation or paraphrase identification [10], [39], [42]. We consider a CNN architecture, as illustrated in Fig. 3-a1, for various article analysis tasks. In the figure, dashed boxes are the inputs to the network, and solid boxes are the layers. Moreover, white boxes denote data and layers that are shared by all tasks (we denote the group by "Shared Textual CNN"), while shaded ones are task specific.

Assuming Word2Vec embeddings have been learnt, an article can be represented by a $D_w \times n_i$ matrix, where D_w is the dimensionality of the embedding space, and n_i is the number of tokens in the i^{th} article. Such matrix is zeros-padded to $D_w \times N$, where N is the number of tokens in the longest article (we add $N - n_i$ columns of zeros). The *convolution* layer convolves the Word2Vec matrices with 256 rectangular kernels of size $D_w \times 5$ at a stride of 1, capturing local correlations in the tokens. The output of the convolution layer are 256 feature channels of dimension $N - 4$ each. Effectively, the convolutional kernels act as learnable feature detectors.

The activations of the convolution layer are max-pooled (*pooling*) along each channel. The resulting 256 dimensional vectors are transformed into 64 channels in a fully connected linear layer (*FC*), before *nonlinearity* is applied. Although Tanh or Absolute nonlinearities were observed to have a small accuracy advantage, we found that ReLU [46] is numerically more stable in terms of gradient computation.

Dropout has been proven a simple yet effective way of preventing overfitting in large networks [71]. We found that a small dropout ratio (e.g. 0.1) is sufficient for our tasks. Finally, after dropout, the 64 dimensional vectors are transformed in a second linear layer (*FCo*) to vectors with appropriate length. Let the output of *FCo* be $\mathbf{z} \in \mathbb{R}^d$ for each article, which will then be used in a task specific loss function, together with a task specific label. The gradient of the loss with respect to $\mathbf{z}$ will be computed and backpropagated to update the CNN parameters. The parameters of the CNN architecture for various tasks are given in Fig. 3-a1.

The network described above takes as input textual features i.e. Word2Vec embeddings. We also propose the simple nonlinear network in Fig. 3-a2 that takes image features as input, including VGG19, Places, and the concatenation of the two. Moreover, in Fig. 3-a3 the text CNN and image CNN are combined in a concatenation layer. Fig. 3-a1,a2,a3 allow us to evaluate the performance of various features and their combinations in the context of different tasks.

We next describe the configuration details of the last layer *FCo* and the loss functions (and when necessary their corresponding derivatives), that make the CNN architecture of Fig. 3-a1,a2,a3 problem specific.

Source detection: For source detection, the label y is in the set $\{1, \cdots, d\}$, being d the number of news agencies where the articles originate from. The *loss* layer implements the multinomial logistic loss that maximizes the cross entropy, whose gradients for backpropagation are readily available.

Article illustration: For article illustration, we use the image representation discussed in Section 4.2 as the label $\mathbf{y} \in \mathbb{R}^{8,192}$, which is a concatenation of ℓ_2 normalized VGG19 and Places activations. The output of the *FCo* layer $\mathbf{z}$ also has a dimensionality $d = 8,192$. For a batch of m data points, the inputs to the *loss* layer are two $d \times m$ matrices Z and Y (see [83] for details).

Let the covariances of Z and Y be Σ_{zz} and Σ_{yy} respectively, and let the cross covariance be Σ_{zy}, all of them $\mathbb{R}^{d\times d}$ matrices. A good loss function for article illustration is the Canonical Correlation Analysis (CCA) loss [33], [27], which seeks pairs of linear projections $\mathbf{w}_z, \mathbf{w}_y \in \mathbb{R}^d$ that maximize the correlation between the

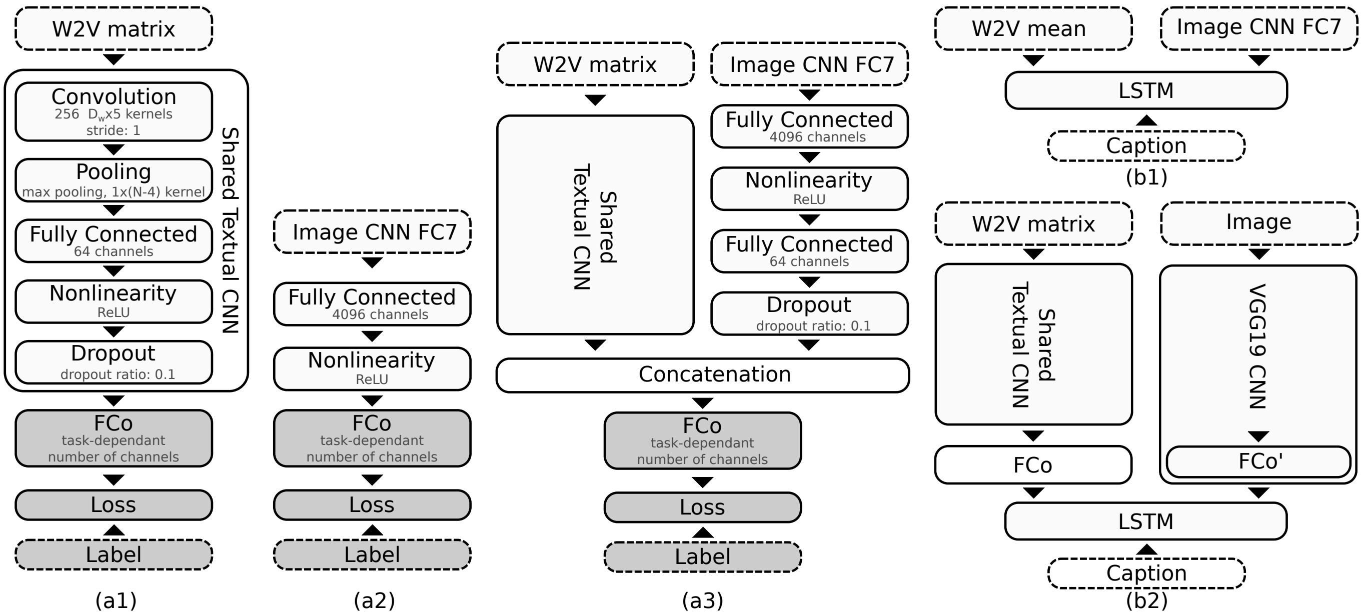


Fig. 3: **CNN and LSTM architectures for article analysis and caption generation.** (a1-3) CNN for article analysis. Dashed boxes are the CNN inputs, and solid boxes correspond to the layers. White boxes (enclosed within the "Shared Textual CNN" box) are shared by all tasks, and shaded boxes are task specific layers. (b1, b2) Models for caption generation: (b1) LSTM with fixed features, and (b2) end-to-end learning with CNNs+LSTM. Note that the "Shared Textual CNN" from (a1) is used in the textual branch.

articles and the images:

$$\begin{aligned}(\mathbf{w}_z^*, \mathbf{w}_y^*) &= \underset{\mathbf{w}_z, \mathbf{w}_y}{\arg\max}\ \operatorname{corr}(\mathbf{w}_z^\top Z, \mathbf{w}_y^\top Y) \\ &= \underset{\mathbf{w}_z, \mathbf{w}_y}{\arg\max} \frac{\mathbf{w}_z^\top \Sigma_{zy} \mathbf{w}_y}{\sqrt{\mathbf{w}_z^\top \Sigma_{zz} \mathbf{w}_z \mathbf{w}_y^\top \Sigma_{yy} \mathbf{w}_y}} \qquad (1)\end{aligned}$$

Using the fact that the objective is invariant to scaling of $\mathbf{w}_x$ and $\mathbf{w}_y$, and assembling the top projection vectors into the columns of projection matrices W_z and W_y, the CCA objective can be written as:

$$\begin{aligned}&\max_{W_z, W_y} \ \operatorname{tr}(W_z^\top \Sigma_{zy} W_y) \qquad (2)\\ &\text{s.t.}: W_z^\top \Sigma_{zz} W_z = W_y^\top \Sigma_{yy} W_y = I_d\end{aligned}$$

where I_d is the $d-$dimensional identity matrix.

Let us define $\Gamma = \Sigma_{zz}^{-1/2} \Sigma_{zy} \Sigma_{yy}^{-1/2}$, and let U_k and V_k be the matrices of the first k left- and right- singular vectors of Γ, respectively. In [1], [55] it is shown that the optimal objective value is the sum of the top k singular values of Γ, and the optimum is attained at

$$(W_z^*, W_y^*) = (\Sigma_{zz}^{-1/2} U_k, \Sigma_{yy}^{-1/2} V_k) \qquad (3)$$

When $k = d$, the total correlation objective in Eq. (2) is equal to the trace norm of Γ:

$$\operatorname{corr}(Z, Y) = ||\Gamma||_{\text{tr}} = \operatorname{tr}((\Gamma^\top \Gamma)^{1/2}) \qquad (4)$$

We define the CCA-based loss as $L = -\operatorname{corr}(Z, Y)$. If we consider the singular value decomposition (SVD) of Γ to be $\Gamma = UDV^\top$, recent works [1], [83], have shown that the gradient of L with respect to Z is given by:

$$\frac{\partial L}{\partial Z} = -\frac{1}{m-1}(2\nabla_{zz}\bar{Z} + \nabla_{zy}\bar{Y}) \qquad (5)$$

where $\bar{Z}$ and $\bar{Y}$ are the centered data matrices, and

$$\begin{aligned}\nabla_{zz} &= -\tfrac{1}{2}\Sigma_{zz}^{-1/2} U D U^\top \Sigma_{zz}^{-1/2} \qquad (6)\\ \nabla_{zy} &= \Sigma_{zz}^{-1/2} U V^\top \Sigma_{yy}^{-1/2} \qquad (7)\end{aligned}$$

Note that with CCA loss we do not directly predict the image feature vector, instead, we maximise the correlation of the text and the image features (after CNN layers).

Geolocation prediction: For geolocation prediction, the label $\mathbf{y} = [y_1, y_2]^\top$ is a pair of latitude and longitude values, where the latitude $y_1 \in [-\pi/2, \pi/2]$ and the longitude $y_2 \in [-\pi, \pi]$. Fig. 2 illustrates the ground truth geolocations in the training set.

The output of the *FCo* layer $\mathbf{z} \in \mathbb{R}^2$ can be thought of as the predicted latitude and longitude. In the *loss* layer, we could minimize the Euclidean distance between the two geolocations $\mathbf{z}$ and $\mathbf{y}$. However, the Euclidean loss does not take into account the fact that the two meridians near $-\pi$ and π respectively are actually close to each other. To address this, we minimize the Great Circle Distance (GCD) between $\mathbf{z}$ and $\mathbf{y}$, which is defined as the geodesic distance on the surface of a sphere, measured along the surface. We use the spherical law of cosines approximation of GCD:

$$\text{GCD} = R \cdot \arccos(\sin y_1 \sin z_1 + \cos y_1 \cos z_1 \cos\delta) \qquad (8)$$

where $\delta = z_2 - y_2$, and $R = 6,137$ km is the radius of the Earth. Ignoring constant R we define our geolocation loss function as

$$L = \arccos(\sin y_1 \sin z_1 + \cos y_1 \cos z_1 \cos\delta) \qquad (9)$$

Using the chain rule, the gradient of L with respect to $\mathbf{z}$ can be shown to be:

$$\frac{\partial L}{\partial \mathbf{z}} = \begin{pmatrix} -\frac{1}{\sqrt{1-\phi^2}}(\sin y_1 \cos z_1 - \cos y_1 \sin z_1 \cos \delta) \\ -\frac{1}{\sqrt{1-\phi^2}}(-\cos y_1 \cos z_1 \sin \delta) \end{pmatrix} \quad (10)$$

where $\phi = \sin y_1 \sin z_1 + \cos y_1 \cos z_1 \cos \delta$. In practice, to ensure numerical stability, the term $1 - \phi^2$ is discarded when it is too close to zero.

5.2 Caption generation

Image and video captioning has made significant progress in the last few years. State-of-the-art approaches use CNNs to encode image and video as a fixed-length vectors, and employ recurrent neural networks (RNNs) in the particular form of Long Short-Term Memory Networks (LSTMs) [30] to decode the vector into a caption [16], [75], [82]. By using an input gate, an output gate and a forget gate, LSTMs are capable of learning long-term temporal dependences between important events. Moreover, LSTMs successfully deal with the vanishing and exploding gradients problem by keeping the errors being backpropagated in their memory.

In this paper we consider the following LSTM-based architectures to tackle the caption generation problem.

LSTM with fixed features: The task of news image captioning differs from most existing works on image captioning in two aspects. First, captions for news images are often only loosely related to what is illustrated in the images, making it much more challenging to learn the mapping between images and captions. Consequently, the approaches designed to improve the description of image, object and action such as the attention mechanism [82] were not used as a baseline here. Second, instead of conditioning only on images, news captions are conditioned both on the articles and images. To address the second difference, we extend the recent techniques in [16], [75], [82]. We take the mean of a Word2Vec matrix as the fixed article representation, and VGG19 and/or Places activations as the fixed image representation, and train an LSTM using either representation, or a concatenation of the two. The architecture of such a scheme is illustrated in Fig. 3-b1.

End-to-end learning with CNNs+LSTM: The language CNN described earlier in this section (Fig. 3-a1), the VGG19 (or Places) image CNN, and the LSTM for caption generation are all neural networks trained with backpropagation. We propose to connect the three networks as illustrated in Fig. 3-b2, achieving end-to-end learning of caption generation from raw articles and raw images. In this setting, the VGG19 CNN and the language CNN have *FCo′* and *FCo* as their last layers respectively, which encode image and article information into a 500 dimensional vector each. The VGG19 CNN is also filled with parameters pretrained on ImageNet. During training, the gradients in LSTM are propagated backward in time, and backward into the two CNNs.

In addition to the LSTM approaches, we also experimented with extractive captioning. To generate a new caption, we selected the sentence from the article that is most compatible with the image according to the CCA projection, as presented for the article illustration. Sentences shorter than 5 words or longer than 50 words were discarded. We termed this approach CCA-Sentence selection (CCA-SS).

6 Results

In this section, we present experimental results on the four tasks we considered. We first discuss CNN results for source detection, geolocation prediction and article illustration, followed by a discussion on LSTM and a mixed LSTM/CNNs model for caption generation.

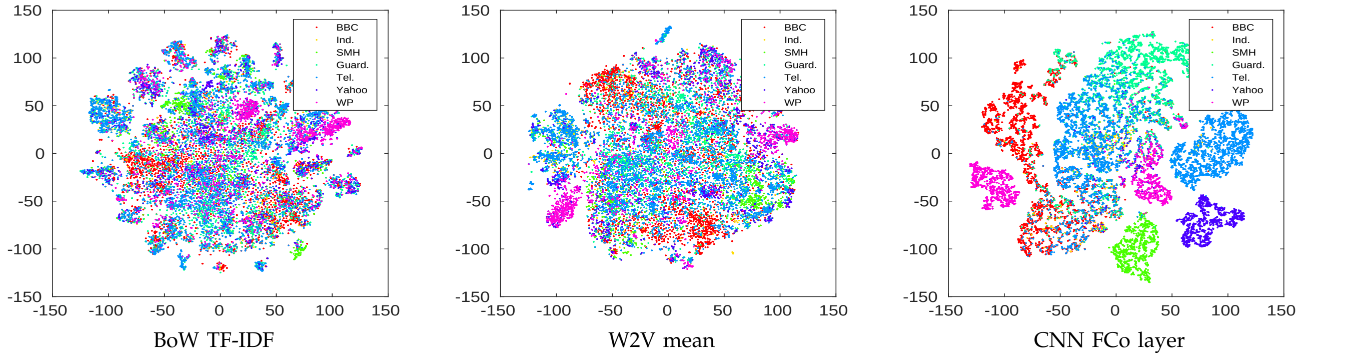
6.1 Implementation Details

Before describing the experimental results, we discuss the technical aspects related to the preparation of the dataset and to the implementation details, including the hyper-parameters set-up for representation and learning.

Dataset considerations for the experimental setup: The BreakingNews dataset is split into train, validation and test sets with 60%, 20% and 20% articles respectively. To ensure fairness in the experiments, we also checked there was almost no overlap of images between the sets using the VGG19 features and a cosine distance. This ratio of near-identical image pairs was in the order of 10^{-6}.

Textual representations: In total there are 47,677 unique tokens with frequency 5 or more in the training set (out of 115,427), so the BoW representation is $D_b = 47{,}677$ dimensional. Regarding the captions, the BoW size is 13,507 (out of 89,262 unique terms). For Word2Vec, we used the BreakingNews training set with the skip-gram method to learn the embedding space. The size of the embedding space was $D_w = 500$, the window size was 30, the sampling rate was set to 1e-5, and the number of negative samples was set to 5. These hyper-parameters were chosen based on the article illustration task for a small subset. We also experimented with a publicly available Word2Vec model trained with the Google 100 billion words dataset, but the performance was worse. Furthermore, we considered using Glove [63] and Doc2Vec [49] instead of Word2Vec. However, we have found these embeddings to perform worse in our dataset. For instance, for the illustration task, when trained on a subset with 5,000 pairs and evaluated on a subset with 1,000 pairs, the Word2Vec features have a median rank (MR) of 60 (lower is better), while Glove features have 100 and Doc2Vec features 200. We therefore focused on Word2Vec features in other tasks.

Article analysis: The proposed CNN architecture including our novel GCD and CCA losses were implemented using the Caffe framework [35]. Unless specified otherwise, in the experiments we used a base learning rate of



BoW TF-IDF

W2V mean

CNN FCo layer

Fig. 4: Source detection: t-SNE embedding of shallow and deep features for the articles in the test set.

0.05, which dropped by a factor of 0.1 after every 1,000 iterations. Stochastic gradient descent was employed as the solver, for which the momentum was set to 0.9. The regularisation parameter weight decay was set to 0.0005. The results reported in this paper were obtained with a batch size of $m = 64$, which was cross-validated on the validation set. For the illustration task the dimensionality of the projection space was also cross-validated, and the cosine distance was used to measure the distance in the projection space. While exhaustive search was not realistic, we experimented as much as possible with network architectures and hyper-parameters of the various layers. For instance, for the convolutional layer we experimented with size and stride, and found that a kernel width of 5 and a stride of 1 produced the best performance. For the source prediction task, with textual features, using a kernel size of 500×7 instead of 500×5 reduced the accuracy by $\sim$5 points, and a kernel size of 500×3 instead of 500×5 reduced it by $\sim$2 points. On the other hand, a stride of 2 and 5 instead of 1 decreased the accuracy by $\sim$1 and $\sim$5 points, respectively. All our experiments were performed on an NVIDIA Tesla K40C GPU with 12G memory. It takes approximately 10 hours for training to converge. Depending on the task, the total number of parameters learnt in the CNN ranges from approximately 0.7M to 1.2M.

LSTMs for caption generation: The two architectures in Fig. 3-b1,b2 were implemented using Neuraltalk2[4]. We set both the input encoding size and LSTM memory size to 256, the learning rate to 0.0002, the batch size to 4, and kept all other parameters default. On the NVIDIA K40C GPU it typically took several days for the training to converge. The total number of parameters was approximately 50M and 200M for the two architectures.

6.2 CNNs for article analysis

Source detection: For this task we compare the proposed CNN trained with the full $500 \times 8,000$ Word2Vec matrix with the popular shallow learner linear SVM. For the latter, we consider four versions of shallow textual features: the full 47,677 dimensional TF-IDF weighted BoW feature; its truncated version with 5,000 dimensions; the mean of Word2Vec matrix; and the max of Word2Vec matrix (direct usage of the full matrix was discarded right away as impractical and not robust to sentence variations). Additionally, note that the source detection problem is unbalanced, as can be seen in Table 2. We therefore use two performance metrics: the overall accuracy, and the balanced accuracy which is the average of per-class accuracies.

Results in Table 3 show that the full BoW feature outperforms its truncated version, and the mean of Word2Vec is much better than the max version. When comparing the two types of text features, BoW's small advantage may be attributed to its much higher dimensionality (47,677/5,000 vs. 500). The proposed CNN architecture produces accuracies significantly higher than the best shallow feature. As expected, image features alone have lower performance than the text features, yet they still produce surprisingly good results, indicating that the different media agencies have distinctive ways of illustrating their articles. Another noteworthy conclusion is that the shallow learners benefit more from the additional data than the CNN, which seems to already have saturated its performance by fully utilizing the text information.

In Fig. 4 we plot the 2D t-SNE [52] embedding of the shallow features and the output of the *FCo* layer of the CNN for the test set, where the points are colour-coded with class labels. It is evident that the deeply-learnt representation in CNN provides a much better separation of the different classes.

Article illustration: We compare the CNN with the Canonical Correlation Analysis (CCA), which is a standard shallow learner for text and image matching [26], [31]. Given textual and visual representations for a pair of article and image, CCA finds the optimal projection that maximises the correlation between the two modalities. For textual representation on the CCA we again consider both BoW and Word2Vec mean; and for image representation we use activations of the pretrained VGG19 and Places models, and the concatenation of the two.

[4] https://github.com/karpathy/neuraltalk2

Source Detection				
Image feat.	**Text feat.**	**Learning**	**Acc.**	**Bal. acc.**
-	BoW TF-IDF	SVM	72.9	72.6
-	BoW TF-IDF 5K-d	SVM	70.8	70.0
-	W2V mean	SVM	68.6	64.4
-	W2V max	SVM	57.3	53.4
VGG19	-	SVM	45.4	33.7
Places	-	SVM	41.8	30.6
VGG19+Places	-	SVM	48.3	37.4
VGG19+Places	BoW TF-IDF	SVM	76.3	75.9
-	W2V matrix	CNN	81.2	80.7
VGG19+Places	-	CNN	36.4	21.0
VGG19+Places	W2V matrix	CNN	**81.8**	**80.9**

TABLE 3: Results of source detection. Due to the unbalance between the classes, balanced accuracy is also reported. CNN shows a clear advantage over shallow learners for this task.

Text to image					
Image feat.	**Text feat.**	**Learning**	**R@1**	**R@10**	**MR**
VGG19	BoW TF-IDF	CCA	5.7	14.6	448
Places	BoW TF-IDF	CCA	5.1	12.1	801
VGG19+Places	BoW TF-IDF	CCA	**6.0**	**15.3**	**397**
VGG19	W2V mean	CCA	2.4	9.6	499
Places	W2V mean	CCA	2.2	7.9	820
VGG19+Places	W2V mean	CCA	3.5	11.9	415
VGG19+Places	W2V mean	CNN	3.3	12.3	411
Text to related images					
Image feat.	**Text feat.**	**Learning**	**R@1**	**R@10**	**MR**
VGG19	W2V mean	CCA	2.9	14.5	161
Places	W2V mean	CCA	2.4	11.8	257
VGG19+Places	W2V mean	CCA	**3.7**	**16.7**	**137**
Text to image+caption					
Image feat.	**Text feat.**	**Learning**	**R@1**	**R@10**	**MR**
VGG19+Orig.	W2V mean	CCA	25.0	55.6	7
VGG19+Anon.	W2V mean	CCA	11.8	33.2	42

TABLE 4: Article Illustration experiments. Three setups to evaluate different alternatives (learning approaches and features) for automatically illustrating images, and exploring the problem.

For each test article we rank the 23,523 test images according to the projection learnt in CCA or CNN, and measure the performance in terms of recall of the ground truth image at the k^{th} rank position (R@k), and the median rank (MR) of the ground truth. Note that a higher R@k or a lower MR score indicates a better performance.

Results in the top of Table 4 demonstrate that when CCA is used, with either textual representation the VGG19-based visual feature outperforms Places-based feature by a significant margin; while combining the two further improves the performance. Comparing the two textual representations, BoW with TF-IDF again has an edge over the mean of Word2Vec.

The CNN only manages to marginally outperform its shallow counterpart (CCA with Word2Vec mean and VGG19+Places). We hypothesize that this is because in news articles, the text is only loosely related to the content of the associated image. This is in contrast to standard image/text matching datasets such as MS-COCO [51] and Flickr30k [26], where the associated caption explicitly describes the content of the image. In any event, note that the MR values are around 400, which, considering a pool of more than 23K images, are very remarkable results.

In order to better understand the nature of the problem and the challenges of the proposed dataset we performed two additional experiments. For efficiency's sake, in these experiments we just considered the CCA, which can be trained much faster (requires computing an SVD of a matrix of size the number of training images) than the back-propagation scheme needed for the CNN.

In the first setting, the related images collected from the Internet replace the original image associated with the article. Visual features are extracted from the related images and averaged as the visual representation. This is used for both the train and test splits. As can be seen in the Table 4-middle, adding multiple views to the image representation significantly outperforms the results . For example, the MR decreases from 499 to 161 for the W2V mean/VGG19 feature combination. This shows that the original images are weakly correlated with the actual articles, and this correlation is improved by considering additional Internet images.

Along the same lines, in the second setting we assumed the images were already equipped with captions. The W2V embeddings were computed for each caption and the mean was concatenated with the visual feature as the final "visual" representation. We considered two variants: one with the original captions and an "anonymized" one, with proper nouns replaced by common nouns, i.e., person names replaced by *someone*, place names by *somewhere* and organizations by *organization*. The results in the bottom of Table 4 show that both variants substantially improve the performance, reducing MR from 499 to 7 and 42, respectively. Again, this confirms that the main challenge of the BreakingNews dataset is the lack of correlation between images and articles. While the results obtained considering the original images are very promising, these last two experiments indicate there is still a big margin for improvement.

Geolocation prediction: As a baseline for this 2D regression problem we used SVR and the nearest neighbor method proposed in [28]. Another baseline was the CNN but with an Euclidean loss, as opposed to the GCD loss. In the datasets there are articles with multiple geolocations. For training, we only used the first geolocation, while for testing we computed the GCD as defined in Eq. (8) between the predicted geolocation and the nearest ground truth.

Geolocation Prediction				
Image feat.	Text feat.	Learning	**Mean**	**Med.**
-	BoW TF-IDF	SVR	2.87	1.72
-	W2V mean	SVR	2.52	1.37
VGG19	-	SVR	3.85	1.60
Places	-	SVR	3.89	1.58
VGG19+Places	-	SVR	3.88	1.60
-	BoW TF-IDF	[28]	3.28	1.00
-	W2V mean	[28]	3.21	1.02
VGG19	-	[28]	4.91	3.51
Places	-	[28]	5.11	3.92
VGG19+Places	-	[28]	4.92	3.51
-	W2V matrix	CNN, Euc. loss	2.40	1.41
-	W2V matrix	CNN, GCD loss	1.92	0.90
VGG19	-	CNN, GCD loss	3.38	0.78
Places	-	CNN, GCD loss	3.40	**0.68**
VGG19+Places	-	CNN, GCD loss	3.38	0.76
VGG19+Places	W2V matrix	CNN, GCD loss	**1.91**	0.88

TABLE 5: Results of the geolocation prediction task. Results are expressed in thousands of kilometers. The deep learning based approaches attain better performance, and the Great Circle Distance is a more suitable objective for this task than the Euclidean distance.

Table 5 reports mean and median GCD of the various learning schemes and combinations of features. CNN with the proposed GCD loss has a clear advantage over not only the shallow learners, but also CNN with Euclidean loss. This again confirms the effectiveness of the proposed CNN architecture, especially when coupled with a proper loss function. It is also interesting to note that although using different approaches, the median GCD achieved in our work (680 kilometers of error) is comparable to that of the recent work [79]. Regarding the choice of features for the CNN approach, using only the *Places* CNN image features yields excellent results under the Median GCD metric; yet, in terms of Mean GCD, the smoothing effect of adding the more stable text features prevents potential gross errors committed when the image is non-informative. Figure 5 shows the distribution of geolocation error for image only features (Places CNN) and text plus image features (last row of the results table), and Fig. 6 shows sample images with low and high geolocation error.

6.3 Caption generation

In Table 6 we report caption generation results with the two models in Fig. 3-b1,b2, i.e., LSTM with fixed feature and end-to-end CNNs/LSTM, respectively, and of CCA Sentence selection (CCA-SS), an extractive captioning system. The performance is reported in terms of METEOR [14] and BLEU [62]. Recently introduced CIDEr (Consensus-based Image Description Evaluation) [74] makes judgments based on a large number of reference descriptions e.g. 50 were used in their original work.

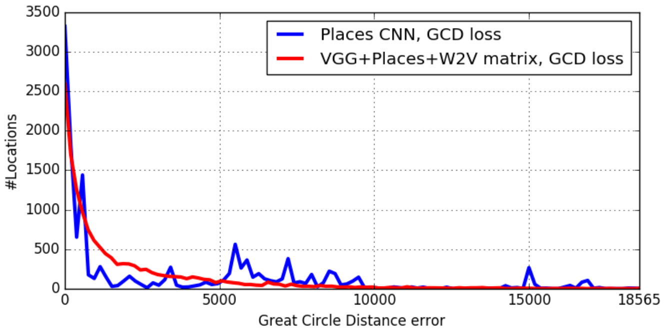


Fig. 5: Comparison of the distribution of errors between the Places CNN with GCD loss (row 11 of Table 5) and the combination of VGG19 and Places CNNs for the image data and the W2V matrix embedding for the text.

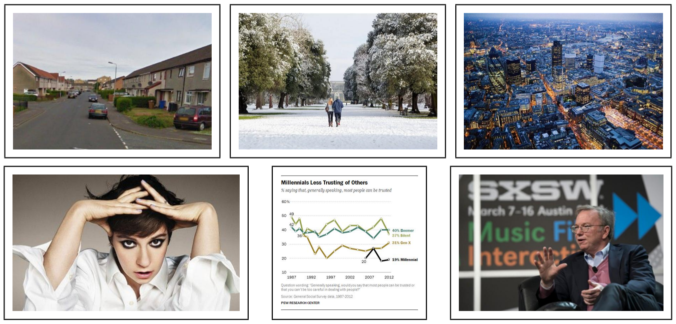


Fig. 6: Random sample images with low geolocation error (top row) and high geolocation error (bottom row) when using only Places CNN features.

BreakingNews has only one caption per image, which is insufficient to run this type of metric.

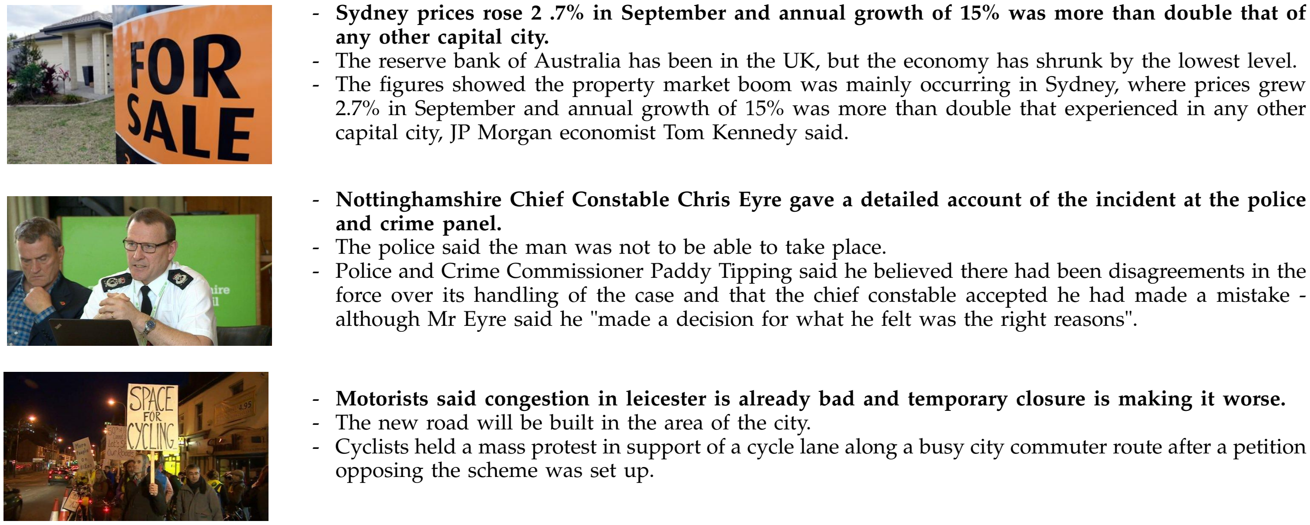
The results indicate that combining textual and visual features, and moreover combining more types of visual features (VGG19 and Places) helps to improve caption generation. On the other hand, finetuning the CNN features in the end-to-end model does not improve the results. We hypothesize that this is again due to the low correlation between the caption and the content of the image. It should also be noted that the performance of all models is poor compared to that achieved by very similar CNN models in standard image caption generation tasks [26], [51]. Again, we argue that captions of news images are of very different nature. Generating such captions is considerably more challenging and requires more careful and specific treatment and evaluation metric, than that done by current state-of-the-art approaches. The superior performance of CCA-SS, which directly takes sentences from the article as captions, also supports this idea. However, one intrinsic limitation of CCA-SS, and extractive methods in general, is that they tend to produce overlong captions: As also noted in [24], rarely an article contains sentences that are ideal for captioning a picture. In our case they were on average 10 words longer than the ground truth ones.

We performed a small-scale human assessment ex-

Caption Generation Results

Text feat.	Image feat.	Learning	**METEOR**	**BLEU-1**	**BLEU-2**	**BLEU-3**	**BLEU-4**
W2V mean	-	LSTM w/ fixed feat.	5.2	16.3	6.6	3.2	1.7
-	VGG19	LSTM w/ fixed feat.	3.9	14.4	4.8	1.8	0.8
-	VGG19+Places	LSTM w/ fixed feat.	4.3	14.4	4.9	2.0	1.0
W2V mean	VGG19	LSTM w/ fixed feat.	4.9	16.2	6.5	3.1	1.6
W2V mean	VGG19+Places	LSTM w/ fixed feat.	**5.3**	17.2	7.0	**3.4**	**1.9**
W2V matrix	VGG19	CNNs + LSTM	5.2	**19.6**	**8.9**	1.6	0.5
W2V mean	VGG19+Places	CCA-SS	8.1	46.3	34.4	24.9	19.5

TABLE 6: Results of caption generation. Performance is overall low, which highlights the difficulty of the task, due to the subtle and high-level relation between the images and the captions.



- **Sydney prices rose 2 .7% in September and annual growth of 15% was more than double that of any other capital city.**
- The reserve bank of Australia has been in the UK, but the economy has shrunk by the lowest level.
- The figures showed the property market boom was mainly occurring in Sydney, where prices grew 2.7% in September and annual growth of 15% was more than double that experienced in any other capital city, JP Morgan economist Tom Kennedy said.

- **Nottinghamshire Chief Constable Chris Eyre gave a detailed account of the incident at the police and crime panel.**
- The police said the man was not to be able to take place.
- Police and Crime Commissioner Paddy Tipping said he believed there had been disagreements in the force over its handling of the case and that the chief constable accepted he had made a mistake - although Mr Eyre said he "made a decision for what he felt was the right reasons".

- **Motorists said congestion in leicester is already bad and temporary closure is making it worse.**
- The new road will be built in the area of the city.
- Cyclists held a mass protest in support of a cycle lane along a busy city commuter route after a petition opposing the scheme was set up.

Fig. 7: Example images, their corresponding captions (first row in each block, boldface), and generated captions (second row: with LSTM; last row: with CCA-based sentence selection). Note that the captions not always explicitly describe the visual content of the images.

periment with the results of our LSTM caption generators and, even though the quantitative results are not representative, we have made several interesting observations: i) In general, the quality of generated captions is much lower in the context of news illustrations compared to other datasets such as MS COCO which may be due to greater diversity of the image content and frequent use of proper nouns. ii) The generated captions more often correspond to the content of the image, while the ground truth is often aligned with the topic of the article. iii) Replacing proper nouns by indefinite pronouns has improved the results which indicates that a model explicitly addressing this issue would be more suitable for similar application scenarios e.g. news articles. The proposed BreakingNews dataset, therefore, poses new challenges for future research. Three example images and associated captions in the training set are shown as examples in Fig. 7.

7 Conclusion

In this paper we have introduced new deep CNN architectures to combine weakly correlated text and image representations and address several tasks in the domain of news articles, including story illustration, source detection, geolocation and automatic captioning. In particular, we propose an adaptive CNN architecture that shares most of its structure for all the tasks. Addressing each problem then requires designing specific loss functions, and we introduce a metric based on the Great Circle Distance for geolocation and Deep Canonical Correlation Analysis for article illustration. All these technical contributions are exhaustively evaluated on a new dataset, BreakingNews, made of approximately 100K news articles, and additionally including a diversity of metadata (like GPS coordinates and popularity metrics) that makes it possible to explore new problems. Overall results are very promising, specially for the tasks of source detection, article illustration and geolocation. The automatic caption generation task, however, is clearly more sensitive to loosely related text and images. Designing new metrics able to handle this situation, is part of our future work.

Acknowledgments

This work was partly funded by the MINECO project RobInstruct TIN2014-58178-R and by the ERA-net CHIS-

TERA project VISEN PCIN-2013-047. Fei Yan and Krystian Mikolajczyk were funded by EPSRC EP/K01904X/2 Visen and EP/N007743/1 FACER2VM projects. The authors would like to thank Nvidia for the hardware donation under the GPU grant program and Blaz Novak for his help with the IJS Newsfeed.

REFERENCES

[1] G. Andrew, R. Arora, J. Bilmes, and K. Livescu. Deep canonical correlation analysis. In *ICML*, 2013.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015.

[3] 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.

[4] K. Barnard and D. Forsyth. Learning the semantics of words and pictures. In *ICCV*, volume 2, pages 408–415. IEEE, 2001.

[5] L. Cao, J. Yu, J. Luo, and T. Huang. Enhancing semantic and geographic annotation of web images via logistic canonical correlation regression. In *ACM International Conference on Multimedia*, pages 125–134. ACM, 2009.

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

[7] D. Chen, G. Baatz, K. Köser, S. Tsai, R. Vedantham, T. Pylvä, K. Roimela, X. Chen, J. Bach, M. Pollefeys, et al. City-scale landmark identification on mobile devices. In *CVPR*, pages 737–744. IEEE, 2011.

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

[9] F. Coelho and C. Ribeiro. Image abstraction in crossmedia retrieval for text illustration. *Lecture Notes in Computer Science*, 7224 LNCS:329–339, 2012.

[10] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. Natural language processing (almost) from scratch. *JMLR*, 12(08):2493–2537, 2011.

[11] B. Coyne and R. Sproat. Wordseye: an automatic text-to-scene conversion system. In *Conference on Computer Graphics and Interactive Techniques*, pages 487–496. ACM, 2001.

[12] D. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. Mapping the world's photos. In *International Conference on World Wide Web*, pages 761–770. ACM, 2009.

[13] J. Deng, W. Dong, R. Socher, K. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

[14] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*, 2014.

[15] J. Ding, L. Gravano, and N. Shivakumar. Computing geographical scopes of web resources. In *International Conference on Very Large Data Bases*, VLDB '00, pages 545–556, 2000.

[16] 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.

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

[18] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[19] H. Fang, S. Gupta, F. Iandola, R. Srivastava, L. Deng, P. Dollar, J. Gao, X. He, M. Mitchell, J. Platt, C. Zitnick, and G. Zweig. From captions to visual concepts and back. In *CVPR*, 2015.

[20] H. Fang, S. Gupta, F. Iandola, R. Srivastava, L. Deng, and P. Dollár others. From captions to visual concepts and back. In *CVPR*, pages 1473–1482, 2015.

[21] A. Farhadi, M. Hejrati, M. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. In *ECCV*, pages 15–29. Springer, 2010.

[22] 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.

[23] Y. Feng and M. Lapata. Automatic caption generation for news images. *PAMI*, 35(4):797–812, 2013.

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

[25] S. Gobron, J. Ahn, G. Paltoglou, M. Thelwall, and D. Thalmann. From sentence to emotion: a real-time three-dimensional graphics metaphor of emotions extracted from text. *The Visual Computer*, 26(6-8):505–519, 2010.

[26] Y. Gong, L. Wang, M. Hodosh, J. Hockenmaier, and S. Lazebnik. Improving image-sentence embeddings using large weakly annotated photo collections. In *ECCV*, pages 529–545. 2014.

[27] D. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.

[28] J. Hays and A. Efros. Im2gps: estimating geographic information from a single image. In *CVPR*, pages 1–8, 2008.

[29] K. Hermann and P. Blunsom. Multilingual distributed representations without word alignment. In *ICLR*, 2014.

[30] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[31] M. Hodosh, P. Young, and J. Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013.

[32] Laura Hollink, Adriatik Bedjeti, Martin van Harmelen, and Desmond Elliott. A corpus of images and text in online news. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may 2016. European Language Resources Association (ELRA).

[33] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

[34] C. Huang, C. Li, and M. Shan. VizStory: Visualization of Digital Narrative for Fairy Tales. *Conference on Technologies and Applications of Artificial Intelligence*, pages 67–72, 2013.

[35] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. arXiv:1408.5093 [cs.CV], 2014.

[36] Y. Jiang, J. Liu, and H. Lu. Chat with illustration. *Multimedia Systems*, 2014.

[37] J. Johnson, A. Karpathy, and L. Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. *ICCV*, 2015.

[38] D. Joshi, J. Wang, and J. Li. The Story Picturing Engine: a system for automatic text illustration. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2(1):68–89, 2006.

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

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

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

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

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

[44] M. Koppel, J. Schler, and S. Argamon. Authorship attribution in the wild. *Language Resources and Evaluation*, 45(1):83–94, 2011.

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

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

[47] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. Berg, and T. Berg. Baby talk: Understanding and generating image descriptions. In *CVPR*. Citeseer, 2011.

[48] N. Kumar, P. Belhumeur, and S. Nayar. Facetracer: A search engine for large collections of images with faces. In *ECCV*, pages 340–353. Springer, 2008.

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

[50] Z. Li, M. Wang, J. Liu, C. Xu, and H. Lu. News contextualization with geographic and visual information. *International Conference on Multimedia*, page 133, 2011.

[51] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. 2014.

[52] L. Maaten and G. Hinton. Visualizing high dimensional data using t-sne. *JMLR*, 9(11):2579–2605, 2008.

[53] Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Advances in Neural Information Processing Systems*, pages 1682–1690, 2014.

[54] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). *ICLR*, 2015.

[55] K. Mardia, J. Kent, and J. Bibby. *Multivariate Analysis*. Academic Press, 1979.

[56] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In *ICLR*, 2013.

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

[58] A. Narayanan, H. Paskov, N. Gong, J. Bethencourt, E. Stefanov, E. Shi, and D. Song. On the feasibility of internet-scale author identification. In *Symposium on Security and Privacy*, pages 300–314. IEEE, 2012.

[59] A. Oliva and A. Torralba. Building the gist of a scene: The role of global image features in recognition. *Progress in brain research*, 155:23–36, 2006.

[60] V. Ordonez, X. Han, P. Kuznetsova, G. Kulkarni, M. Mitchell, K. Yamaguchi, K. Stratos, et al. Large scale retrieval and generation of image descriptions. *International Journal of Computer Vision*, pages 1–14, 2015.

[61] V. Ordonez, G. Kulkarni, and T. Berg. Im2text: Describing images using 1 million captioned photographs. In *NIPS*, pages 1143–1151, 2011.

[62] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

[63] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014.

[64] K. Perlin and A. Goldberg. Improv: A system for scripting interactive actors in virtual worlds. In *Conference on Computer Graphics and Interactive Techniques*, pages 205–216. ACM, 1996.

[65] A. Quattoni, A. Ramisa, P. Swaroop, E. Simo-Serra, and F. Moreno-Noguer. Structured prediction with output embeddings for semantic image annotation. In *Conference of the North American Chapter of the Association for Computational Linguistics*, 2016.

[66] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier. Collecting image annotations using amazon's mechanical turk. In *NAACL: Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 139–147, 2010.

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

[68] 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.

[69] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 [cs.CV], 2014.

[70] R. Socher and L Fei-Fei. Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In *CVPR*, pages 966–973. IEEE, 2010.

[71] N. Srivastava, J. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *JMLR*, 15(6):1929–1958, 2014.

[72] E. Stamatatos. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556, 2009.

[73] M. Trampuš and B. Novak. Internals of an aggregated web news feed. In *International Information Science Conference IS*, pages 431–434, 2012.

[74] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4566–4575, 2015.

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

[76] L. Wang, Y. Li, and S. Lazebnik. Learning deep structure-preserving image-text embeddings. In *CVPR*, 2016.

[77] Peng Wang, Qi Wu, Chunhua Shen, Anton van den Hengel, and Anthony Dick. Explicit knowledge-based reasoning for visual question answering. *arXiv preprint arXiv:1511.02570*, 2015.

[78] Z. Wang, P. Cui, L. Xie, W. Zhu, Y. Rui, and S. Yang. Bilateral Correspondence Model for Words-and-Pictures Association in Multimedia-Rich Microblogs. *ACM Trans. Multimedia Computing, Communications, and Applications*, 10(4):1–21, 2014.

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

[80] B. Wing and J. Baldridge. Simple supervised document geolocation with geodesic grids. In *Annual Meeting of the ACL: Human Language Technologies*, pages 955–964. Association for Computational Linguistics, 2011.

[81] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, pages 3485–3492, 2010.

[82] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015.

[83] F. Yan and K. Mikolajczyk. Deep correlation for matching images and text. In *CVPR*, 2015.

[84] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

[85] Licheng Yu, Eunbyung Park, Alexander C Berg, and Tamara L Berg. Visual madlibs: Fill in the blank description generation and question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2461–2469, 2015.

[86] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *NIPS*, 2014.

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

[88] C. Zitnick, D. Parikh, and L. Vanderwende. Learning the visual interpretation of sentences. In *ICCV*, pages 1681–1688, 2013.

Arnau Ramisa received the MSc degree in computer science (computer vision) from the Autonomous University of Barcelona (UAB) in 2006, and in 2009 completed a PhD at the Artificial Intelligence Research Institute (IIIA-CSIC) and the UAB. From 2009 to the start of 2011, he was a postdoctoral fellow at the LEAR team in INRIA Grenoble / Rhone-Alpes, and since 2011 he is a research fellow at the Institut de Robòtica i Informàtica Industrial in Barcelona (IRI). His research interests include robot vision, object classification and detection, image retrieval and natural language processing.

Fei Yan is a senior research fellow at University of Surrey in the United Kingdom. His research interests focus on machine learning, in particular kernel methods, structured learning, and deep neural networks. He is also interested in the application of machine learning to computer vision and natural language processing, such as object recognition, object tracking, natural language analysis and generation, and joint modelling of vision and language.

Francesc Moreno-Noguer received the MSc degrees in industrial engineering and electronics from the Technical University of Catalonia (UPC) and the Universitat de Barcelona in 2001 and 2002, respectively, and the PhD degree from UPC in 2005. From 2006 to 2008, he was a postdoctoral fellow at the computer vision departments of Columbia University and the École Polytechnique Fédérale de Lausanne. In 2009, he joined the Institut de Robòtica i Informàtica Industrial in Barcelona as an associate researcher of the Spanish Scientific Research Council. His research interests include retrieving rigid and nonrigid shape, motion, and camera pose from single images and video sequences. He received UPC's Doctoral Dissertation Extraordinary Award for his work.

Krystian Mikolajczyk is an Associate Professor at Imperial College London. He completed his PhD degree at the Institute National Polytechnique de Grenoble and held a number of research positions at INRIA, University of Oxford and Technical University of Darmstadt, as well as faculty positions at the University of Surrey, and Imperial College London. His main area of expertise is in image and video recognition, in particular methods for image representation and learning. He has served in various roles at major international conferences co-chairing British Machine Vision Conference 2012 and IEEE International Conference on Advanced Video and Signal-Based Surveillance 2013. In 2014 he received Longuet-Higgins Prize awarded by the Technical Committee on Pattern Analysis and Machine Intelligence of the IEEE Computer Society.