Good News, Everyone! Context driven entity-aware captioning for news images

Ali Furkan Biten, Lluis Gomez, Marçal Rusiñol, Dimosthenis Karatzas

Computer Vision Center, UAB, Spain
{abiten, lgomez, marcal, dimos}@cvc.uab.es

Abstract

Current image captioning systems perform at a merely descriptive level, essentially enumerating the objects in the scene and their relations. Humans, on the contrary, interpret images by integrating several sources of prior knowledge of the world. In this work, we aim to take a step closer to producing captions that offer a plausible interpretation of the scene, by integrating such contextual information into the captioning pipeline. For this we focus on the captioning of images used to illustrate news articles. We propose a novel captioning method that is able to leverage contextual information provided by the text of news articles associated with an image. Our model is able to selectively draw information from the article guided by visual cues, and to dynamically extend the output dictionary to out-of-vocabulary named entities that appear in the context source. Furthermore we introduce "GoodNews", the largest news image captioning dataset in the literature and demonstrate state-of-the-art results.

1. Introduction

People understand scenes by building causal models and employing them to compose stories that explain their perceptual observations [19]. This capacity of humans is associated with intelligent behaviour. One of the cognitive tasks in the Binet-Simon intelligence test [34] is to describe an image. Three performance levels are defined, going from enumeration of objects in the scene, to basic description of contents and finally interpretation, where contextual information is drawn upon to compose an explanation of the depicted events.

Current image captioning systems [37, 2, 16, 31, 23, 11] can at best perform at the description level, if not restricted at the enumeration part, while failing to integrate any prior world knowledge in the produced caption. Prior world knowledge might come in the form of social, political, geographic or temporal context, behavioural cues, or previously built knowledge about entities such as people, places or landmarks. In this work, we aim to take a step closer to producing captions that offer a plausible interpretation of the scene, by integrating such contextual information into the captioning pipeline.

Ground Truth: JoAnn Falletta leading a performance of the Buffalo Philharmonic Orchestra at Kleinhans Music Hall.

Show & Tell [37]: A group of people standing around a table.

Ours: JoAnn Falletta performing at the Buffalo Philharmonic Orchestra.

Figure 1: Standard approaches to image captioning cannot properly take any contextual information into account. Our model is capable of producing captions that include out-of-vocabulary named entities by leveraging information from available context knowledge.

This introduces numerous new challenges. On one hand, the context source needs to be encoded and information selectively drawn from it, guided by the visual scene content. On the other hand, explicit contextual information, typically found in the form of named entities such as proper names, prices, locations, dates, etc, which are typically out-of-dictionary terms or at best underrepresented in the statistics of the dictionary used, need to be properly injected in the produced natural language output.

Currently available image captioning datasets are not

fit for developing captioning models with the aforementioned characteristics, as they provide generic, dry, repetitive and non-contextualized captions, while at the same time there is no contextual information available for each image. For the task at hand, we considered instead other image sources, such as historical archive images or images illustrating newspaper articles, for which captions (i.e. descriptions provided by archivists, captions provided by journalists) and certain contextual information (i.e. history texts, news articles) is readily available or can be collected with reasonable effort.

In this work, we focus on the captioning of images used to illustrate news articles. Newspapers are an excellent domain for moving towards human-like captions, as they provide readily available contextual information that can be modelled and exploited. In this case contextual information is provided by the text of the associated news article, along with other metadata such as titles and keywords. At the same time, there is readily available ground truth in the form of the existing caption written by domain experts (journalists), which is invaluable in itself. Finally, data is freely available at a large scale online. To this end, we have put together GoodNews the biggest news-captioning dataset in the literature with more than 466,000 images and their respective captions and associated articles.

To the best of our knowledge, generative news image captioning has been scarcely explored in the literature [12, 33, 30]. Similarly to [30] we draw contextual information about the image from the associated article. Unlike [30] which uses world-level encoding, we encode the article at the sentence level, as semantic similarity is easier to establish at this granularity. In addition, we introduce an attention mechanism in order to selectively draw information from the article guided by the visual content of the image.

News articles and their respective news image captions, unlike common image captioning datasets such as MSCOCO [21], or Flickr [29], contain a significant amount of named entities. Named entities[1] pose serious problems to current captioning systems that have no mechanism to deal with out-of-vocabulary (OOV) words. This includes [30] where named entity usage is implicitly restricted to the ones that appear in adequate statistics in the training set. Unlike existing approaches, we propose here an end-to-end, two-stage process, where first template captions are produced in which named entities placeholders are indicated along with their respective tags. These are subsequently substituted by selecting the best matching entities from the article, allowing our model to produce captions that include out-of-vocabulary words.

The contributions of this work are as follows:

- We propose a novel captioning method, able to leverage contextual information to produce image captions at the scene interpretation level.
- We propose a two-stage, end-to-end architecture, that allows us to dynamically extend the output dictionary to out-of-vocabulary named entities that appear in the context source.
- We introduce GoodNews, the largest news image captioning dataset in the literature, comprising 466,000 image-caption pairs, along with metadata.

We compare the performance of our proposed method against existing methods and demonstrate state-of-the-art results. Comparative studies demonstrate the importance of properly treating named entities, and the benefits of considering contextual information. Finally, comparisons against human performance highlight the difficulty of the task and limitations of current evaluation metrics.

2. Related Work

Automatic image captioning has received increased attention lately as a result of advances in both computer vision and natural language processing stemming from deep learning [4, 5]. Latest state-of-the-art models [39, 23, 31, 2] usually follow an attention guided encoder-decoder strategy, in which visual information is extracted from images by deep CNNs and then natural language descriptions are generated with RNNs. Despite the good results current state-of-the-art models start to yield according to standard performance evaluation metrics, automatic image captioning is still a challenging problem. Present-day methods tend to produce repetitive, simple sentences [9] written in a consistent style, generally limited on enumerating or describing visual contents, and not offering any deeper semantic interpretation.

The latest attempts of producing richer human-like sentences, are centered in gathering new datasets that might be representative of different writing styles. For example, using crowd-sourcing tools to collect different styles of captions (negative/positive, romantic, humorous, etc.) as in [25, 13], or leveraging the usage of romance novels to change the style of captions to story-like sentences like in [24]. Even though gathering annotations with heterogeneous styles helps mitigating the repetitiveness of the outputs' tone, content-wise captions remain detailed descriptions of the visual content. Automatic captioning still suffers from a huge semantic gap referring to the lack of correlation between images and semantic concepts [33].

The particular domain of news image captioning, has been explored in the past towards incorporating contextual information to the produced captions. In [12] 3K news articles were gathered from BBC News. Image captions were then produced by either choosing the closest sentence in the article or using a template-based linguistic method. In [33], 100K images were collected from TIME magazine, and refined the captioning strategy proposed by Feng et. al. [12].

[1] Named entities are the objects that can be denoted with a proper name such as persons, organizations, places, dates, percentages, etc. [26]

Closer to our work, Ramisa et. al. [30] (BreakingNews) used pre-trained word2vec representations of the news articles concatenated with CNN visual features to feed the generative LSTM. A clear indicator of whether contextual information is correctly incorporated in such cases, is to check to what extent the produced image captions include the correct named entities given the context. This is a challenging task, as in most of the cases such named entities are only becoming available at test time. Although this is particularly important in the case of news image captioning, to the best of our knowledge none of the existing methods addresses named entity inclusion, employing instead closed dictionaries.

Nevertheless, the problem of dealing with named entities has been explored in generic (not context-driven) image captioning. In [35] after gathering Instagram data, a CNN is used to recognize celebrities and landmarks as well as visual concepts such as water, mountain, boat, etc. Afterwards, a confidence model is used to choose whether or not to produce captions with proper names or with visual concepts. In [22] template captions were created using named entity tags, that were later filled by the usage of a knowledge-base graph. The aforementioned methods require a predefined set of named entities. Unlike these methods, our approach looks in the text while producing a caption and "attends" to different sentences for entity extraction, which makes our model consider the context in which the named entities appear to incorporate new, out-of-vocabulary named entities in the produced captions.

3. The GoodNews Dataset

To assemble the *GoodNews* dataset, we have used the New York Times API to retrieve the URLs of news articles ranging from 2010 to 2018. We will provide the URLs of the articles and the script to download images and related metadata, also the released scripts can be used to obtain 167 years worth of news. However, for image captioning purposes, we have restricted our collection to the last 8 years of data, mainly because it covers a period when images were widely used to illustrate news articles. In total, we have gathered $466,000$ images with captions, headlines and text articles, randomly split into $424,000$ for training, $18,000$ for validation and $23,000$ for testing.

GoodNews exhibits important differences to current benchmark datasets for generic captioning like MSCOCO, while it is similar in nature, but about five times larger than BreakingNews, the largest currently available dataset for news image captioning. Key aspects are summarized in Table 1. The *GoodNews* dataset, similarly to BreakingNews, exhibits longer average caption lengths than generic captioning datasets like MSCOCO, indicating that news captions tend to be more descriptive.

GoodNews only includes a single ground truth caption per image, while MSCOCO offers 5 different ground truth captions per image. However, *GoodNews* captions were written by expert journalists, instead of being crowdsourced, which has implications to the style and richness of the text.

Table 1: Comparison of captioning datasets.

	MSCOCO	BreakingNews	GoodNews
Number of Samples	120k	110k	466k
Average Caption Length (words)	11.30	28.09	18.21
Named Entities(Word)	0%	15.66%	19.59%
Named Entities (Sentence)	0%	90.79%	95.51%
Nouns	33.45%	55.59%	46.70%
Adjectives	27.23%	7.21%	5%
Verbs	10.72%	12.57%	11.22%
Pronouns	1.23%	1.36%	2.54%

Named entities represent 20% of the words in the captions of *GoodNews*, while named entities are by design completely absent from the captions of MSCOCO. At the level of sentences, 95% of caption sentences and 73% of article sentences in *GoodNews* contain at least one named entity. Moreover, we observe that *GoodNews* has more named entities than BreakingNews at both token level and sentence level. Analyzing the part of speech tags, we observe that both *GoodNews* and BreakingNews have less amount of adjectives but a higher amount of verbs and significantly higher amount of pronouns and nouns than MSCOCO. Given the nature of news image captions, this is expected, since they do not describe scene objects, but rather offer a contextualized interpretation of the scene.

A key difference between our dataset and BreakingNews, apart from the fact that *GoodNews* has five times more samples, is that our dataset includes a wider range of events and stories since *GoodNews* spans a much longer time period. On the other hand, we must point out that BreakingNews offers a wider range of metadata as it aims to more tasks than news image captioning.

4. Model

As illustrated in Figure 2 our model for context driven entity-aware captioning consists of two consecutive stages. In the first stage, given an image and the text of the corresponding news article, our model generates a template caption where placeholders are introduced to indicate the positions of named entities. In a subsequent stage our model selects the right named entities to fill those placeholders with the help of an attention mechanism over the text of the news article.

We have used SpaCy's named entity recognizer [15] to recognize named entities in both captions and articles of the *GoodNews* dataset. We create template captions by replacing the named entities with their respective tags. At the article level, we store the named entities to be used later in

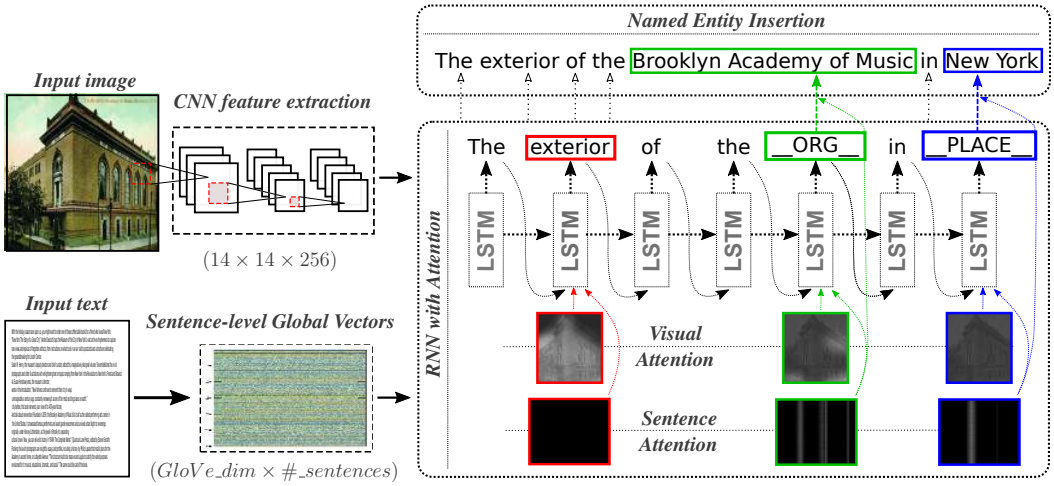

Figure 2: Overview of our model where we combine the visual and textual features to generate first the template captions. Afterwards, we fill these templates with the attention values obtained over the input text. (Best viewed in color)

the named entity insertion stage (see subsection 4.3). As an example, the caption "Albert Einstein taught in Princeton University in 1921" is converted into the following template caption: "PERSON_ taught in ORGANIZATION_ in DATE_". The template captions created this way comprise the training set ground truth we use to train our models. Our model is designed as a two-stream architecture, that combines a visual input (the image) and a textual input (the encoding of the news article).

Our model's main novelty comes from the fact that it encodes the text article associated with each input image and uses it as a second input stream, while employing an attention mechanism over the textual features. For encoding the input text articles we have used the Global Vectors (GloVe) word embedding [28] and an aggregation technique to obtain the article sentence level features. The attention mechanism provides our model with the ability to focus on distinct parts (sentences) of the article at each timestep. Besides, it makes our model end-to-end, capable of inserting the correct named entity in the template caption at each timestep using attention, see Figure 2.

4.1. Template Caption Generation

For the template caption generation stage we follow the same formulation as in state-of-the-art captioning systems [39, 23, 2] which is to produce a word at each timestep given the previously produced word and the attended image features in each step, trained with cross entropy. More formally, we produce a sentence $s_i := \{w_0, w_1, ..., w_N\}$, where w_i is a one-hot vector for the ith word, as follows:

$$x_t = W_e * w_t, \ where\ t \in \{0, 1, ..., N-1\},$$
$$o_t = LSTM(concat(x_t, I_t, A_t)),$$
$$w_{t+1} = softmax(W_{ie} o_t),$$
$$L = -\sum_{t=0}^{N} log(w_{t+1}) \quad (1)$$

where W_e, W_{ie} are learnable parameters, A_t denotes attended article features, and I_t the attended image features. The attended image features at timestep t are obtained as a function of the hidden state of previous timestep and the image features extracted using a Deep CNN model:

$$I_f = CNN(I),$$
$$I_t = Att(h_{t-1}, I_f) \quad (2)$$

where h_{t-1} is the hidden state at time $t-1$, I is the input image, and I_f are features of the input image extracted from a ResNet [14] network pretrained on ImageNet [32].

In the next section we describe three different article encoding techniques that we have used to obtain a fixed size matrix A_f with the sentence level features of the input article. Later, we will explain in detail how we calculate the attended article features, A_t, at every timestep t.

4.2. Article Encoding Methods

Inspired by the state of the art on semantic textual similarity tasks [3], we use a sentence level encoding to represent the news articles in our model, as domain, purpose and context are better preserved at the sentence level.

By using a sentence level encoding, we overcome two shortcomings associated with word level encodings. First, encoding the article at the word granularity requires a higher dimensional matrix which makes the models slower to train and converge. Second, a word level encoding cannot encode the flow (or context) that sentences provide, e.g. "He graduated from Massachusetts" and "He is from Massachusetts":

the former is for MIT which is an organization while the latter one is a state.

Formally, to obtain the sentence level features for the i^{th} article, $A_i := \{s_0^{art}, s_1^{art}, ..., s_M^{art}\}$, where $s_j^{art} = \{w_0, w_1, ..., w_{N_j}\}$ is the j^{th} sentence of article and w_k is the word vector obtained from the pre-trained GloVe model, we have first used a simple average of words for each sentence of the article:

$$A_{f_j}^{avg} = \frac{1}{N_j} \sum_{i=0}^{N_j} w_i, where\ j = 0, 1, ..., M \tag{3}$$

As an alternative we have also considered the use of a weighted average of word vectors according to their smoothed inverse frequency because the simple average of word vectors has huge components along semantically meaningless directions [3]:

$$A_{f_j}^{wAvg} = \frac{1}{N_j} \sum_{i=0}^{N_j} p(w_i) * w_i,\ p(w) = \frac{a}{a + tf(w)} \tag{4}$$

Finally, we have explored the use of the tough-to-beat baseline (TBB) [3], which consists in subtracting the first component of the PCA from the weighted average of the article encoding since empirically the top singular vectors of the datasets seem to roughly correspond to the syntactic information or common words:

$$\begin{aligned} &A_{f_j}^{wAvg} = U\ \Gamma\ V, \\ &X = U^*\ \Gamma^*\ V^*,\ where\ X\ is\ the\ 1^{st}\ component \\ &A_{f_j}^{TBB} = A_f^{wAvg} - X \end{aligned} \tag{5}$$

Article Encoding with Attention: After obtaining the article sentence level features, $A_f \in R^{M \times D_w}$, where M is the fixed sentence length and D_w is the dimension of the word embedding, we have designed an attention mechanism that works by multiplying the sentence level features with an attention vector $\beta_t \in R^M$:

$$\begin{aligned} A_f &= GloVe(A_i), \\ A_t &= \beta_t * A_f \end{aligned} \tag{6}$$

where given the previous timestep of the LSTM, h_{t-1} and article features, A_f, we learn the attention with a fully connected layer:

$$\begin{aligned} \theta_t &= FC(h_{t-1}, A_f), \\ \beta_t &= softmax(\theta_t) \end{aligned} \tag{7}$$

As explained next, apart from improving the generation of the template captions, the usage of attention enables us to also to select the correct named entities to include on the basis of the attention vector.

4.3. Named Entity Insertion

After generating the template captions, we insert named entities according to their categories. If there are more than one tag of PERSON, ORGANIZATION, LOCATION, etc. in the top ranked sentence, we select the named entity in order of appearance in the sentence. In order to compare our method with standard image captioning models we came up with there different insertion techniques, from which two can be used with visual-only architectures (i.e. without considering the article text features): Random Insertion (RandIns) and Context-based Insertion (CtxIns). Whereas the third one is based on an attention mechanism over the article that guides the insertion (AttIns).

The random insertion (RandIns) offers a baseline for the other insertion methods explored, and it consists of randomly picking a named entity of the same category from the article, for each named entity placeholder that is produced in the template captions.

For the Context Insertion (CtxIns) we make use of a pre-trained GloVe embedding to rank the sentences of articles with cosine similarity according to the produced template caption embedding and afterwards insert the named entities on the basis of this ranking.

Finally, for our insertion by attention method (AttIns), we use the article attention vector β_t that is produced at each timestep t of the template caption generation to insert named entities without using any external insertion method.

4.4. Implemention Details

We coded our models in PyTorch. We have used the 5th layer of ResNet-152 [14] for image attention and a single-layer LSTM with dimension size 512. We re-sized each image into 256×256 and then randomly cropped them to 224×224. We created our vocabulary by removing words that occur less than 4 times, resulting in 35K words while we also truncated long sentences to a maximum length of 31 words. For the article encoding, we used SpaCy's pre-trained GloVe embedding with dimension size of 300 and set the maximum sentence length to 55. In 95% of the cases, articles have less than 55 sentences. In the case of articles with more than 55 sentences, we encode the average representation of the rest of the sentences at the 55^{th} dimension. In all of our models, we used Adam [18] optimizer with 0.002 learning rate with learning rate decay 0.8 after 10 epochs for every 8 epochs with dropout probability set to 0.2. We have produced our captions with beam size 1. The code and dataset are available online[2].

5. Experiments

In this section we provide several experiments in order to evaluate the quality of the image captions generated with

[2] https://github.com/furkanbiten/GoodNews

Table 2: Results on the intermediate task of template caption generation for state-of-the-art captioning models without using any Article Encoding (top) and for our method using different Article Encoding strategies (bottom).

	Bleu-1	Bleu-2	Bleu-3	Bleu-4	Meteor	Rouge-L	CIDEr	Spice
Show Attend Tell [39]	**11.537**%	**5.757**%	**2.983**%	**1.711**%	**13.559**%	**20.468**%	**17.317**%	22.864%
Att2in2 [31]	10.536%	5.176%	2.716%	1.542%	12.962%	19.934%	16.511%	23.789%
Up-Down [2]	10.812%	5.201%	2.649%	1.463%	12.546%	19.424%	15.345%	23.112%
Adaptive Att [23]	7.916%	3.858%	1.941%	1.083%	12.576%	19.638%	15.928%	**25.017**%
Ours (Average)	**13.419**%	**6.530**%	**3.336**%	**1.869**%	**13.752**%	**20.468**%	**17.577**%	22.699%
Ours (Weighted Average)	11.898%	5.857%	3.012%	1.695%	13.645%	20.355%	17.132%	**23.251**%
Ours (TBB)	12.236%	5.817%	2.950%	1.662%	13.530%	20.353%	16.624%	22.766%

Table 3: Results on news image captioning. RandIns: Random Insertion; CtxIns: GloVe Insertion; AttIns: Insertion by Attention; No-NE: without named entity insertion.

		Bleu-1	Bleu-2	Bleu-3	Bleu-4	Meteor	Rouge	CIDEr	Spice
Visual only	**Show Attend Tell - No-NE**	8.80%	3.01%	0.97%	0.43%	2.47%	9.06%	1.67%	0.69%
	Show Attend Tell + RandIns	7.37%	2.94%	1.34%	0.70%	3.77%	11.15%	10.03%	3.48%
	Att2in2 + RandIns	6.88%	2.82%	1.35%	0.73%	3.57%	10.84%	9.68%	3.57%
	Up-Down + RandIns	6.92%	2.77%	1.29%	0.67%	3.40%	10.38%	8.94%	3.60%
	Adaptive Att + RandIns	5.22%	2.11%	0.97%	0.51%	3.28%	10.21%	8.68%	3.56%
	Show Attend Tell + CtxIns	7.63%	3.03%	1.39%	0.73%	4.14%	11.88%	12.15%	4.03%
	Att2in2 + CtxIns	7.11%	2.91%	1.39%	0.76%	3.90%	11.58%	11.58%	4.12%
	Up-Down + CtxIns	7.21%	2.87%	1.34%	0.71%	3.74%	11.06%	11.02%	3.91%
	Adaptive Att + CtxIns	5.30%	2.11%	0.98%	0.51%	3.59%	10.94%	10.55%	4.13%
Visual & Textual	**BreakingNews*** - No-NE [30]	5.06%	1.70%	0.60%	0.31%	1.66%	6.38%	1.28%	0.49%
	Ours (Avg.) + CtxIns	**8.92**%	**3.54**%	**1.60**%	**0.83**%	**4.34**%	12.10%	12.75%	4.20%
	Ours (Wavg.) + CtxIns	7.99%	3.22%	1.50%	0.79%	4.21%	11.86%	12.37%	**4.25**%
	Ours (TBB) + CtxIns	8.32%	3.31%	1.52%	0.80%	4.27%	**12.11**%	12.70%	4.19%
	Ours (Avg.) + AttIns	8.63%	3.45%	1.57%	0.81%	4.23%	11.72%	12.70%	4.20%
	Ours (Wavg.) + AttIns	7.70%	3.13%	1.44%	0.74%	4.11%	11.54%	12.53%	**4.25**%
	Ours (TBB) + AttIns	8.04%	3.23%	1.47%	0.76%	4.17%	11.81%	**12.79**%	4.19%
	$Human^{\dagger}$- (Estimation)	14.24%	7.70%	4.76%	3.22%	10.03%	15.98%	39.58%	13.87%

*: Reported results are based on our own implementation.
†: Indicative performance, based on two subjects' captions over a subset of 20 samples.

our model on the *GoodNews* dataset. First, we compare the obtained results with the state of the art on image captioning using standard metrics. Then we analyze the precision and recall of our method for the specific task of named entity insertion. Finally we provide a human evaluation study and show some qualitative results.

As discussed extensively in the literature [8, 10, 17, 38, 6] standard evaluation metrics for image captioning have several flaws and in many cases they do not correlate with human judgments. Although we present the results in Bleu [27], METEOR [7], ROUGE [20], CIDEr [36] and SPICE [1], we believe the most suitable metric for the specific scenario of image captioning for news images is CIDEr. This is because both METEOR and SPICE use synonym matching and lemmatization, and named entities rarely have any meaningful synonyms or lemmas. For Bleu and ROUGE, every word alters the metric equally: e.g. missing a stop word has the same impact as the lack of a named entity. That is why we believe CIDEr, although it has its own drawbacks, is the most informative metric to analyze our results since it downplays the stop words and puts more importance to the "unique" words by using a tf-idf weighting scheme.

5.1. News Image Captioning

Our pipeline for news image captioning operates at two levels. First it produces template captions, before substituting the placeholders with named entities from the text.

Table 2 shows the results on the intermediate task of template caption generation for state-of-the-art captioning models without using any contextual information ("Visual only", i.e. ignoring the news articles), and compares them with our method's results using different Article Encoding strategies ("Visual & Textual"). We appreciate that the "Show, Attend and Tell" [39] model outperforms the rest of the baselines [2, 31, 23] on the intermediate task of template caption generation. This outcome differs from the results obtained on other standard benchmarks for image captioning like MSCOCO, where [2, 31, 23] are known to improve over the "Show, Attend and Tell" model. We believe this discrepancy can be explained because those architectures are better at recognizing objects in the input image and their relations, but when the image and its caption are loosely related at the object level, as is the case in the many of the *GoodNews* samples, these models fail to capture the underlying semantic relationships between images and captions.

(a)		**GT**: Sidney Crosby celebrated his goal in the second period that seemed to deflate Sweden. **V**: `Crosby of Vancouver won the Crosby in several seasons.` **V+T**: `Crosby of Canada after scoring the winning goal in the second period.`
(b)		**GT**: Ms Ford and her husband Erik Allen Ford in their cabin. **V**: `Leanne Ford and Ford in the kitchen.` **V+T**: `Ford and Ford in their home in Echo Park.`
(c)		**GT**: Ismail Haniya the leader of the Hamas government in Gaza in Gaza City last month. **V**: `Haniya left and Mahmoud Abbas in Gaza City.` **V+T**: `Haniya the Hamas speaker leaving a meeting in Gaza City on Wednesday.`
(d)		**GT**: Supreme Court nominee Robert Bork testifying before the Senate Judiciary Committee. **V**: `Bork left and the Bork Battle in GPE.` **V+T**: `Bork the the Supreme Court director testifying before Senate on 1987.`

Figure 3: Qualitative Result; V: Visual Only, V+T: Visual and Textual, GT: Ground Truth

Therefore, we have decided to use the architecture of "Show Attend and Tell" as the basis for our own model design. We build our two stream architecture, that combines a visual input and a textual input. From Table 2, we can see that encoding the article by simply averaging the GloVe descriptors of its sentences achieves slightly better scores on the intermediate task of template-based captioning than the weighted average and tough-to-beat baseline (TBB) approaches. Overall, the performance of our two-stream (visual and textual) architecture is on par with the baseline results in this task.

In Table 3, we produce the full final captions for both approaches (visual only and visual+textual) by using different strategies for the named entity insertion: random insertion (RandIns), GloVe based context insertion (CtxIns), and insertion by attention (AttIns). Our architecture consistently outperforms the "Visual only" pipelines on every metric. Moreover, without the two-stage formulation we introduced (template-based and full captions), current captioning systems (see "Show Attend Tell - No-NE" in Table 3) as well as BreakingNews [30] perform rather poorly.

Despite the fact that the proposed approach yields better results than previous state of the art, and properly deals with out-of-dictionary words (named entities), the overall low results, compared with the typical results on simpler datasets such as MSCOCO, are indicative of the complexity of the problem and the limitations of current captioning approaches. To emphasize this aspect we provide in Table 3 an estimation of human performance in the task of full caption generation on the *GoodNews* dataset. The reported numbers indicate the average performance of 2 subjects tasked with creating captions for a small subset of 20 images and their associated articles.

Finally, we provide in Figure 3 a qualitative comparison for the best performing model of both "visual only" (Show, Attend and Tell+CtxIns) and "visual+textual" (Avg+AttIns) architectures. We appreciate that taking the textual content into account results in more contextualized captions. We also present some failure cases in which incorrect named entities have been inserted.

5.2. Evaluation of Named Entity Insertion

Results of Table 2 represent a theoretical maximum, since a perfect named entity insertion would give us those same results for the full caption generation task. However, from Table 2 results to Table 3 there is a significant drop ranging from 4 to 18 points in each metric. To further quantify the differences between context insertion and insertion by attention, we provide in Table 4 their precision and recall for exact and partial match named entity insertion. In the exact match evaluation, we only accept the insertion of the names as true positive if they match the ground truth character by character, while on the partial match setting, we do consider token level match as being correct (i.e. "Falletta" is considered a true positive for the "JoAnn Falletta" entity).

In Table 4, we observe that the proposed insertion by attention ("AttIns") clearly outperforms the "CtxIns" strategy at both exact and partial match evaluations. The use of the proposed text attention mechanism allows us to deal with named entity insertion in an end-to-end fashion, eliminating the need for any separate processing.

However, notice that this was not revealed by the anal-

Table 4: Precision and Recall for named entity insertion.

	Exact match		Partial match	
	P	R	P	R
Show Attend Tell + CtxIns	8.19	7.10	19.39	17.33
Ours (Avg.) + CtxIns	8.17	7.23	19.53	17.88
Ours (WAvg.) + CtxIns	7.80	6.68	19.14	17.08
Ours (TBB) + CtxIns	7.84	6.64	19.60	17.11
Ours (Avg.) + AttIns	**9.19**	**8.21**	21.17	**19.48**
Ours (WAvg.) + AttIns	8.88	7.74	21.11	19.00
Ours (TBB) + AttIns	9.09	7.81	**21.71**	19.19

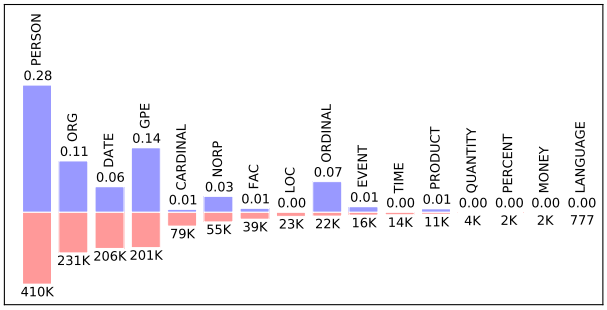

Figure 4: Named entity insertion recall (blue) and number of training samples (red) for each named entity category.

ysis of Table 3, where all insertion strategies seem to have a similar effect. This is partly explained by the fact that image captioning evaluation metrics fail to put any special weight to named entities. Intuitively, humans would prefer captions where the named entities are correctly inserted. To further analyze the results of this experiment we provide in Figure 4 the named entity insertion recall of our method (Avg+AttIns) on each of the individual named entity tags. We observe a correlation of the recall values with the number of training samples for each named entity category. This suggests that the overall named entity insertion performance can be potentially improved with more training data.

5.3. Human Evaluation

In order to provide a more fair evaluation we have conducted a human evaluation study. We asked 20 human evaluators to compare the outputs of the best performing "visual + textual" model (Avg. + AttIns) with the ones of the best performing "visual only" model ("Show Attend and Tell" with Ctx named entity insertion) on a subset of 106 randomly chosen images. Evaluators were presented an image, its ground-truth caption, and the two captions generated by those methods, and were asked to choose the one they considered most similar to the ground truth. In total we collected 2,101 responses.

The comparative study revealed that our model was perceived as better than "Show Attend and Tell + CtxIns" in 53% of the cases. In Figure 5 we analyze the results as a function of the degree of consensus of the evaluators for each image. Our aim is to exclude from the analysis those images in which there is no clear consensus about the better caption between the evaluators. To do this we define the degree of consensus $C = 1 - \frac{min(votes_v, votes_{v+t})}{max(votes_v, votes_{v+t})}$, where $votes_v$ and $votes_{v+t}$ denote the evaluator votes for each method. At each value of C We reject all images that have smaller consensus. Then we report on how many samples the majority vote was for the "visual" or "visual+textual" method. As can be appreciated the results indicate a consistent preference for the "visual+textual" variant.

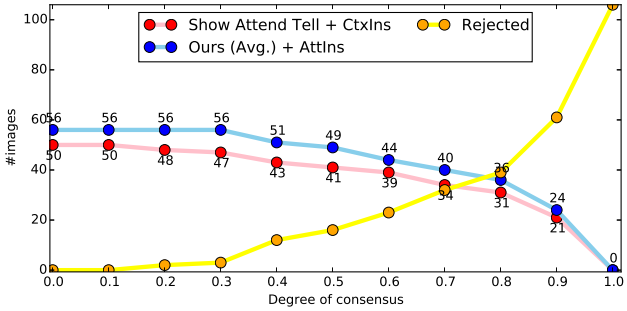

Figure 5: Comparison of "visual only" and "visual+textual" models regarding human judgments.

6. Conclusion

In this paper we have presented a novel captioning pipeline that aims to take a step closer to producing captions that offer a plausible interpretation of the scene, and applied it to the particular case of news image captioning. In addition, we presented *GoodNews*, a new dataset comprising 466K samples, the largest news-captioning dataset to date. Our proposed pipeline integrates contextual information, given here in the form of a news article, introducing an attention mechanism that permits the captioning system to selectively draw information from the context source, guided by the image. Furthermore, we proposed a two-stage procedure implemented in an end-to-end fashion, to incorporate named entities in the captions, specifically designed to deal with out-of-dictionary entities that are only made available at test time. Experimental results demonstrate that the proposed method yields state-of-the-art performance, while it satisfactorily incorporates named entity information in the produced captions.

Acknowledgements

This work has been supported by projects TIN2017-89779-P, Marie-Curie (712949 TECNIOspring PLUS), aBSINTHE (Fundación BBVA 2017), the CERCA Programme / Generalitat de Catalunya, NVIDIA Corporation and a UAB PhD scholarship.

References

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, 2016.

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and VQA. *arXiv preprint arXiv:1707.07998*, 2017.

[3] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *International Conference on Learning Representations*, 2017.

[4] Shuang Bai and Shan An. A survey on automatic image caption generation. *Neurocomputing*, 311:291–304, 2018.

[5] Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artificial Intelligence Research*, 55:409–442, 2016.

[6] Yin Cui, Guandao Yang, Andreas Veit, Xun Huang, and Serge Belongie. Learning to evaluate image captioning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

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

[8] Jacob Devlin, Hao Cheng, Hao Fang, Saurabh Gupta, Li Deng, Xiaodong He, Geoffrey Zweig, and Margaret Mitchell. Language models for image captioning: The quirks and what works. *arXiv preprint arXiv:1505.01809*, 2015.

[9] Jacob Devlin, Saurabh Gupta, Ross Girshick, Margaret Mitchell, and C Lawrence Zitnick. Exploring nearest neighbor approaches for image captioning. *arXiv preprint arXiv:1505.04467*, 2015.

[10] Desmond Elliott and Frank Keller. Comparing automatic evaluation measures for image description. In *Annual Meeting of the Association for Computational Linguistics*, 2014.

[11] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K. Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, and Geoffrey Zweig. From captions to visual concepts and back. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

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

[13] Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. Stylenet: Generating attractive visual captions with styles. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[15] Matthew Honnibal and Ines Montani. SpaCy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. 2017.

[16] Andrej Karpathy and Li Fei-Fei. Deep Visual-Semantic Alignments for Generating Image Descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):664–676, 2017.

[17] Mert Kilickaya, Aykut Erdem, Nazli Ikizler-Cinbis, and Erkut Erdem. Re-evaluating Automatic Metrics for Image Captioning. 2016.

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[19] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40, 2017.

[20] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *ACL Workshop on Text Summarization*, 2004.

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, 2014.

[22] Di Lu, Spencer Whitehead, Lifu Huang, Heng Ji, and Shih-Fu Chang. Entity-aware image caption generation. *arXiv preprint arXiv:1804.07889*, 2018.

[23] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Conference on Computer Vision and Pattern Recognition*, 2017.

[24] Alexander Mathews, Lexing Xie, and Xuming He. Semstyle: Learning to generate stylised image captions using unaligned text. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[25] Alexander Patrick Mathews, Lexing Xie, and Xuming He. Senticap: Generating image descriptions with sentiments. In *Conference of the Association for the Advancement of Artificial Intelligence*, 2016.

[26] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

[27] Kishore Papineni, Salim Roukos, Todd Ward, and Wj Zhu. BLEU: a method for automatic evaluation of machine translation. *Annual Meeting on Association for Computational Linguistics*, 2002.

[28] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Conference on Empirical Methods in Natural Language Processing*, 2014.

[29] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *IEEE International Conference on Computer Vision*, 2015.

[30] Arnau Ramisa, Fei Yan, Francesc Moreno-Noguer, and Krystian Mikolajczyk. Breakingnews: Article annotation by image and text processing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(5):1072–1085, 2018.

[31] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. Self-critical sequence training for

image captioning. In *Conference on Computer Vision and Pattern Recognition*, 2017.

[32] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[33] Amara Tariq and Hassan Foroosh. A context-driven extractive framework for generating realistic image descriptions. *IEEE Transactions on Image Processing*, 26(2):619–632, 2017.

[34] Lewis Madison Terman. *The measurement of intelligence: An explanation of and a complete guide for the use of the Stanford revision and extension of the Binet-Simon intelligence scale*. Houghton Mifflin, 1916.

[35] Kenneth Tran, Xiaodong He, Lei Zhang, Jian Sun, Cornelia Carapcea, Chris Thrasher, Chris Buehler, and Chris Sienkiewicz. Rich image captioning in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016.

[36] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[37] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[38] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and Tell: Lessons Learned from the 2015 MSCOCO Image Captioning Challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):652–663, 2017.

[39] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *International Conference on Machine Learning*, 2015.