Transform and Tell: Entity-Aware News Image Captioning

Alasdair Tran, Alexander Mathews, Lexing Xie
Australian National University
{alasdair.tran,alex.mathews,lexing.xie}@anu.edu.au

Abstract

We propose an end-to-end model which generates captions for images embedded in news articles. News images present two key challenges: they rely on real-world knowledge, especially about named entities; and they typically have linguistically rich captions that include uncommon words. We address the first challenge by associating words in the caption with faces and objects in the image, via a multi-modal, multi-head attention mechanism. We tackle the second challenge with a state-of-the-art transformer language model that uses byte-pair-encoding to generate captions as a sequence of word parts. On the GoodNews dataset [3], our model outperforms the previous state of the art by a factor of four in CIDEr score ($13 \rightarrow 54$). This performance gain comes from a unique combination of language models, word representation, image embeddings, face embeddings, object embeddings, and improvements in neural network design. We also introduce the NYTimes800k dataset which is 70% larger than GoodNews, has higher article quality, and includes the locations of images within articles as an additional contextual cue.

1. Introduction

The Internet is home to a large number of images, many of which lack useful captions. While a growing body of work has developed the capacity to narrate the contents of generic images [10, 49, 12, 19, 39, 30, 1, 6], these techniques still have two important weaknesses. The first weakness is in world knowledge. Most captioning systems are aware of generic object categories but unaware of names and places. Also generated captions are often inconsistent with commonsense knowledge. The second weakness is in linguistic expressiveness. The community has observed that generated captions tend to be shorter and less diverse than human-written captions [50, 24]. Most captioning systems rely on a fixed vocabulary and cannot correctly place or spell new or rare words.

News image captioning is an interesting case study for tackling these two challenges. Not only do news captions describe specific people, organizations and places, but the associated news articles also provide rich contextual information. The language used in news is evolving, with both the vocabulary and style changing over time. Thus news captioning approaches need to adapt to new words and concepts that emerge over a longer period of time (e.g. *walkman* in the 1990s or *mp3 player* in the 2000s). Existing approaches [44, 37, 3] rely on text extraction or template filling, which prevents the results from being linguistically richer than the template generator and are error-prone due to the difficulty in ranking entities for gap filling. Successful strategies for news image captioning can be generalized to images from domains with other types of rich context, such as web pages, social media posts, and user comments.

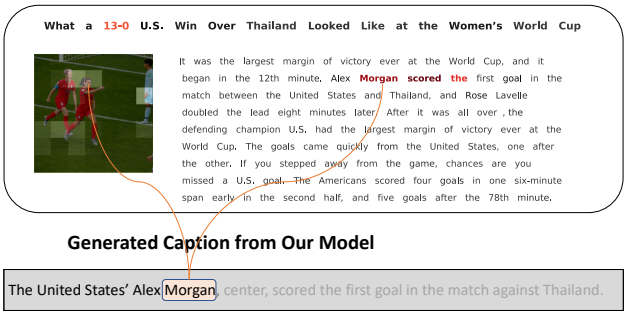

Figure 1: An example of entity-aware news image captioning. Given a news article and an image (top), our model generates a relevant caption (bottom) by attending over the contexts. Here we show the attention scores over the image patches and the article text as the decoder generates the word "Morgan". Image patches with higher attention have a lighter shade, while highly-attended words are in red. The orange lines point to the highly attended regions.

We propose an end-to-end model for news image captioning with a novel combination of sequence-to-sequence neural networks, language representation learning, and vision subsystems. In particular, we address the knowledge gap by computing multi-head attention on the words in the article, along with faces and objects that are extracted from the image. We address the linguistic gap with a flexible byte-pair-encoding that can generate unseen words. We

use dynamic convolutions and mix different linguistic representation layers to make the neural network representation richer. We also propose a new dataset, NYTimes800k, that is 70% larger than GoodNews [3] and has higher-quality articles along with additional image location information. We observe a performance gain of $6.8\times$ in BLEU-4 ($0.89 \rightarrow 6.05$) and $4.1\times$ in CIDEr ($13.1 \rightarrow 53.8$) compared to previous work [3]. On both datasets we observe consistent gains for each new component in our language, vision, and knowledge-aware system. We also find that our model generates names not seen during training, resulting in linguistically richer captions, which are closer in length (mean 15 words) to the ground truth (mean 18 words) than the previous state of the art (mean 10 words).

Our main contributions include:

1. A new captioning model that incorporates transformers, an attention-centric language model, byte-pair encoding, and attention over four different modalities (text, images, faces, and objects).
2. Significant performance gains over all metrics, with associated ablation studies quantifying the contributions of our main modeling components using BLEU-4, CIDEr, precision & recall of named entities and rare proper nouns, and linguistic quality metrics.
3. NYTimes800k, the largest news image captioning dataset to date, containing 445K articles and 793K images with captions from The New York Times spanning 14 years. NYTimes800k builds and improves upon the recently proposed GoodNews dataset [3]. It has 70% more articles and includes image locations within the article text. The dataset, code, and pre-trained models are available on GitHub[1].

2. Related Works

A popular design choice for image captioning systems involves using a convolutional neural network (CNN) as the image encoder and a recurrent neural network (RNN) with a closed vocabulary as a decoder [19, 10, 49]. Attention over image patches using a multilayer perception was introduced in "Show, Attend and Tell" [53]. Further extensions include having the option to not attend to any image region [30] using a bottom-up approach to propose a region to attend to [1], and attending specifically to object regions [51] and visual concepts [55, 25, 51] identified in the image.

News image captioning includes the article text as input and focuses on the types of images used in news articles. A key challenge here is to generate correct entity names, especially rare ones. Existing approaches include extractive methods that use n-gram models to combine existing phrases [13] or simply retrieving the most representative sentence [44] in the article. Ramisa *et al*. [37] built an end-to-end LSTM decoder that takes both the article and image as inputs, but the model was still unable to produce names that were not seen during training.

To overcome the limitation of a fixed-size vocabulary, template-based methods have been proposed. An LSTM first generates a template sentence with placeholders for named entities, e.g. "PERSON speaks at BUILDING in DATE." [3]. Afterwards the best candidate for each placeholder is chosen via a knowledge graph of entity combinations [29], or via sentence similarity [3]. One key difference between our proposed model and previous approaches [3, 29] is that our model can generate a caption with named entities directly without using an intermediate template.

One tool that has seen recent successes in many natural language processing tasks are transformer networks. Transformers have been shown to consistently outperform RNNs in language modeling [36], story generation [11], summarization [43], and machine translation [4]. In particular, transformer-based models such as BERT [9], XLM [22], XLNet [54], RoBERTa [27], and ALBERT [23] are able to produce high level text representations suitable for transfer learning. Furthermore, using byte-pair encoding (BPE) [41] to represent uncommon words as a sequence of subword units enables transformers to function in an open vocabulary setting. To date the only image captioning work that uses BPE is [57], but they did not use it for rare named entities as these were removed during pre-processing. In contrast we explicitly examine BPE for generating rare names and compare it to template-based methods.

Transformers have been shown to yield competitive results in generating generic MS COCO captions [58, 25]. Zhao *et al*. [57] have gone further and trained transformers to produce some named entities in the Conceptual Captions dataset [42]. However, the authors used web-entity labels, extracted using Google Cloud Vision API, as inputs to the model. In our work, we do not explicitly give the model a list of entities to appear in the caption. Instead our model automatically identifies relevant entities from the provided news article.

3. The Transform and Tell Model

Our model consists of a set of pretrained encoders and a decoder, as illustrated in Figure 2. The encoders (Section 3.1) generate high-level vector representations of the images, faces, objects, and article text. The decoder (Section 3.2) attends over these representations to generate a caption at the sub-word level.

3.1. Encoders

Image Encoder: An overall image representation is obtained from a ResNet-152 [17] model pre-trained on Im-

[1] https://github.com/alasdairtran/transform-and-tell

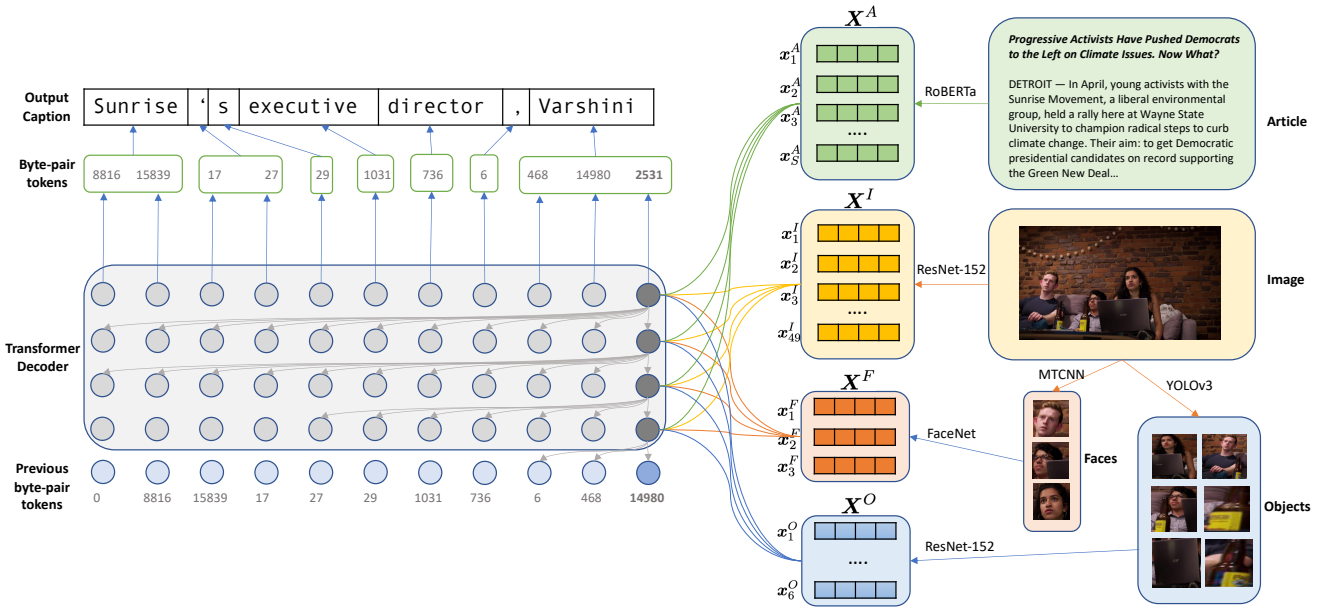

Figure 2: Overview of the Transform and Tell model. Left: Decoder with four transformer blocks; Right: Encoder for article, image, faces, and objects. The decoder takes embeddings of byte-pair tokens as input (blue circles at the bottom). For example, the input in the final time step, 14980, represents “arsh” in “Varshini” from the previous time step. The grey arrows show the convolutions in the final time step in each block. Colored arrows show attention to the four domains on the right: article text (green lines), image patches (yellow lines), faces (orange lines), and objects (blue lines). The final decoder outputs are byte-pair tokens, which are then combined to form whole words and punctuations.

ageNet. We use the output of the final block before the pooling layer as the image representation. This is a set of 49 different vectors $\boldsymbol{x}_i^I \in \mathbb{R}^{2048}$ where each vector corresponds to a separate image patch after the image is divided into equally-sized 7 by 7 patches. This gives us the set $\boldsymbol{X}^I = \{\boldsymbol{x}_i^I \in \mathbb{R}^{D^I}\}_{i=1}^{M^I}$, where $D^I = 2048$ and $M^I = 49$ for ResNet-152. Using this representation allows the decoder to attend to different regions of the image, which is known to improve performance in other image captioning tasks [53] and has been widely adopted.

Face Encoder: We use MTCNN [56] to detect face bounding boxes in the image. We then select up to four faces since the majority of the captions contain at most four people’s names (see Section 4). A vector representation of each face is obtained by passing the bounding boxes to FaceNet [40], which was pre-trained on the VGGFace2 dataset [5]. The resulting set of face vectors for each image is $\boldsymbol{X}^F = \{\boldsymbol{x}_i^F \in \mathbb{R}^{D^F}\}_{i=1}^{M^F}$, where $D^F = 512$ for FaceNet and M^F is the number of faces. If there are no faces in the image, $\boldsymbol{X}^F$ is an empty set.

Even though the faces are extracted from the image, it is useful to consider them as a separate input domain. This is because a specialized face embedding model is tuned for identifying people and thus can help the decoder to generate more accurate named entities.

Object Encoder: We use YOLOv3 [38] to detect object bounding boxes in the image. We filter out objects with a confidence less than 0.3 and select up to 64 objects with the highest confidence scores to feed through a ResNet-152 pretrained on ImageNet. In contrast to the image encoder, we take the output after the pooling layer as the representation for each object. This gives us a set of object vectors $\boldsymbol{X}^O = \{\boldsymbol{x}_i^O \in \mathbb{R}^{D^O}\}_{i=1}^{M^O}$, where $D^O = 2048$ for ResNet-152 and M^O is the number of objects.

Article Encoder: To encode the article text we use RoBERTa [27], a recent improvement over the popular BERT [9] model. RoBERTa is a pretrained language representation model providing contextual embeddings for text. It consists of 24 layers of bidirectional transformer blocks.

Unlike GloVe [35] and word2vec [31] embeddings, where each word has exactly one representation, the bidirectionality and the attention mechanism in the transformer allow a word to have different vector representations depending on the surrounding context.

The largest GloVe model has a vocabulary size of 1.2 million. Although this is large, many rare names will still get mapped to the unknown token. In contrast, RoBERTa uses BPE [41, 36] which can encode any word made from Unicode characters. In BPE, each word is first broken down into a sequence of bytes. Common byte sequences are then merged using a greedy algorithm. Following [36], our vocabulary consists of 50K most common byte sequences.

Inspired by Tenney *et al.* [46] who showed that different layers in BERT represent different steps in the tradi-

tional NLP pipeline, we mix the RoBERTa layers to obtain a richer representation. Given an input of length M^T, the pretrained RoBERTa encoder will return 25 sequences of embeddings, $\boldsymbol{G} = \{\boldsymbol{g}_{\ell i} \in \mathbb{R}^{2048} : \ell \in \{0, 1, ..., 24\}, i \in \{1, 2, ..., M^T\}\}$. This includes the initial uncontextualized embeddings and the output of each of the 24 transformer layers. We take a weighted sum across all layers to obtain the article embedding $\boldsymbol{x}_i^A$:

$$\boldsymbol{x}_i^A = \sum_{\ell=0}^{24} \alpha_\ell \, \boldsymbol{g}_{\ell i} \tag{1}$$

where α_ℓ are learnable weights.

Thus our RoBERTa encoder produces the set of token embeddings $\boldsymbol{X}^A = \{\boldsymbol{x}_i^A \in \mathbb{R}^{D^T}\}_{i=1}^{M^T}$, where $D^T = 1024$ in RoBERTa.

3.2. Decoder

The decoder is a function that generates caption tokens sequentially. At time step t, it takes as input: the embedding of the token generated in the previous step, $\boldsymbol{z}_{0t} \in \mathbb{R}^{D^E}$ where D^E is the hidden size; embeddings of all other previously generated tokens $\boldsymbol{Z}_{0<t} = \{\boldsymbol{z}_{00}, \boldsymbol{z}_{01}, ..., \boldsymbol{z}_{0t-1}\}$; and the context embeddings $\boldsymbol{X}^I$, $\boldsymbol{X}^A$, $\boldsymbol{X}^F$, and $\boldsymbol{X}^O$ from the encoders. These inputs are then fed through L transformer blocks:

$$\boldsymbol{z}_{1t} = \text{Block}_1(\boldsymbol{z}_{0t} | \boldsymbol{Z}_{0<t}, \boldsymbol{X}^I, \boldsymbol{X}^A, \boldsymbol{X}^F, \boldsymbol{X}^O) \tag{2}$$

$$\boldsymbol{z}_{2t} = \text{Block}_2(\boldsymbol{z}_{1t} | \boldsymbol{Z}_{1<t}, \boldsymbol{X}^I, \boldsymbol{X}^A, \boldsymbol{X}^F, \boldsymbol{X}^O) \tag{3}$$

$$\ldots$$

$$\boldsymbol{z}_{Lt} = \text{Block}_L(\boldsymbol{z}_{L-1t} | \boldsymbol{Z}_{L-1<t}, \boldsymbol{X}^I, \boldsymbol{X}^A, \boldsymbol{X}^F, \boldsymbol{X}^O) \tag{4}$$

where $\boldsymbol{z}_{\ell t}$ is the output of the ℓ^{th} transformer block at time step t. The final block's output $\boldsymbol{z}_{Lt}$ is used to estimate $p(y_t)$, the probability of generating the tth token in the vocabulary via adaptive softmax [16]:

$$p(y_t) = \text{AdaptiveSoftmax}(\boldsymbol{z}_{Lt}) \tag{5}$$

By dividing the vocabulary into three clusters based on frequency—5K, 15K, and 30K—adaptive softmax makes training more efficient since most of the time, the decoder only needs to compute the softmax over the first cluster containing the 5,000 most common tokens.

In the following two subsections, we will describe the transformer block in detail. In each block, the conditioning on past tokens is achieved using dynamic convolutions, and the conditioning on the contexts is achieved using multi-head attention.

Dynamic Convolutions: Introduced by Wu *et al*. [52], the goal of dynamic convolution is to provide a more efficient alternative to self-attention [47] when attending to past tokens. At block $\ell + 1$ and time step t, we have the input $\boldsymbol{z}_{\ell t} \in \mathbb{R}^{D^E}$. Given kernel size K and H attention heads, for each head $h \in \{1, 2, ..., H\}$, we first project the current and last $K - 1$ steps using a feedforward layer to obtain $\boldsymbol{z}'_{\ell hj} \in \mathbb{R}^{D^E/H}$:

$$\boldsymbol{z}'_{\ell hj} = \text{GLU}(\boldsymbol{W}_{\ell h}^Z \, \boldsymbol{z}_{\ell j} + \boldsymbol{b}_{\ell h}^Z) \tag{6}$$

for $j \in \{t - K + 1, t - K + 2, ..., t\}$. Here GLU is the gated linear unit activation function [7]. The output of each head's dynamic convolution is the weighted sum of these projected values:

$$\tilde{\boldsymbol{z}}_{\ell ht} = \sum_{j=t-K+1}^{t} \gamma_{\ell hj} \, \boldsymbol{z}'_{\ell hj} \tag{7}$$

where the weight $\gamma_{\ell hj}$ is a linear projection of the input (hence the term "dynamic"), followed by a softmax over the kernel window:

$$\gamma_{\ell hj} = \text{Softmax}\left((\boldsymbol{w}_{\ell h}^\gamma)^T \, \boldsymbol{z}'_{\ell hj}\right) \tag{8}$$

The overall output is the concatenation of all the head outputs, followed by a feedforward with a residual connection and layer normalization [2], which does a z-score normalization across the feature dimension (instead of the batch dimension as in batch normalization [18]):

$$\tilde{\boldsymbol{z}}_{\ell t} = [\tilde{\boldsymbol{z}}_{\ell 1t}, \tilde{\boldsymbol{z}}_{\ell 2t}, ..., \tilde{\boldsymbol{z}}_{\ell Ht}] \tag{9}$$

$$\boldsymbol{d}_{\ell t} = \text{LayerNorm}\left(\boldsymbol{z}_{\ell t} + \boldsymbol{W}_\ell^{\tilde{z}} \, \tilde{\boldsymbol{z}}_{\ell t} + \boldsymbol{b}_\ell^{\tilde{z}}\right) \tag{10}$$

The output $\boldsymbol{d}_{\ell t}$ can now be used to attend over the context embeddings.

Multi-Head Attention: The multi-head attention mechanism [47] has been the standard method to attend over encoder outputs in transformers. In our setting, we need to attend over four context domains—images, text, faces, and objects. As an example, we will go over the image attention module, which consists of H heads. Each head h first does a linear projection of $\boldsymbol{d}_{\ell t}$ and the image embeddings $\boldsymbol{X}^I$ into a query $\boldsymbol{q}_{\ell ht}^I \in \mathbb{R}^{D^E/H}$, a set of keys $\boldsymbol{K}_{\ell ht}^I = \{\boldsymbol{k}_{\ell hti} \in \mathbb{R}^{D^E/H}\}_{i=1}^{M^I}$, and the corresponding values $\boldsymbol{V}_{\ell ht}^I = \{\boldsymbol{v}_{\ell hti}^I \in \mathbb{R}^{D^E/H}\}_{i=1}^{M^I}$:

$$\boldsymbol{q}_{\ell ht}^I = \boldsymbol{W}_{\ell h}^{IQ} \, \boldsymbol{d}_{\ell t} \tag{11}$$

$$\boldsymbol{k}_{\ell hi}^I = \boldsymbol{W}_{\ell h}^{IK} \, \boldsymbol{x}_i^I \qquad \forall i \in \{1, 2, ..., M^I\} \tag{12}$$

$$\boldsymbol{v}_{\ell hi}^I = \boldsymbol{W}_{\ell h}^{IV} \, \boldsymbol{x}_i^I \qquad \forall i \in \{1, 2, ..., M^I\} \tag{13}$$

Then the attended image for each head is the weighted sum of the values, where the weights are obtained from the dot product between the query and key:

$$\lambda_{\ell hti}^I = \text{Softmax}\left(\boldsymbol{K}_{\ell h}^I \, \boldsymbol{q}_{\ell ht}^I\right)_i \tag{14}$$

$$\boldsymbol{x}'^I_{\ell ht} = \sum_{i=1}^{M^I} \lambda_{\ell hti}^I \, \boldsymbol{v}_{\ell hti}^I \tag{15}$$

The attention from each head is then concatenated into $\boldsymbol{x}'^{I}_{\ell t} \in \mathbb{R}^{D^E}$:

$$\boldsymbol{x}'^{I}_{\ell t} = [\tilde{\boldsymbol{x}}^{I}_{\ell 1 t}, \tilde{\boldsymbol{x}}^{I}_{\ell 2 t}, ..., \tilde{\boldsymbol{x}}^{I}_{\ell H t}] \tag{16}$$

and the overall image attention $\tilde{\boldsymbol{x}}^{I}_{\ell t} \in \mathbb{R}^{D^E}$ is obtained after adding a residual connection and layer normalization:

$$\tilde{\boldsymbol{x}}^{I}_{\ell t} = \text{LayerNorm}(\boldsymbol{d}_{\ell t} + \boldsymbol{x}'^{I}_{\ell t}) \tag{17}$$

We use the same multi-head attention mechanism (with different weight matrices) to obtain the attended article $\tilde{\boldsymbol{x}}^{A}_{\ell t}$, faces $\tilde{\boldsymbol{x}}^{F}_{\ell t}$, and objects $\tilde{\boldsymbol{x}}^{O}_{\ell t}$. These four are finally concatenated and fed through a feedforward layer:

$$\tilde{\boldsymbol{x}}^{C}_{\ell t} = [\tilde{\boldsymbol{x}}^{I}_{\ell t}, \tilde{\boldsymbol{x}}^{A}_{\ell t}, \tilde{\boldsymbol{x}}^{F}_{\ell t}, \tilde{\boldsymbol{x}}^{O}_{\ell t}] \tag{18}$$

$$\tilde{\boldsymbol{x}}^{C'}_{\ell t} = \boldsymbol{W}^{C}_{\ell}\, \tilde{\boldsymbol{x}}^{C}_{\ell t} + \boldsymbol{b}^{C}_{\ell} \tag{19}$$

$$\tilde{\boldsymbol{x}}^{C''}_{\ell t} = \text{ReLU}(\boldsymbol{W}^{C'}_{\ell}\, \tilde{\boldsymbol{x}}^{C'}_{\ell t} + \boldsymbol{b}^{C'}_{\ell}) \tag{20}$$

$$\boldsymbol{z}_{\ell+1\, t} = \text{LayerNorm}(\tilde{\boldsymbol{x}}^{C'}_{\ell t} + \boldsymbol{W}^{C''}_{\ell}\, \tilde{\boldsymbol{x}}^{C''}_{\ell t} + \boldsymbol{b}^{C''}_{\ell}) \tag{21}$$

The final output $\boldsymbol{z}_{\ell+1\, t} \in \mathbb{R}^{D^E}$ is used as the input to the next transformer block.

4. News Image Captioning Datasets

We describe two datasets that contain news articles, images, and captions. The first dataset, GoodNews, was recently proposed in Biten *et al*. [3], while the second dataset, NYTimes800k, is our contribution.

GoodNews: The GoodNews dataset was previously the largest dataset for news image captioning [3]. Each example in the dataset is a triplet containing an article, an image, and a caption. Since only the article text, captions, and image URLs are publicly released, the images need to be downloaded from the original source. Out of the 466K image URLs provided by [3], we were able to download 463K images, or 99.2% of the original dataset—the remaining are broken links.

We use this 99.2% sample of GoodNews and the train-validation-test split provided by [3]. There are 421K training, 18K validation, and 23K test captions. Note that this split was performed at the level of captions, so it is possible for a training and test caption to share the same article text (since articles have multiple images).

We observe several issues with GoodNews that may limit a system's ability to generate high-quality captions. Many of the articles in GoodNews are partially extracted because the generic article extraction library failed to recognize some of the HTML tags specific to The New York Times. Importantly, the missing text often included the first few paragraphs which frequently contain important information for captioning images. In addition GoodNews contains some non-English articles and captioned images from the recommendation sidebar which are not related to the main article.

Table 1: Summary of news captioning datasets

	GoodNews	NYTimes800k
Number of articles	257 033	444 914
Number of images	462 642	792 971
Average article length	451	974
Average caption length	18	18
Collection start month	Jan 10	Mar 05
Collection end month	Mar 18	Aug 19
% of caption words that are		
– nouns	16%	16%
– pronouns	1%	1%
– proper nouns	23%	22%
– verbs	9%	9%
– adjectives	4%	4%
– named entities	27%	26%
– people's names	9%	9%
% of captions with		
– named entities	97%	96%
– people's names	68%	68%

NYTimes800k: The aforementioned issues motivated us to construct NYTimes800k, a 70% larger and more complete dataset of New York Times articles, images, and captions. We used The New York Times public API[2] for the data collection and developed a custom parser to resolve the missing text issue in GoodNews. The average article in NYTimes800k is 963 words long, whereas the average article in GoodNews is 451 words long. Our parser also ensures that NYTimes800k only contains English articles and images that are part of the main article. Finally, we also collect information about where an image is located in the corresponding article. Most news articles have one image at the top that relates to the key topic. However 39% of the articles have at least one more image somewhere in the middle of text. The image placement and the text surrounding the image is important information for captioning as we will show in our evaluations. Table 1 presents a comparison between GoodNews and NYTimes800k.

Entities play an important role in NYTimes800k, with 97% of captions containing at least one named entity. The most popular entity type are names of people, comprising a third of all named entities (see the supplementary material for a detailed breakdown of entity types). Furthermore, 71% of training images contain at least one face and 68% of training captions mention at least one person's name. Figure 3 provides a further breakdown of the co-occurrence of faces and people's names. One important observation is that 99% of captions contain at most four names.

[2] https://developer.nytimes.com/apis

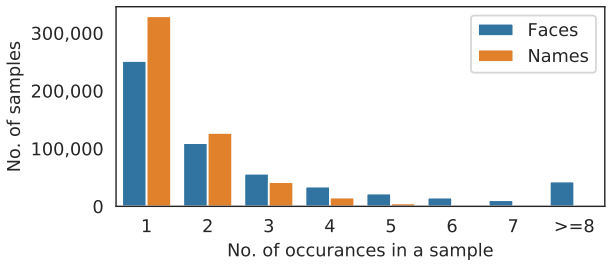

Figure 3: Co-occurrence of faces and people's names in NYTimes800k training data. The blue bars count how many images containing a certain number of faces. The orange bars count how many captions containing a certain number of people's names.

We split the training, validation, and test sets according to time, as shown in Table 2. Compared to the random split used in GoodNews, splitting by time allows us to study the model performance on novel news events and new names, which might be important in a deployment scenario. Out of the 100K proper nouns in our test captions, 4% never appear in any training captions.

5. Experiments

This section describes settings for neural network learning, baselines and evaluation metrics, followed by a discussion of key results.

5.1. Training Details

Following Wu *et al*. [52], we set the hidden size D^E to 1024; the number of heads H to 16; and the number of transformer blocks L to four with kernel sizes 3, 7, 15, and 31, respectively. For parameter optimization we use the adaptive gradient algorithm Adam [21] with the following parameter: $\beta_1 = 0.9, \beta_2 = 0.98, \epsilon = 10^{-6}$. We warm up the learning rate in the first 5% of the training steps to 10^{-4}, and decay it linearly afterwards. We apply L_2 regularization to all network weights with a weight decay of 10^{-5} and using the fix [28] that decouples the learning rate from the regularization parameter. We clip the gradient norm at 0.1. We use a maximum batch size of 16 and training is stopped after the model has seen 6.6 million examples. This is equivalent to 16 epochs on GoodNews and 9 epochs on NYTimes800k.

The training pipeline is written in PyTorch [34] using the AllenNLP framework [15]. The RoBERTa model and dynamic convolution code are adapted from fairseq [32]. Training is done with mixed precision to reduce the memory footprint and allow our full model to be trained on a single GPU. The full model takes 5 days to train on one Titan V GPU and has 200 million trainable parameters—see the supplementary material for the size of each model variant.

Table 2: NYTimes800k training, validation, and test splits

	Training	Validation	Test
Number of articles	433 561	2 978	8 375
Number of images	763 217	7 777	21 977
Start month	Mar 15	May 19	Jun 19
End month	Apr 19	May 19	Aug 19

5.2. Evaluation Metrics

We use BLEU-4 [33] and CIDEr [48] scores as they are standard for evaluating image captions. These are obtained using the COCO caption evaluation toolkit[3]. The supplementary material additionally reports BLEU-1, BLEU-2, BLEU-3, ROUGE [26], and METEOR [8]. Note that CIDEr is particularly suited for evaluating news captioning models as it puts more weight than other metrics on uncommon words. In addition, we evaluate the precision and recall on named entities, people's names, and rare proper names. Named entities are identified in both the ground-truth captions and the generated captions using SpaCy. We then count exact string matches between the ground truths and generated entities. For people's names we restrict the set of named entities to those marked as PERSON by the SpaCy parser. Rare proper nouns are nouns that appear in a test caption but not in any training caption.

5.3. Baselines and Model Variants

We show two previous state-of-the-art models: *Biten (Avg + CtxIns)* and *Biten (TBB + AttIns)* [3]. To provide a fair comparison we used the full caption results released by Biten *et al*. [3] and re-evaluated with our evaluation pipeline on a slightly smaller test set (a few test images are no longer available due to broken URLs). The final metrics are the same as originally reported if rounded to the nearest whole number.

We evaluate a few key modeling choices: the decoder type (*LSTM* vs *Transformer*), the text encoder type (*GloVe* vs *RoBERTa* vs *weighted RoBERTa*), and the additional context domains (*location-aware*, *face attention*, and *object attention*). The *location-aware* models select the 512 tokens surrounding the image instead of the first 512 tokens of the article. Note that all our models use BPE in the decoder with adaptive softmax. We ensure that the total number of trainable parameters for each model is within 7% of one another (148 million to 159 million), with the exception of *face attention* (171 million) and *object attention* (200 million) since the latter two have extra multi-head attention modules. The results reported over GoodNews are based on a model trained solely on GoodNews, using the original random split of [3] for easier comparison to previous work.

[3] https://github.com/tylin/coco-caption

Table 3: Results on GoodNews (rows 1–10) and NYTimes800k (rows 11–19). We report BLEU-4, ROUGE, CIDEr, and precision (P) & recall (R) of named entities, people's names, and rare proper nouns. Precision and recall are expressed as percentages. Rows 1–2 contain previous state-of-the-art results [3]. Rows 3–5 and 11–13 are ablation studies where we swap the Transformer with an LSTM and/or RoBERTa with GloVe. These models only have the image attention (IA). Rows 6 & 14 are our baseline RoBERTa transformer language model that only has the article text (and not the image) as inputs. Building on top of this, we first add attention over image patches (rows 7 & 15). We then take a weighted sum of the RoBERTa embeddings (rows 8 & 16) and attend to the text surrounding the image instead of the first 512 tokens of the article (row 17). Finally we add attention over faces (rows 9 & 18) and objects (rows 10 & 19) in the image.

		BLEU-4	ROUGE	CIDEr	Named entities		People's names		Rare proper nouns	
					P	R	P	R	P	R
GoodNews	(1) Biten (Avg + CtxIns) [3]	0.89	12.2	13.1	8.23	6.06	9.38	6.55	1.06	12.5
	(2) Biten (TBB + AttIns) [3]	0.76	12.2	12.7	8.87	5.64	11.9	6.98	1.58	12.6
	(3) LSTM + GloVe + IA	1.97	13.6	13.9	10.7	7.09	9.07	5.36	0	0
	(4) Transformer + GloVe + IA	3.48	17.0	25.2	14.3	11.1	14.5	10.5	0	0
	(5) LSTM + RoBERTa + IA	3.45	17.0	28.6	15.5	12.0	16.4	12.4	2.75	8.64
	(6) Transformer + RoBERTa	4.60	18.6	40.9	19.3	16.1	24.4	18.7	10.7	18.7
	(7) + image attention	5.45	20.7	48.5	21.1	17.4	26.9	20.7	12.2	20.9
	(8) + weighted RoBERTa	6.0	21.2	53.1	21.8	18.5	28.8	22.8	**16.2**	26.0
	(9) + face attention	**6.05**	**21.4**	**54.3**	22.0	18.6	**29.3**	**23.3**	15.5	24.5
	(10) + object attention	**6.05**	**21.4**	53.8	**22.2**	**18.7**	29.2	23.1	15.6	**26.3**
NYTimes800k	(11) LSTM + GloVe + IA	1.77	13.1	12.1	10.2	7.24	8.83	5.73	0	0
	(12) Transformer + GloVe + IA	2.75	15.9	20.3	13.2	10.8	13.2	9.66	0	0
	(13) LSTM + RoBERTa + IA	3.29	16.1	24.9	15.1	12.9	17.7	14.4	7.47	9.50
	(14) Transformer + RoBERTa	4.26	17.3	33.9	17.8	16.3	23.6	19.7	21.1	16.7
	(15) + image attention	5.01	19.4	40.3	20.0	18.1	28.2	23.0	24.3	19.3
	(16) + weighted RoBERTa	5.75	19.9	45.1	21.1	19.6	29.7	25.4	29.6	22.8
	(17) + location-aware	6.36	21.4	52.8	24.0	21.9	35.4	30.2	33.8	**27.2**
	(18) + face attention	6.26	21.5	53.9	24.2	22.1	36.5	30.8	33.4	26.4
	(19) + object attention	**6.30**	**21.7**	**54.4**	**24.6**	**22.2**	**37.3**	**31.1**	**34.2**	27.0

5.4. Results and Discussion

Table 5 summarizes evaluation metrics on GoodNews and NYTimes800k, while Figure 4 compares generated captions from different model variants. Our full model (row 10) performs substantially better than the existing state of the art [3] across all evaluation metrics. On GoodNews, the full model yields a CIDEr score of 53.8, whereas the previous state of the art [3] achieved a CIDEr score of only 13.1.

Our most basic LSTM model (row 3) differs from Biten *et al*. [3] in that we use BPE in the caption decoder instead of template generation and filling. The slight improvement in CIDEr (from 13.1 to 13.9) shows that BPE offers a competitive end-to-end alternative to the template filling method. This justifies the use of BPE in the remaining experiments.

Models that encode articles using GloVe embeddings (rows 3–4 and 11–12) are unable to generate rare proper nouns, giving a precision and recall of 0. This is because the encoder skips words that are not part of the fixed GloVe vocabulary. This motivates the switch from GloVe to RoBERTa, which has an unbounded vocabulary. This switch shows a clear advantage in rare proper noun generation. On NYTimes800k, even the worst performing model that uses RoBERTa (row 13) achieves a precision of 7.47%, a recall of 9.50%, and a CIDEr gap of 12.8 points over the model without RoBERTa (row 11).

Another important modeling choice is the functional form of the caption decoder. We find that the Transformer architecture provides a substantial improvement over the LSTM with respect to all evaluation metrics. For example, when we swap the LSTM with a Transformer (from row 13 to 15), the CIDEr score on NYTimes800k jumps from 24.9 to 40.3.

Adding attention over faces improves both the recall and precision of people's names. It has no significant effect on other entity types (see the supplementary material for

Japan Desperately Needs More Day Care Workers. New Mothers Need Not Apply.

TOKYO — Ever since she was a young girl, all Erica Takato wanted to do was work with small children. A few weeks into her term, she requested time off for bed rest ordered by her doctor.

....

Union officials and former teachers cite a major obstacle to the aspirations:

....

Ground-truth caption	A nursery school teacher showing a bug to his class.
Transformer + RoBERTa	Ms. Takato, who was born in Japan, was forced out of the day care program because she was pregnant.
+ image attention	Ms. Takato with her son, Kishiko, and their children, from left, Kaiti, 3, and Kaitama, 3, at a day care center in Tokyo.
+ weighted RoBERTa	Ms. Takato, with her son, Shiro, and son, at home in Tokyo. Ms. Takato, who was pregnant, said she was "so frustrated and lost hope of being able to work."
+ location-aware	A day care center in Tokyo.
+ face attention	A child care center in Tokyo. The government is eager to bring more women into the work force, and is trying to come up with enough child care for mothers.
+ object attention	A day care worker in Tokyo. The government is trying to bring more women into the work force, and the government is trying to come up with enough child care for mothers to go back to work.

Figure 4: An example article (left) and the corresponding news captions (right) from the NYTimes800k test set. The model with no access to the image makes a sensible but incorrect guess that the image is about Ms. Takato. Since the image appears in the middle of the article, only the location-aware models correctly state that the focus of the image is on a day care center.

a detailed breakdown). Importantly, people's names are the most common entity type in news captions and so we also see an improvement in CIDEr. Attention over objects also improves performance on most metrics, especially on NYTimes800k. More broadly, this result suggests that introducing specialized vision models tuned to the common types of objects such as organizations (via logos or landmarks) is a promising future direction to improve the performance on news image captioning.

The location-aware models (rows 17–19) focus the article context using the image location in the article, information which is only available in our NYTimes800k dataset. This simple focusing of context offers a big improvement to CIDEr, from 45.1 (row 16) to 52.8 (row 17). This suggests a strong correspondence between an image and the closest text that can be easily exploited to generate better captions.

The supplementary material additionally reports three caption quality metrics: caption length, type-token ratio (TTR) [45], and Flesch reading ease (FRE) [14, 20]. TTR is the ratio of the number of unique words to the total number of words in a caption. The FRE takes into account the number of words and syllables and produces a score between 0 and 100, where higher means being easier to read. As measured by FRE, captions generated by our model exhibit a level of language complexity that is closer to the ground truths. Additionally, captions generated by our model are 15 words long on average, which is closer to the ground-truths (18 words) than those generated by the previous state of the art (10 words) [3].

6. Conclusion

In this paper, we have shown that by using a carefully selected novel combination of the latest techniques drawn from multiple sub-fields within machine learning, we are able to set a new SOTA for news image captioning. Our model can incorporate real-world knowledge about entities across different modalities and generate text with better linguistic diversity. The key modeling components are byte-pair encoding that can output any word, contextualized embeddings for article text, specialized face & object encoding, and transformer-based caption generation. This result provides a promising step for other image description tasks with contextual knowledge, such as web pages, social media feeds, or medical documents. Promising future directions include specialized visual models for a broader set of entities like countries and organizations, extending the image context from the current article to recent or linked articles, or designing similar techniques for other image and text domains.

Acknowledgement

This research was supported in part by the Data to Decisions Cooperative Research Centre whose activities are funded by the Australian Commonwealth Government's Cooperative Research Centres Programme. The research was also supported in part by the Australian Research Council through project number DP180101985. We thank NVIDIA for providing us with Titan V GPUs through their GPU Grant Program.

References

[1] 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 visual question answering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 2

[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 4

[3] Ali Furkan Biten, Lluis Gomez, Marcal Rusinol, and Dimosthenis Karatzas. Good news, everyone! context driven entity-aware captioning for news images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1, 2, 5, 6, 7, 8, 12, 13, 14, 15

[4] Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels, Oct. 2018. Association for Computational Linguistics. 2

[5] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. VGGFace2: A dataset for recognising faces across pose and age. *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 67–74, 2017. 3

[6] Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Show, control and tell: A framework for generating controllable and grounded captions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1

[7] Yann N. Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 933–941, Sydney, Australia, 06–11 Aug 2017. PMLR. 4

[8] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics. 6, 12

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 2, 3

[10] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 1, 2

[11] Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia, July 2018. Association for Computational Linguistics. 2

[12] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K. Srivastava, Li Deng, Piotr Dollar, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, and Geoffrey Zweig. From captions to visual concepts and back. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 1

[13] Yansong Feng and Mirella Lapata. Automatic caption generation for news images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(4):797–812, April 2013. 2

[14] Rudolph Flesch. A new readability yardstick. *Journal of applied psychology*, 32(3):221, 1948. 8, 12

[15] Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia, July 2018. Association for Computational Linguistics. 6

[16] Édouard Grave, Armand Joulin, Moustapha Cissé, David Grangier, and Hervé Jégou. Efficient softmax approximation for GPUs. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1302–1310, Sydney, Australia, 06–11 Aug 2017. PMLR. 4

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2

[18] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 448–456, Lille, France, 07–09 Jul 2015. PMLR. 4

[19] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 1, 2

[20] J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. 1975. 8, 12

[21] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. 6

[22] Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. *ArXiv*, abs/1901.07291, 2019. 2

[23] Zhen-Zhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite

BERT for self-supervised learning of language representations. *ArXiv*, abs/1909.11942, 2019. 2

[24] Dianqi Li, Qiuyuan Huang, Xiaodong He, Lei Zhang, and Ming-Ting Sun. Generating diverse and accurate visual captions by comparative adversarial learning. *arXiv preprint arXiv:1804.00861*, 2018. 1

[25] Jiangyun Li, Peng Yao, Longteng Guo, and Weicun Zhang. Boosted transformer for image captioning. *Applied Sciences*, 9(16):3260, 2019. 2

[26] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. 6, 12

[27] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar S. Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke S. Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *ArXiv*, abs/1907.11692, 2019. 2, 3

[28] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 6

[29] Di Lu, Spencer Whitehead, Lifu Huang, Heng Ji, and Shih-Fu Chang. Entity-aware image caption generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4013–4023, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics. 2

[30] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1, 2

[31] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., 2013. 3

[32] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 6

[33] 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 of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. 6, 12

[34] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. In *NIPS Autodiff Workshop*, 2017. 6

[35] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. 3

[36] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. 2, 3

[37] 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*, 40:1072–1085, 2016. 1, 2

[38] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *ArXiv*, abs/1804.02767, 2018. 3

[39] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1

[40] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 3

[41] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, Aug. 2016. Association for Computational Linguistics. 2, 3

[42] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia, July 2018. Association for Computational Linguistics. 2

[43] Sandeep Subramanian, Raymond Li, Jonathan Pilault, and Christopher Joseph Pal. On extractive and abstractive neural document summarization with transformer language models. *ArXiv*, abs/1909.03186, 2019. 2

[44] A. Tariq and H. Foroosh. A context-driven extractive framework for generating realistic image descriptions. *IEEE Transactions on Image Processing*, 26(2):619–632, Feb 2017. 1, 2

[45] Mildred C Templin. Certain language skills in children; their development and interrelationships. 1957. 8, 12

[46] Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy, July 2019. Association for Computational Linguistics. 3

[47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017. 4

[48] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 6, 12

[49] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 1, 2

[50] 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, 2016. 1

[51] Weixuan Wang, Zhihong Chen, and Haifeng Hu. Hierarchical attention network for image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8957–8964, 2019. 2

[52] Felix Wu, Angela Fan, Alexei Baevski, Yann Dauphin, and Michael Auli. Pay less attention with lightweight and dynamic convolutions. In *International Conference on Learning Representations*, 2019. 4, 6

[53] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015. 2, 3

[54] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. XLNet: Generalized autoregressive pretraining for language understanding. *ArXiv*, abs/1906.08237, 2019. 2

[55] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2

[56] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23:1499–1503, 2016. 3

[57] Sanqiang Zhao, Piyush Sharma, Tomer Levinboim, and Radu Soricut. Informative image captioning with external sources of information. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6485–6494, Florence, Italy, July 2019. Association for Computational Linguistics. 2

[58] Xinxin Zhu, Lixiang Li, Jing Liu, Haipeng Peng, and Xinxin Niu. Captioning transformer with stacked attention modules. *Applied Sciences*, 8(5):739, 2018. 2

7. Supplementary Material

7.1. Live Demo

A live demo of our model is available at https://transform-and-tell.ml. In the demo, the user is able to provide the URL to a New York Times article. The server will then scrape the web page, extract the article and image, and feed them into our model to generate a caption.

7.2. Entity Distribution

Figure 5 shows how different name entity types are distributed in the training captions of the NYTimes800k dataset. The four most popular types are people's names (PERSON), geopolitical entities (GPE), organizations (ORG), and dates (DATE). Out of these, people's names comprise a third of all named entities. This motivates us to add a specialized face attention module to the model.

7.3. Model Complexity

Table 4: Model complexity. See Table 3 caption in the main paper for more explanation of each model variant.

	No. of Parameters
LSTM + GloVe + IA	157M
Transformer + GloVe + IA	148M
LSTM + RoBERTa + IA	159M
Transformer + RoBERTa	125M
+ image attention (IA)	154M
+ weighted RoBERTa	154M
+ location-aware	154M
+ face attention	171M
+ object attention	200M

Table 4 shows the number of training parameters in each of our model variants. We ensure that the total number of trainable parameters for each model is within 7% of one another (148 million to 159 million), with the exception of the model with face attention (171 million) and with object attention (200 million) since the latter two have extra multi-head attention modules.

7.4. Further Experimental Results

Table 5 reports BLEU-1, BLEU-2, BLEU-3, BLEU-4 [33] ROUGE [26], METEOR [8], and CIDEr [48]. Our results display a strong correlation between all the metrics—a method that performs well on one metric tends to perform well on them all. Of particular interest is CIDEr since it uses Term Frequency Inverse Document Frequency (TF-IDF) to put more importance on less common words such as entity names. This makes CIDEr particularly well suited for evaluating news captions where uncommon words tend to be vitally important, e.g. people's names.

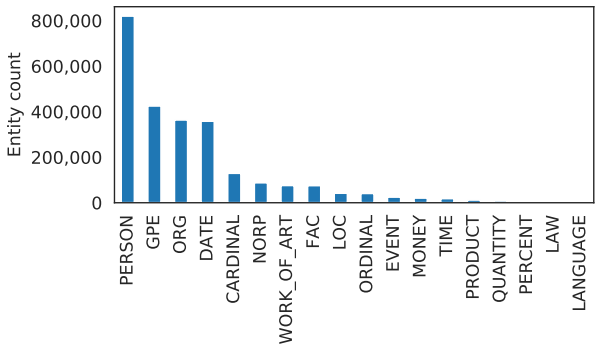

Figure 5: Entity distribution in NYTimes800k training captions. The four most common entity types are people's names, geopolitical entities, organizations, and dates.

Table 6 further reports metrics on the entities. In particular, we show the precision and recall of all proper nouns and new proper nouns. We define a proper noun to be new if it has never appeared in any training caption or article text. This is in contrast to the rare proper noun metrics reported in the main paper, which are proper nouns that are not present in any training caption but might have appeared inside a training article context.

The three rightmost columns of Table 6 show the linguistic quality metrics, including caption length (CL), type-token ratio (TTR) [45], and Flesch readability ease (FRE) [14, 20]. The TTR is measured as

$$\text{TTR} = \frac{U}{W} \tag{22}$$

where U is the number of unique words and W is the total number of words in the caption. FRE is measured as

$$\text{FRE} = 206.835 - 1.015\left(\frac{W}{S}\right) - 84.6\left(\frac{B}{W}\right) \tag{23}$$

where W is the number of words, S is the number of sentences, and B is the number of syllables in the caption.

The higher TTR corresponds to a higher vocabulary variation in the text, while a higher FRE indicates that the text uses simpler words and thus is easier to read. Overall our models produce captions that are closer in length to the ground truths than the previous state of the art *Biten* [3]. Moreover, our captions exhibit a level of language complexity (as measured by Flesch score) that is closer to the ground truths. However, there is still a gap in TTR, Flesch, and length, between captions generated by our model and the human-written ground-truth captions.

Finally Figure 6 and Figure 7 show two further set of generated captions.

Table 5: BLEU, ROUGE, METEOR, and CIDEr metrics on the GoodNews and NYTimes800k datasets.

		BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	METEOR	CIDEr
GoodNews	Biten (Avg + CtxIns) [3]	9.04	3.66	1.71	0.89	12.2	4.37	13.1
	Biten (TBB + AttIns) [3]	8.10	3.26	1.48	0.76	12.2	4.17	12.7
	LSTM + GloVe + IA	14.1	6.50	3.36	1.97	13.6	5.54	13.9
	Transformer + GloVe + IA	18.8	9.72	5.55	3.48	17.0	7.63	25.2
	LSTM + RoBERTa + IA	18.0	9.54	5.51	3.45	17.0	7.68	28.6
	Transformer + RoBERTa	19.7	11.3	6.96	4.60	18.6	8.82	40.9
	+ image attention	21.6	12.7	8.09	5.45	20.7	9.74	48.5
	+ weighted RoBERTa	22.3	13.4	8.72	6.0	21.2	10.1	53.1
	+ face attention	**22.4**	**13.5**	8.77	**6.05**	**21.4**	10.2	**54.3**
	+ object attention	**22.4**	**13.5**	**8.80**	**6.05**	**21.4**	**10.3**	53.8
NYTimes800k	LSTM + GloVe + IA	13.4	6.0	3.06	1.77	13.1	5.34	12.1
	Transformer + GloVe + IA	16.8	8.28	4.56	2.75	15.9	6.94	20.3
	LSTM + RoBERTa + IA	17.0	8.92	5.19	3.29	16.1	7.31	24.9
	Transformer + RoBERTa	18.2	10.2	6.37	4.26	17.3	8.14	33.9
	+ image attention	20.0	11.6	7.38	5.01	19.4	9.05	40.3
	+ weighted RoBERTa	20.9	12.5	8.18	5.75	19.9	9.56	45.1
	+ location-aware	21.8	13.5	8.96	6.36	21.4	10.3	52.8
	+ face attention	**21.6**	13.3	8.85	6.26	21.5	**10.3**	53.9
	+ object attention	**21.6**	**13.4**	**8.90**	**6.30**	**21.7**	**10.3**	**54.4**

Table 6: All proper noun and new proper noun precision (P) & recall (R) on the GoodNews and NYTimes800k datasets. Linguistic measures on the generated captions: caption length (CL), type-token ratio (TTR), and Flesch readability ease (FRE).

		All proper nouns		New proper nouns		CL	TTR	FRE
		P	R	P	R			
GoodNews	Ground truths	–	–	–	–	18.1	94.9	65.4
	Biten (Avg + CtxIns) [3]	16.5	12.2	2.70	12.0	9.89	92.2	78.3
	Biten (TBB + AttIns) [3]	19.2	11.0	4.21	12.3	9.14	90.7	77.6
	LSTM + GloVe + IA	16.1	11.3	0	0	14.0	89.5	77.2
	Transformer + GloVe + IA	22.7	18.4	0	0	16.0	88.4	73.9
	LSTM + RoBERTa + IA	25.1	20.8	1.68	7.86	15.0	89.0	75.7
	Transformer + RoBERTa	30.7	26.0	7.69	16.4	15.1	90.0	73.0
	+ image attention	33.4	28.0	8.53	19.3	15.2	90.0	72.5
	+ weighted RoBERTa	33.9	29.6	**15.2**	**24.4**	15.5	90.8	71.8
	+ face attention	34.3	29.8	13.6	22.2	15.4	90.8	71.8
	+ object attention	**34.7**	**29.9**	13.3	23.6	15.3	90.9	72.0
NYTimes800k	Ground truths	–	–	–	–	18.4	94.6	63.9
	LSTM + GloVe + IA	15.8	12.4	0	0	13.9	88.7	76.1
	Transformer + GloVe + IA	21.5	18.2	0	0	14.8	88.8	71.9
	LSTM + RoBERTa + IA	24.1	21.8	3.28	7.18	14.8	89.3	73.3
	Transformer + RoBERTa	28.0	26.0	13.4	14.5	15.2	90.4	71.4
	+ image attention	31.1	28.7	15.6	17.2	15.1	90.1	71.5
	+ weighted RoBERTa	31.8	30.5	21.7	20.2	15.5	91.6	70.1
	+ location-aware	36.4	34.1	26.3	**25.3**	15.1	91.7	70.8
	+ face attention	36.8	34.2	26.2	24.2	14.9	91.8	70.9
	+ object attention	**37.2**	**34.5**	**26.7**	25.1	14.8	91.9	71.2

Table 7: Geopolitical entity (GPE), organization (ORG), and date (DATE) precision (P) & recall (R) on the GoodNews and NYTimes800k datasets.

		GPE		ORG		DATE	
		P	R	P	R	P	R
GoodNews	Biten (Avg + CtxIns) [3]	12.0	11.5	5.67	7.45	6.12	4.03
	Biten (TBB + AttIns) [3]	12.8	8.41	5.81	7.36	5.86	4.06
	LSTM + GloVe + IA	15.6	12.8	14.0	8.58	11.0	8.20
	Transformer + GloVe + IA	20.8	18.8	16.6	11.8	12.0	10.1
	LSTM + RoBERTa + IA	20.8	19.2	16.9	12.3	13.4	10.9
	Transformer + RoBERTa	22.6	22.5	20.4	16.3	13.8	12.6
	+ image attention	**25.8**	24.5	21.0	17.3	14.4	13.0
	+ weighted RoBERTa	25.0	24.2	22.0	**18.7**	14.3	13.1
	+ face attention	24.9	24.4	21.6	18.5	14.7	**13.3**
	+ object attention	25.6	**24.7**	**22.4**	**18.7**	**15.1**	**13.3**
NYTimes800k	LSTM + GloVe + IA	16.0	14.7	8.60	4.89	11.3	8.31
	Transformer + GloVe + IA	19.1	21.8	12.1	7.95	11.3	10.1
	LSTM + RoBERTa + IA	20.2	22.2	13.1	8.95	11.8	11.1
	Transformer + RoBERTa	21.4	25.4	15.8	12.2	12.0	12.5
	+ image attention	23.9	27.3	17.6	13.6	12.8	13.2
	+ weighted RoBERTa	24.2	28.2	19.2	15.6	13.9	14.3
	+ location-aware	26.8	30.1	20.9	**17.3**	**14.1**	**14.1**
	+ face attention	**26.9**	**30.6**	20.7	16.5	13.9	**14.1**
	+ object attention	26.8	**30.6**	**21.9**	17.2	13.7	13.8

An Artist Making a Powerful Statement — by Creating Work About Herself

During the final days of her solo show at Kravets Wehby Gallery in Manhattan this past spring, the mixed-media artist Theresa Chromati had something to confess about her latest body of work. “I realized that you can’t hide from anything,” she said, staring up at the 2019 painting “We All Look Back At It (Morning Ride).” In it, a nude figure squats with her glittering, butterfly-adorned buttocks in the air and her unobscured face turned to look directly at the viewer.

For much of her career, Chromati, 26, depicted the naked bodies in her powerful portraits of black women behind protective disguises….

Ground-truth caption	The mixed-media artist Theresa Chromati sits in front of an unfinished and currently untitled acrylic painting at her Brooklyn studio.
LSTM + GloVe + IA	“Untitled (Bubs),” 2017, oil on canvas.
Transformer + GloVe + IA	“Untitled (The Red Rose)” (2015), a painting by Nina Arianda.
LSTM + RoBERTa + IA	“The B-N-1,” by the artist and artist Ms. Chastain.
Transformer + RoBERTa	The artist Theresa Cromati in her studio in Manhattan.
+ image attention	The artist Theresa Cromati in her studio in New York.
+ weighted RoBERTa	“I’m a woman who’s not going to be a woman,” said the artist Theresa Cromati, who has been working with her own work since 2017.
+ location-aware	Theresa Nemati, who has created a new work, in her studio in Brooklyn.
+ face attention	Theresa Cromati in her studio in Brooklyn.
+ object attention	The artist Theresa Cromati in her studio in Manhattan.

Figure 6: An example article (left) and the corresponding news captions (right) from the NYTimes800k test set. The name “Chromati” has never appeared in the training data, and none of the models can spell the artist’s name correctly. They all miss the letter “h” in her name. Captions from models that use an LSTM or GloVe contain made-up names for both the painting and the artist. Finally the model that has no access to the image, *Transformer + RoBERTa*, still guesses correctly that the image is about the artist being in her studio. This shows that NYTimes article images can have a predictable theme.

Bernie Sanders Heads to Canada for Affordable Insulin

WINDSOR, Ontario — Bernie Sanders wanted to make a point about a crippling injustice. So he crossed the border.

Well, the northern one.

On Sunday, he took about a dozen people with diabetes on a bus from Detroit to Windsor to get insulin at a Canadian pharmacy, just minutes from the border. Because of traffic, and multiple stops along the way, it took an hour and 17 minutes to get there and about the same time to get back. But the duration and the mileage were not really the main points.
....

Ground-truth caption	Mr. Sanders spoke for about four minutes outside the pharmacy.
LSTM + GloVe + IA	Senator Bernie Sanders of Vermont at a rally in the Bronx on Sunday.
Transformer + GloVe + IA	Senator Bernie Sanders of Vermont at a rally in Ottawa on Sunday.
LSTM + RoBERTa + IA	Senator Bernie Sanders of Vermont, center, with his wife, Ann, and son, Michael, at the Canadian border in Canada on Sunday.
Transformer + RoBERTa	Senator Bernie Sanders of Vermont, center, with a group of people with diabetes, at a Canadian pharmacy in Windsor, Ontario, on Sunday.
+ image attention	Senator Bernie Sanders of Vermont, who spoke to reporters on Sunday, took a group of people with diabetes to get a drug from Canada to Ottawa.
+ weighted RoBERTa	Senator Bernie Sanders of Vermont, center, at a rally in Detroit on Sunday.
+ location-aware	Mr. Sanders, center, and Ms. Nystrom, right, at a rally in Windsor, Ontario, on Wednesday.
+ face attention	Mr. Sanders spoke to reporters outside the pharmacy where he spoke about his son, Hunter, 22.
+ object attention	Mr. Sanders spoke to reporters outside the pharmacy in Windsor, Ontario, on Saturday.

Figure 7: An example article (left) and the corresponding news captions (right) from the NYTimes800k test set. The model that has no access to the image, *Transformer + RoBERTa*, is correct in predicting that the image is about Bernie Sanders. However it guesses that he is with a group of people with diabetes, which is not correct but is sensible given the article content. Some of the models manage to override the strong prior that he is at a rally (which is what many of Bernie Sanders images in the training set are about) and correctly say that he is outside a pharmacy. The caption from the model with object attention is the most accurate because it generates all three entities correctly: Windsor in Ontario, the reporters, and the pharmacy.