# MT: Multi-Modal Transformer for News Image Captions


**Fuxiao Liu**
University of Virginia
fl3es@virginia.edu

**Tianlu Wang**
University of Virginia
tianlu@virginia.edu

**Yinghan Wang**
University of Virginia
yw9fm@virginia.edu

**Vicente Ordonez**
University of Virginia
vicente@virginia.edu



## Abstract

In this paper, we develop a lightweight transformer model which can efficiently generate captions given images and associated news articles. Previous works present two limitations: they ignore the connection between multi-modal inputs during encoding; uncommon words, especially named entities, are waited to be predicted more accurately. We tackle the first challenge via introducing the visual selective layer and multi-modal attention. We address the second challenge by proposing the entity guide and tag cleaning operation. Empirical results on both the GoodNews and VisualNews datasets demonstrate the proposed architecture achieves state-of-the-art results while having significantly fewer parameters than competing methods (200M → 93M).



Figure 1: Examples from VisualNews dataset (Liu et al., 2020) (left) and COCO (Chen et al., 2015) (right). VisualNews provides more informative captions with name entities, whereas COCO contains more generic captions.

## 1 Introduction

Image captioning is a vision and language task which attracted considerable attention. While important progress has been made in recent years (Vinyals et al., 2015; Fang et al., 2015; Xu et al., 2015; Lu et al., 2018b; Anderson et al., 2018), these techniques working on generic captions lack real-world knowledge. For example, a caption such as *"A bunch of people who are holding red umbrellas."* properly describes the image at some level, but it fails to capture the higher level situation that is taking place in this picture i.e. *"why are people gathering with red umbrellas and what role do they play?"* This type of language is typical in describing events in news text.

News image captions are typically more complex than pure image captions and thus make them harder to generate. News captions describe the contents of images at a higher degree of specificity and as such contain many named entities referring to specific people, places, and organizations. Such named entities convey key information regarding the events presented in the images, and conversely events are often used to predict what types of entities are involved. e.g. if the news article mentions a baseball game then a picture might involve a baseball player or a coach, conversely if the image contains someone wearing baseball gear, it might imply that a game of baseball is taking place.

Previous works (Lu et al., 2018a; Biten et al., 2019) have attempted news image captioning by adopting a two-stage pipeline. They first replace all specific named entities with entity type tags to create templates and train a model to generate template captions with fillable placeholders. Then, these methods search in the input news articles for entities to fill placeholders. Such approach reduces the vocabulary size and eases the burden on the template generator network. However, just replacing all named entities will miss some key information since not all named entities are uncommon words. For example, "LeBron James" are popular word which contains rich information. (Tran et al., 2020) applied a transformer method and byte-pair-encoding (Sennrich et al., 2015) to address the

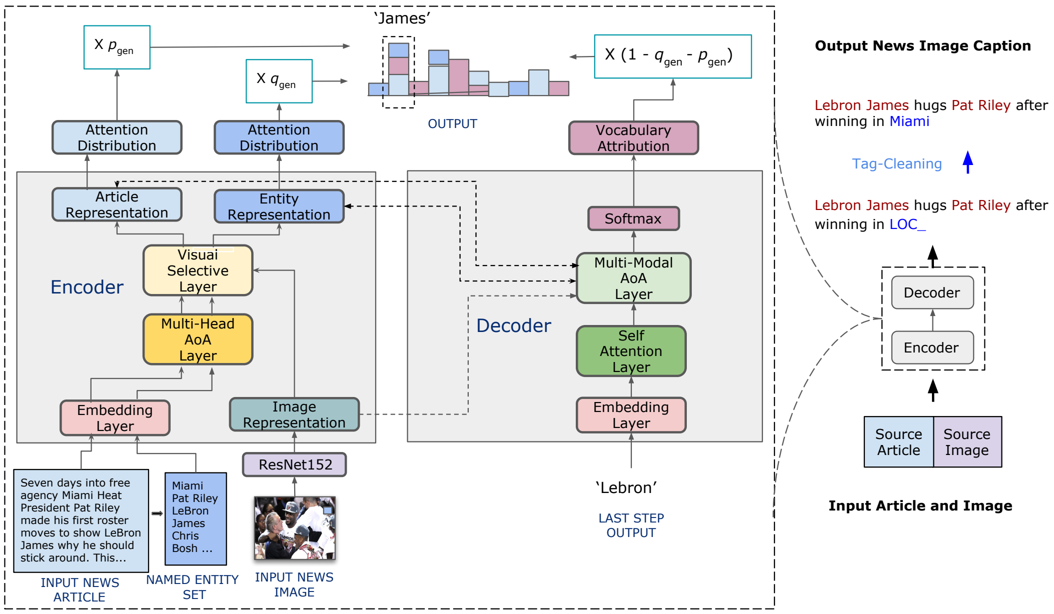



Figure 2: Overview of our model. Left: Details of the the encoder and decoder; Right: The workflow of our model. The input news article and news image are fed into the encoder-decoder system. The blue arrow denotes Tag-Cleaning step, which is a post-processing step to further improves the result during testing. Multi-Head AoA Layer means our Multi-Head Attention on Attention Layer. Multi-Modal AoA Layer means our Multi-Modal Attention on Attention Layer. Self Attention Layer denotes our Masked Multi-Head Attention on Attention Layer.

knowledge gap and linguistic gap. However, it ignores the connection between the article and image during encoding by only using pretrained encoders separately.

To overcome these challenges, we adapt the existing Tranformer (Vaswani et al., 2017) to news image datasets by integrating several critical components. we first propose a novel Visual Selective Layer and Multi-Modal attention mechanism to strengthen the connection between multi-modal features. To effectively attend to important named entities in news articles, we apply Attention on Attention (Huang et al., 2019) technique on attention layers and introduce a new position encoding method to model the relative position relationships of words. To avoid missing rare named entities, we introduce the entity guide mechanism and build our decoder upon the multi-source pointer-generator model.

In addition, news captions also contain a significant amount of words falling either in the long tail of the distribution, or resulting in out-of-vocabulary words at test time. In order to alleviate this, we introduce a tag cleaning post-processing step to further improve our model.

Our main contributions can be summarized as:

- We propose Multi-Modal Transformer, a captioning method for news images, showing superior results on the GoodNews (Biten et al., 2019) and VisualNews datasets (Liu et al., 2020) with much fewer parameters than competing methods.
- Our proposed Visual Selective Layer and Multi-Modal Attention Layer improves the generation of named entities for new image captions.
- We benchmarked both template-based and end-to-end captioning methods on two large scale news image datasets, revealing the challenges in the task of news image captioning.

## 2 Methodology

Figure 2 presents an overview of Multi-Modal Transformer. We first introduce the image encoder and the text encoder. We then explain the decoder in 2.3. To solve the out-of-vocabulary issue, we propose Tag-Cleaning in 2.4.

### 2.1 Image Encoder

We use a ResNet152 (He et al., 2016) pretrained on ImageNet (Deng et al., 2009) to extract visual features. The output of the convolutional layer before the final pooling layer gives us a set of vectors corresponding to different patches in the image. Specifically, we obtain features $V = \{v_1, \ldots, v_K\}, v_i \in \mathbb{R}^D$ from every image $I$, where $K = 49$ and $D = 2048$. With these features, we can selectively attend to different regions at different time steps.

## 2.2 Text Encoder

As the length of the associated article could be very long, we focus on the first 300 tokens in each article following (See et al., 2017). We also used the spaCy (Honnibal and Montani, 2017) named entity recognizer to extract named entities from news articles inspired by (Li et al., 2018). We encode the first 300 tokens and the extracted named entities using the same encoder. Given the input text $T = \{t_1, \ldots, t_L\}$ where $t_i$ denotes the i-th token in the text and $L$ is the text length, we use following layers to obtain textual features:

**Word Embedding and Position Embedding.** For each token $t_i$, we first obtain word embedding $w_i \in \mathbb{R}^H$ and positional embedding $p_i \in \mathbb{R}^H$ through two embedding layers, $H$ is the hidden state size and is set to 512. To better model the relative position relationships, we further feed position embeddings into a LSTM (Hochreiter and Schmidhuber, 1997) to get the updated position embedding $p_i^l \in \mathbb{R}^H$. We then add up $p_i^l$ and $w_i$ to obtain the final input embedding $w_i'$.

$$p_i^l = \text{LSTM}(p_i) \tag{1}$$

$$w_i' = w_i + p_i^l \tag{2}$$

**Multi-Head Attention on Attention Layer.** The Multi-Head Attention Layer (Vaswani et al., 2017) operates on three sets of vectors: queries $Q$, keys $K$ and values $V$, and takes a weighted sum of value vectors according to a similarity distribution between $Q$ and $K$. In our implementation, for each query $w_i'$, $K$ and $Q$ are all input embeddings $T'$ In addition, we have the "Attention on Attention" (AoA) module (Huang et al., 2019) to assist the generation of attended information:

$$v_{att} = \text{MHAtt}(w_i', T', T') \tag{3a}$$

$$g_{att} = \sigma(W_g[v_{att}; T']) \tag{3b}$$

$$v_{att}' = W_a[v_{att}; T'] \tag{3c}$$

$$\tilde{w}_i = g_{att} \odot v_{att}' \tag{3d}$$

where $\odot$ represents the element-wise multiplication operation and $\sigma$ is the sigmoid function. $W_g$ and $W_a$ are trainable parameters.

**Visual Selective Layer.** One limitation of previous works (Tran et al., 2020; Biten et al., 2019) is that they separately encode the image and article, ignoring the connection between them during encoding. In order to generate representations which can capture contextual information from both images and articles, we propose a novel Visual Selective Layer which updates textual embeddings with a visual information gate:

$$\overline{T} = \text{AvgPool}(\tilde{T}) \tag{4}$$

$$g_v = tanh(W_v(\text{MHAtt}_{\text{AoA}}(\overline{T}, V, V)) \tag{5}$$

$$w_i^* = g_v \odot \tilde{w}_i \tag{6}$$

$$w_i^a = \text{LayerNorm}(w_i^* + \text{FFN}(w_i^*)) \tag{7}$$

where $\text{MHAtt}_{\text{AoA}}$ corresponds to Eq 3. To obtain fixed-length article representations, we apply the average pooling operation to get $\overline{T}$, which can be used as the query to attend to different regions of the image. FFN is a two-layer feed-forward network with ReLU as the activation function. $w_i^a$ is the final output embedding from the text encoder. For the sake of simplicity, in the following text, we use $A = \{a_1, \ldots, a_L\}, a_i \in \mathbb{R}^H$ to represent the final embeddings ($w_i^a$) of article tokens, where $H$ is the embedding size and $L$ is the article length. Similarly, $E = \{e_1, \ldots, e_M\}, e_i \in \mathbb{R}^H$ represent the final embeddings of extracted named entities, where $M$ is the number of named entities.

## 2.3 Decoder

Our decoder generates the next token conditioned on previously generated tokens and contextual information. We propose Masked Multi-Head Attention on Attention Layer to flexibly attend to the previous tokens and Multi-Modal Attention on Attention Layer to fuse contextual information. We first use the encoder to obtain embeddings of ground truth captions $X = \{x_0, \ldots, x_N\}$, $x_i \in \mathbb{R}^H$, where $N$ is the caption length and $H$ is the embedding size. Instead of using the Masked Multi-Head Attention Layer in (Tran et al., 2020) to collect the information from past tokens, we use the more efficient Masked Multi-Head Attention on Attention Layer. At the time step $t$, the output embedding $x_t^a$ is used as the query to attend over the context information:

$$x_t^a = \text{MHAtt}_{\text{AoA}}^{\text{Masked}}(x_t, X, X) \tag{8}$$

**Multi-Modal Attention on Attention Layer.** Our Multi-Modal AoA Layer contains three context sources: images $\tilde{V}$, articles $A$ and name entity sets $E$. We use a linear layer to resize features in $V$ into $\tilde{V}$, where $\tilde{v} \in \mathbb{R}^{512}$. In each step, $x_i^a$ is the query that attends over them separately:

$$V'_t = \text{MHAtt}_{\text{AoA}}(x^a_t, \tilde{V}, \tilde{V}) \tag{9}$$
$$A'_t = \text{MHAtt}_{\text{AoA}}(x^a_t, A, A) \tag{10}$$
$$E'_t = \text{MHAtt}_{\text{AoA}}(x^a_t, E, E) \tag{11}$$

We combine the attended image feature $V'_t$, the attended article feature $A'_t$ and the attended named entity feature $E'_t$, and feed them into a residual connection, layer normalization and a two-layer feed-forward layer FFN.

$$C_t = V'_t + A'_t + E'_t \tag{12}$$
$$x'_t = \text{LayerNorm}(x^a_t + C_t) \tag{13}$$
$$x^*_t = \text{LayerNorm}(x'_t + \text{FFN}(x'_t)) \tag{14}$$
$$P_{s_t} = \text{softmax}(x^*_t) \tag{15}$$

The final output $P_{s_t}$ will be used to predict token $s_t$ in the Multi-Head Pointer-Generator Module.

**Multi-Head Pointer-Generator Module.** For the purpose of obtaining more related named entities from the associated article and the extracted named entity set, we adapt the pointer-generator (See et al., 2017). Our pointer-generator contains two sources: the article and named entity set. We first generate $a^V$ and $a^E$ over the source article tokens and extracted named entities by averaging the attention distributions from the multiple heads of the Multi-Modal Attention on Attention layer in the last decoder layer. Next, $p_{gen}$ and $q_{gen}$ are calculated as two soft switches to choose between generating a word from the vocabulary distribution $P_{s_t}$, or copying words from the attention distribution $a^V$ or $a^E$:

$$p_{gen} = \sigma(W_p([x_t; A_t; \tilde{V}_t])) \tag{16}$$
$$q_{gen} = \sigma(W_q([x_t; E_t; \tilde{V}_t])) \tag{17}$$

where $A_i$, $V_i$ and $E_i$ are attended context vector. $W_p$ and $W_q$ are learnable parameters. $\sigma$ is the sigmoid function. $P^*_{s_i}$ provides us with the final distribution to predict the next word.

$$\begin{aligned} P^*_{s_t} = p_{gen}a^V + q_{gen}a^E + \\ (1 - p_{gen} - q_{gen})P_{s_t} \end{aligned} \tag{18}$$

Finally, our loss can be computed as the sum of the negative log likelihood of the target word at each time step:

$$Loss = -\sum_{t=1}^{N} \log P^*_{s_i} \tag{19}$$

### 2.4 Tag-Cleaning

To solve out-of-vocabulary (*OOV*) problem, we replace *OOV* named entities with named entity tags instead of using a single “UNK” token, e.g. if “John Paul Jones Arena” is a *OOV* named entity, we replace it with “LOC_”, which represents location entities. During testing, if the model predicts entity tags, we further replace those tags with specific named entities. More specifically, we select a named entity with the same entity category and the highest frequency from the named entity set.

## 3 Experiments

In this section, we first introduce details of implementation. Then baselines and competing methods will be discussed. Lastly we present comprehensive experiment results on both GoodNewsdataset and our VisualNews dataset.

### 3.1 Implemention Details

**Datasets.** We conduct experiments on two large scale news image datasets: GoodNews (Biten et al., 2019) and VisualNews. For GoodNews, we follow the same splits introduced in Biten et al. (2019), which consists of $424,000$ training, $18,000$ validation and $23,000$ test samples. For VisualNews (Liu et al., 2020), we randomly sample $100,000$ images from each news agency, leading to a training set of $400,000$ samples. Similarly, we get a $40,000$ validation set and a $40,000$ test set, both evenly sampled from four news agencies of Visual News.

Throughout our experiments, we first resize images into $256 \times 256$, and randomly crop patches with size $224 \times 224$ as input. To preprocess captions and articles, we remove noisy HTML labels, brackets, non-ASCII characters and some special tokens. We use spaCy’s named entity recognizer (Honnibal and Montani, 2017) to recognize named entities in both captions and articles.

**Model Training.** We set the embedding size $H$ to 512. For dropout layers, we set the dropout rate as $0.1$. Models are optimized using Adam (Kingma and Ba, 2015) optimizer with a warming up learning rate set to $0.0005$. We use a batch size of 64 and stop training when the CIDEr (Vedantam et al., 2015) score on dev set is not improving for 20 epochs. Since we replace $OOV$ named entities with tags, we add 18 named entity tags provided by spaCy into our vocabulary, including "PERSON_", "LOC_", "ORG_", "EVENT_", etc.

| Model | Solve OOV | BLEU-4 | METEOR | ROUGE | CIDEr | P | R |
|---|---|---|---|---|---|---|---|
| TextRank (Barrios et al., 2016) | ✗ | 1.7 | 7.5 | 11.6 | 9.5 | 1.7 | 5.1 |
| Show Attend Tell (Xu et al., 2015) | ✗ | 0.7 | 4.1 | 11.9 | 12.2 | – | – |
| Tough-to-beat (Biten et al., 2019) | ✗ | 0.8 | 4.2 | 11.8 | 12.8 | 9.1 | 7.8 |
| Pooled Embeddings (Biten et al., 2019) | ✗ | 0.8 | 4.3 | 12.1 | 12.7 | 8.2 | 7.2 |
| Transform and Tell (Tran et al., 2020) | BPE | 6.0 | – | **21.4** | 53.8 | 22.2 | 18.7 |
| Our Transformer | ✗ | 5.2 | 7.9 | 19.5 | 48.4 | 20.8 | 17.5 |
| Our Transformer+EG | ✗ | 5.4 | 7.9 | 19.7 | 49.9 | 21.9 | 18.4 |
| Our Transformer+EG+Pointer | ✗ | 5.5 | 8.0 | 20.1 | 51.1 | 22.4 | 18.7 |
| Our Transformer+EG+Pointer+VG | ✗ | 5.7 | 8.1 | 20.2 | 52.5 | 22.4 | 18.8 |
| Our Transformer+EG+Pointer+VG+PE | ✗ | 6.0 | 8.2 | 20.5 | 53.7 | 22.5 | 18.9 |
| Our Transformer+EG+Pointer+VG+PE+TC | Tag-Cleaning | **6.1** | **8.3** | 20.9 | **55.4** | **22.9** | **19.3** |

Table 1: News image captioning results (%) on GoodNews dataset. EG means adding the named entity set as another text source guiding the generation of captions. Pointer means pointer-generator module. VS means the Visual Selective Layer. PE means adding our Position Embedding. TC means the Tag-Cleaning step.

| Model | Solve OOV | BLEU-4 | METEOR | ROUGE | CIDEr | P | R |
|---|---|---|---|---|---|---|---|
| TextRank (Barrios et al., 2016) | ✗ | 2.1 | 8.0 | 12.0 | 8.4 | 4.1 | 6.1 |
| Tough-to-beat (Biten et al., 2019) | ✗ | 1.7 | 4.6 | 13.2 | 12.4 | 4.9 | 4.8 |
| Pooled Embeddings (Biten et al., 2019) | ✗ | 2.1 | 5.2 | 13.5 | 13.2 | 5.3 | 5.3 |
| Our Transformer | ✗ | 4.9 | 7.7 | 16.8 | 45.6 | 18.5 | 16.1 |
| Our Transformer+EG | ✗ | 5.0 | 7.9 | 17.4 | 46.8 | 19.2 | 16.7 |
| Our Transformer+EG+Pointer | ✗ | 5.1 | 8.0 | 17.7 | 48.0 | 19.3 | 17.0 |
| Our Transformer+EG+Pointer+VS | ✗ | 5.1 | 8.1 | 17.8 | 48.6 | 19.4 | 17.1 |
| Our Transformer+EG+Pointer+VS+PE | ✗ | 5.2 | 8.2 | 17.8 | 49.2 | 19.4 | 17.2 |
| Our Transformer+EG+Pointer+VS+PE+TC | Tag-Cleaning | **5.3** | **8.2** | **17.9** | **50.5** | **19.7** | **17.6** |

Table 2: News image captioning results (%) on our VisualNews dataset.

**Evaluation Metrics.** Following previous literature, we evaluate the models' performance on two categories of metrics. To measure the overall similarity between generated captions and ground truth, we report BLEU-4 (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2014), ROUGE (Ganesan, 2018) and CIDEr (Vedantam et al., 2015) scores. Among these scores, CIDEr is the most suitable one for measuring news captioning since it downweighs stop words and focuses more on uncommon words through a TF-IDF weighting mechanism. On the other hand, to evaluate models ability to predict named entities, we compute the exact match precision and recall scores for named entities following Biten et al. (2019).

### 3.2 Competing Methods and Model Variants

We compare our proposed Multi-Modal Transformer with various baselines and competing methods.

**TextRank** (Barrios et al., 2016) is a graph-based extractive summarization algorithm. This baseline only takes the associated articles as input.

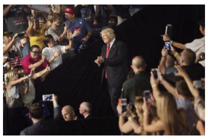
**Show Attend Tell** (Xu et al., 2015) tries to attend to certain image patches during caption generation. This baseline only takes images as input.

**Pooled Embeddings** and **Tough-to-beat** (Arora et al., 2017) are two template-based models proposed in Biten et al. (2019)[1]. They try to encode articles at the sentence level and attend to certain sentence at different time steps. *Pooled Embeddings* method computes sentence representations by averaging word embeddings and adopts context insertion in the second stage. *Tough-to-beat* obtains sentence representations from the tough-to-beat method introduced in Arora et al. (2017) and uses sentence level attention weights (Biten et al., 2019) to insert named entities.

**Transform and Tell** (Tran et al., 2020) is the transformer-based attention model, which uses a pretrained RoBERTa (Liu et al., 2019) model as the article encoder and a transformer as the decoder. It uses byte-pair encoding (BPE) to represent out-of-vocabulary named entities.

**Multi-Modal Transformer (MT)** is our proposed model, which is based on transformer (Vaswani et al., 2017). Our transformer adopts Multi-Head Attention on Attention (AoA). EG (Entity-Guide)

[1]Named as Avg+CtxIns and TBB+AttIns in the original paper.



Ground Truth:
republican presidential candidate **donald trump** enters **germain arena** to a packed house on **monday**
Multi-Modal Transformer:
**donald trump** supporters cheer as republiaan presidential candidate **donald trump** speaks in **germain arena**
Pooled Embeddings:
obama and his wife obama celebrate during the recent weeks EVENT_

Ground Truth:
**virginia cavaliers** fans celebrate on the court after the cavaliers game against the **duke blue devils** at **john paul jones arena**
Multi-Modal Transformer::
**virginia cavaliers** forward anthony gill celebrates with fans after the game against the **duke blue devils** at **john paul jones arena**
Pooled Embeddings
krzyzewski fans celebrate after the krzyzewski win over north carolina in the semifinals


Ground Truth:
president **obama** delivered his annual state of the union address on **tuesday** in **washington**
Multi-Modal Transformer::
president **obama** delivers the state of the union address on **tuesday** jan 20
Pooled Embeddings:
waldman speaks during a the white house news conference on year in **washington**

Ground Truth:
**sidney crosdy** celebrated his goal in the second period that seemed to deflate sweden
Multi-Modal Transformer::
**sidney crosby** of canada celebrating a goal in the men's gold medal game
Pooled Embeddings
**crosby** of canada after scoring the winning goal in the second period

Figure 3: Examples of captions generated by different models on two datasets. First three are from Visual News and the last one is from GoodNews. Correct named entities entities are highlighted in bold. Our Multi-Modal Transformer is able to predict the named entities more accurately and completely than the competing method.

| Model | Number of Parameters |
|---|---|
| Transform and Tell | 200M |
| MT | **93M** |
| MT (w/o PE) | 89M |
| MT (w/o Pointer) | 91M |
| MT (w/o EG) | 91M |

Table 3: We compare the number of training parameters of our model variants and the model from Transform and Tell (Tran et al., 2020). Note that our proposed Multi-Modal Transformer is much more lightweight. MT means our final Multi-Modal Transformer.

adds named entities as another text source to help predict named entities more accurately. VS (Visual Selective Layer) tries to strengthen the connection between the image and text. PE (Position Embedding) provides the trainable positional embeddings added to the word embeddings. Pointer stands for the updated multi-head pointer-generator module. To overcome the limitation of a fixed-size vocabulary, we examine TC, the Tag-Cleaning operation handling *OOV* problem.

### 3.3 Results and Discussion

Table 1 and Table 2 summarize our quantitative results on the GoodNews and VisualNews datasets respectively. On GoodNews, our Multi-Modal Transformer outperforms the state-of-the-art methods on 5 out of 6 metrics and reaches a comparably good performance in ROUGE score. On our Visual News dataset, our model outperforms baseline methods by a large margin, from 13.5 to 50.5 in CIDEr score. In addition, as revealed by Table 3, our final model outperforms *Transform and Tell(transformer)* with much fewer parameters so that our training time is only a half of the model used in Tran et al. (2020). This demonstrates that our proposed model is able to generate better captions in a more efficient way.

Our Entity-Guide (EG) brings improvement in both datasets, demonstrating the named entity set indeed contains key information guiding the generation of news captions. In addition, our Position Embedding (PE) also shows its effectiveness by providing additional positional information to the token embedding. Pointer-generator mechanism builds a stronger connection between the final distribution of the predicted tokens and the Multi-Modal AoA Layer. More importantly, our Visual Selective Layer (VS) improves the caption generation results by providing extra visual context to text features.

Furthermore, our Tag-Cleaning (TC) method is able to effectively retrieve uncommon named entities and thus improves the CIDEr score by $1.7\%$ and $1.3\%$ respectively on the GoodNews and VisualNews datasets. We present qualitative results of different models on both datasets in Figure 3. Our model shows the ability to generate high quality captions with more accurate named entities.

We also observe that our models and *Transform and Tell* methods achieve best performances are directly trained on raw captions rather than following a two-stage template-based manner. Although template-based methods normally handle a much smaller vocabulary, these methods also suffer from losing rich contextual information brought by uncommon named entities.

## 4 Conclusion and Future Work

In this paper, we study the task of news image captioning. We propose Multi-Modal Transformer, an entity-aware captioning method leveraging both visual and textual information. We validate the effectiveness of our method on VisualNews and another large-scale benchmark dataset through extensive experiments. Multi-Modal Transformer outperforms state-of-the-art methods across multiple metrics with fewer parameters.

## References


Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. In *ICLR (Poster)*.

Federico Barrios, Federico López, Luis Argerich, and Rosa Wachenchauzer. 2016. Variations of the similarity function of textrank for automated summarization.

Ali Furkan Biten, Lluis Gomez, Marçal Rusinol, and Dimosthenis Karatzas. 2019. Good news, everyone! context driven entity-aware captioning for news images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12466–12475.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380.

Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. 2015. From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1473–1482.

Kavita Ganesan. 2018. Rouge 2.0: Updated and improved measures for evaluation of summarization tasks. *arXiv preprint arXiv:1803.01937*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7(1).

Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. 2019. Attention on attention for image captioning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4634–4643.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR (Poster)*.

Chenliang Li, Weiran Xu, Si Li, and Sheng Gao. 2018. Guiding generation for abstractive text summarization based on key information guide network. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 55–60.

Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. 2020. Visualnews : Benchmark and challenges in entity-aware image captioning.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Di Lu, Spencer Whitehead, Lifu Huang, Heng Ji, and Shih-Fu Chang. 2018a. Entity-aware image caption generation. In *EMNLP*, pages 4013–4023. Association for Computational Linguistics.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2018b. Neural baby talk. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7219–7228.

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

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *ACL (1)*, pages 1073–1083. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Alasdair Tran, Alexander Mathews, and Lexing Xie. 2020. Transform and tell: Entity-aware news image captioning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057.