Image Caption Generation for News Articles

Zhishen Yang
Tokyo Institute
of Technology
zhishen.yang
@nlp.c.titech.ac.jp

Naoaki Okazaki
Tokyo Institute
of Technology
okazaki
@c.titech.ac.jp

Abstract

In this paper, we address the task of news-image captioning, which generates a description of an image given the image and its article body as input. This task is more challenging than the conventional image captioning, because it requires a joint understanding of image and text. We present a Transformer model that integrates text and image modalities and attends to textual features from visual features in generating a caption. Experiments based on automatic evaluation metrics and human evaluation show that an article text provides primary information to reproduce news-image captions written by journalists. The results also demonstrate that the proposed model outperforms the state-of-the-art model. In addition, we also confirm that visual features contribute to improving the quality of news-image captions.

1 Introduction

Image captioning, i.e., automatic generation of a natural language description from an image, has received much attention from both fields of Computer Vision (CV) and Natural Language Processing (NLP) (Vinyals et al., 2015; Xu et al., 2015; Karpathy and Fei-Fei, 2015). Image captioning is important not only in practical application (e.g., automatic indexing of images) but also as a challenge of image understanding (e.g., recognition of objects, object-object relationships, and scenes).

In this paper, we address a more advanced task, *news-image captioning*: this task generates a description of an image, given the image and its article body as input. The news-image captioning task is different from the conventional image captioning, which receives only an image as input. In other words, news-image captioning requires a mutual understanding of image and text.

Early work proposed a two-stage approach for news-image captioning (Feng and Lapata, 2013; Tariq and Foroosh, 2017). The first stage annotates keywords to a given image and text, and the second stage realizes a description based on the extracted keywords. Recent work presented an end-to-end approach that integrates image and text features in deep neural networks (Ramisa et al., 2018; Batra et al., 2018; Biten et al., 2019). However, the previous studies did not focus on the usefulness of text in the news-image captioning task, extending the conventional models for image captioning to incorporate text features.

Figure 1 shows an example of a news-image caption. It may be difficult to recognize which is the central object in the image, for example, people, violins, and stick (bow). Also, the caption includes much information (e.g., *Juilliard Orchestra*, *Vladimir Jurowsky*, and *Alice Tully Hall*) that may be hard to tell only from the image. This kind of example is rather common in news articles, where a text is the major medium of information, and an image and its caption provide additional explanations that support the text. However, no previous work explored a method that integrated an article text seamlessly with an image in the task of news-image captioning.

In this paper, we present a method for news-image captioning based on Transformer (Vaswani et al., 2017), a successful architecture for various NLP tasks, including machine translation, abstractive summarization, contextualized word embeddings. We propose a Transformer model that integrates text and

Proceedings of the 28th International Conference on Computational Linguistics, pages 1941–1951
Barcelona, Spain (Online), December 8-13, 2020

The Silent-Era Shostakovich

When prominent conductors visit New York for important engagements with local institutions, their schedules are usually very full. So it says good things about the priorities of the Russian maestro Vladimir Jurowski that while in New York to conduct a six-performance run of Strauss's "Die Frau Ohne Schatten" at the Metropolitan Opera, he made time to work with the Juilliard Orchestra on an adventurous Shostakovich program. That concert took place on Monday at Alice Tully Hall, the night before Mr. Jurowski's final "Frau" at the Met. … …

Juilliard Orchestra Vladimir Jurowski conducting at Alice Tully Hall on Monday evening.

Figure 1: An image and its caption in a news article.

image modalities and attends to textual features from visual features in generating a caption. The experimental results show that an article text provides primary information to reproduce news-image captions written by journalists. The results also demonstrate that the proposed model outperforms the state-of-the-art model (Biten et al., 2019). We report the results of human evaluation and discuss challenges in news-image captioning. The dataset and code used in this work are publicly available[1].

2 Multimodal Transformer Model

Given an image i and a text as a sequence of n tokens $(x_1, x_2, \ldots x_n)$, the task of news-image captioning generates a caption as a sequence of m tokens $(y_1, y_2, \ldots y_m)$. As the base architecture for realizing the task, we use the Transformer model (Vaswani et al., 2017), which has been a popular architecture for headline generation for news articles (Takase and Okazaki, 2019; Duan et al., 2019; Dong et al., 2019). Because headline generation is a sequence-to-sequence task from $(x_1, \ldots x_n)$ to $(y_1, \ldots y_m)$, we consider incorporating features from an image i to the architecture.

Figure 2 illustrates the proposed model. The model consists of *image encoder*, *image-article encoder*, and *decoder*. The image encoder converts an image i into a feature vector $p_i \in R^d$,

$$p_i = \mathrm{CNN}(i). \tag{1}$$

Here, d presents the number of dimensions of hidden feature vectors, and $\mathrm{CNN}(.)$ is a Convolutional Neural Network to convert an image into a feature vector. In this study, we compare two CNN models trained on the datasets of object recognition and scene recognition as $\mathrm{CNN}(.)$.

The image-article encoder computes joint representations for an input image and text. The input to the first layer of the image-article encoder is a $d \times n$ matrix, whose column vectors present an addition of token embedding and positional encoding. The image-article encoder stacks N layers of two sub-modules: *encoder module* and *image-attending module*. The encoder module is identical to the encoder component of the original Transformer model: it is a mapping from $R^{d \times n}$ to $R^{d \times n}$ by using a multi-head self-attention followed by a feed-forward layer.

The image-attending module is similar to the decoder component of the original Transformer model: it is a mapping from $R^{d \times n}$ to $R^{d \times n}$ by using a multi-head target-source attention followed by a feed-forward layer. Here, we provide the image vector p_i as keys and values for the scaled dot-product attention so that this module can build intermediate representations by attending to the input image. The image-attending module adds the outputs from the encoder module and the target-source attention via the residual connection. Thus, we expect that target-source attention can focus on implicit contextual correlations between visual and textual features, modifying input representations based on the input image.

The decoder has M layers, each of which is exactly the same as the original Transformer decoder. A decoder layer consists of multi-head masked self-attention over output tokens, followed by the target-source attention that receives the encoder output as keys and values of the scaled dot-product attention. In

[1] https://github.com/nlp-titech/news_image_captioning_for_news_articles

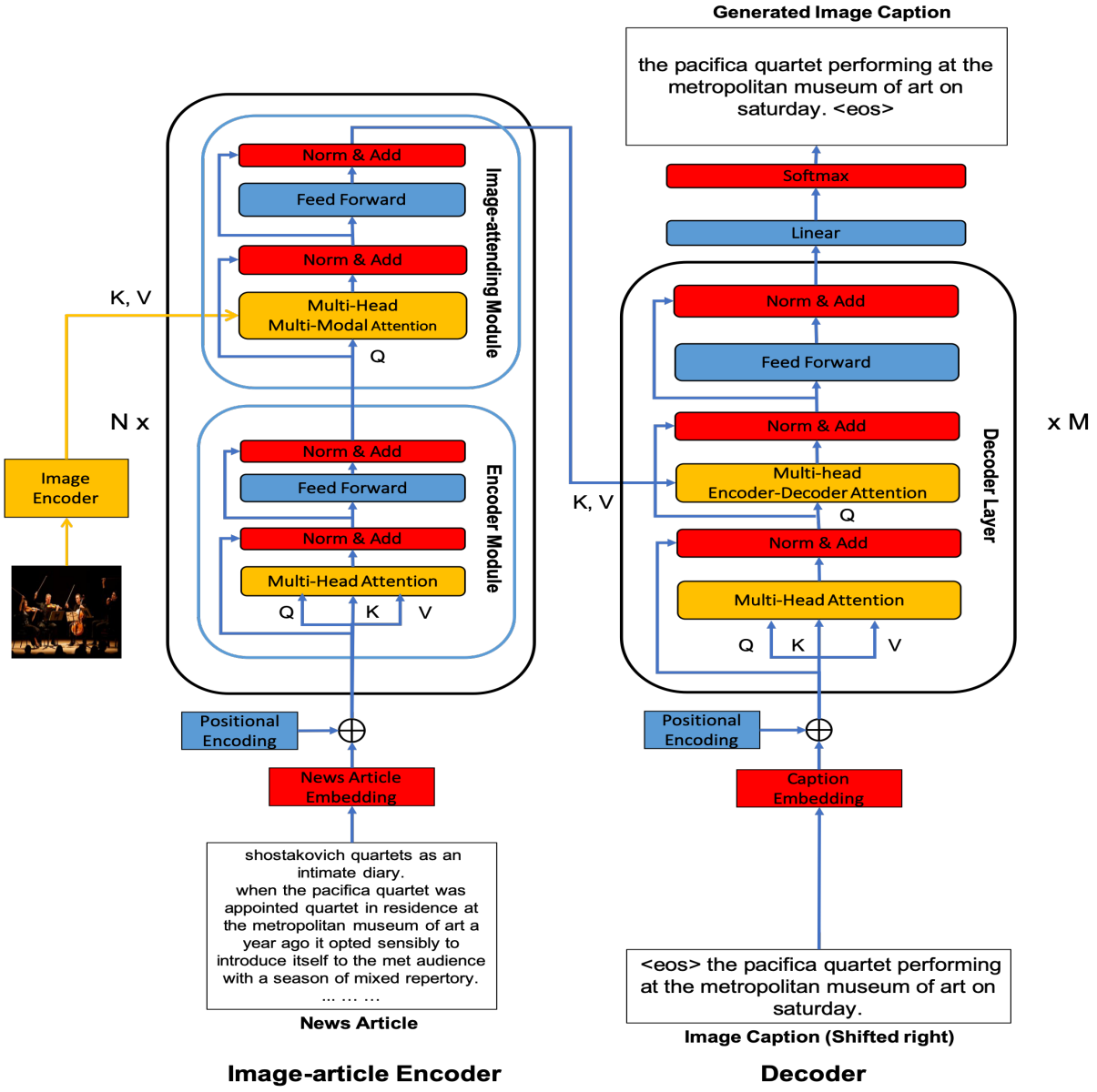

Figure 2: Overall architecture of the proposed multimodal Transformer model.

this way, the overall architecture can generate a caption by selectively drawing attention to the important information of the image and text from the news article.

3 Experiments

The questions we want to explore in this section are two-folds:

- Can the multimodal Transformer model presented in Section 2 successfully integrate visual and textual features?
- Are image captions mostly influenced by textual or visual features in news articles?

To answer these questions, we select the news-image captioning dataset: GoodNews, and present fixes for its issues in Section 3.1. After explaining the experimental settings (in Sections 3.2 to 3.5), we report evaluation results by using the automatic evaluation (in Section 3.6) followed by the human evaluation (in Section 3.7).

3.1 Dataset for News-image Captioning

Biten et al. (2019) released the GoodNews dataset, which is the largest dataset for news-image captions so far. We considered using this dataset but found some issues. The biggest issue was that many instances in the dataset contain incomplete text. More concretely, many instances lack several leading sentences, which usually convey essential information on news articles. The issue is critical to our work, which aims to explore the use of vision and textual information in news-image captioning. Also, Biten et al. (2019) split the dataset into training and test sets at the image level; for this reason, the same news text may appear in both the training and test sets.

To solve these issues, we crawled complete text for each news article and split the dataset randomly at the article level. Our dataset for news-image captions includes 269K articles and 489K images in total. An article contains 1.8 images on average, and 59% of articles have only one image. The average lengths of body text, headlines, and image captions of the articles are 963.29, 8.57, and 17.55 words, respectively. We split the dataset into 245K articles for the training set, 10K for the validation set, and 13K for the test set.

3.2 Data Preprocessing

Some articles are too long to store them in the GPU memory, and a headline and leading paragraphs of an article provide the most useful information, we truncated each news article in the dataset to keep the first 416 words at most (including the headline). In addition to truncating articles, we also removed punctuation and non-ASCII characters from text. We left a dot '.' as a delimiter of sentences. Applying the algorithm of Byte-Pair-Encoding (BPE) implemented in SentencePiece[2] (Kudo and Richardson, 2018) to the preprocessed news articles of the training set, we built a vocabulary of 32,000 subwords.

3.3 Baselines and Model Variants

Biten et al. (2019) presented a model that achieved the state-of-art performance on the GoodNews dataset. Using the publicly-available implementation[3], we trained their models on our dataset with six different settings: Avg + AttIns, Avg + CtxIns, TBB + AttIns, TBB + CtxIns, Wavg + AttIns, and Wavg + CtxIns. These settings are described in detail in Biten et al. (2019). For the sake of clarity, we briefly explain some keywords: Avg: article embeddings computed by averages of GloVe word vectors for each sentence; Wavg: article embeddings computed by weighted averages of GloVe word vectors for each sentence; TBB: article embedding computed by the tough-to-beat baseline (Arora et al., 2019); AttIns: named entity insertion guided by the attention mechanism over the article; and ctxIns: context-based named entity insertion.

To compare the importance of visual and textual features for news-image captioning, we also prepared variants of the proposed model.

- **Transformer (Text)**: We trained the original Transformer model without image features; this is a baseline setting to see how well a Transformer model can generate captions without looking at images.
- **Transformer (ImageNet)**: We trained a decoder-only Transformer model where keys and values of the multi-head target-source attention are image features computed by the object recognition model (trained on ImageNet[4]); this is a baseline setting to generate captions only from images.
- **Transformer (Places 365)**: This model is identical to Transformer (ImageNet), but image features are computed by the scene recognition model (trained on Places 365[5]) (Zhou et al., 2017).
- **Multimodal Transformer (ImageNet)**: The proposed model with image features obtained using the object recognition model (trained on ImageNet).

[2] https://github.com/google/sentencepiece
[3] https://github.com/furkanbiten/GoodNews/
[4] http://www.image-net.org/
[5] http://places2.csail.mit.edu/

- **Multimodal Transformer (Places 365)**: The proposed model with image features obtained using the scene recognition model (trained on Places 365).
- **Multimodal Transformer (ImageNet & Places 365)**: The proposed model with the average of two image features obtained using the object recognition model (trained on ImageNet) and the scene recognition model (trained on Places 365).

In this experiment, we focused on visual features trained for two different tasks: object recognition (ImageNet) and scene recognition (Places 365). We used ResNet-18 (He et al., 2016) pre-trained on ImageNet and Places365 to obtain visual features from an image.

In addition to these variants, we also included two simple baselines: **Lead** and **Headline** assume lead sentences and headlines as image captions. In other words, these baselines reveal the similarity between image captions and lead/headline sentences.

3.4 Implementation and Training

We implemented all Transformer models using PyTorch (Paszke et al., 2019) based on Fairseq (Ott et al., 2019).

Hyper-parameters All Transformer-based models in our experiments used the same hyper-parameters: the number of dimensions of hidden vectors $d = 512$; the number of attention heads $H = 8$; the number of encoder blocks $N = 3$; the number of decoder blocks $M = 6$. Table 1 shows the number of parameters trained for each Transformer-based model.

Model	# Parameters
Transformer (Text)	93M
Transformer (Image)	57M
Multimodal Transformer	93M

Table 1: Number of parameters trained in the Transformer-based models.

Training We used Adam (Kingma and Ba, 2015) with $\beta_1 = 0.9, \beta_2 = 0.98, \epsilon = 10^{-8}$ for parameter estimation. We linearly increased the learning rate from initial rate 10^{-7} until 0.0005 in the 4,000 warm-up steps, and then decreased it proportionally to the inverse of the square root of the step number with the minimum learning rate of 10^{-9}. The objective function is the cross-entropy loss with label smoothing of 0.1 (Szegedy et al., 2016). In each layer of the model, we applied dropout (Srivastava et al., 2014) with the rate of 0.3 after the layer normalization and before the residual connection. In both the encoder and decoder, we also applied dropout with the rate of 0.3 after taking the sums of token embeddings and positional encoding. We also applied dropout with the rate of 0.1 to the attention weights.

Training all models for 50 epochs, we stored model parameters that yielded the minimum loss. It took about 1.2 days to train a Multimodal Transformer model on four NVIDIA Tesla V100 for NVLink (16GiB HBM2).

3.5 Evaluation Metrics

We used five automatic evaluation metrics: BLEU (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2014), ROUGE (Lin and Och, 2004), CIDEr (Vedantam et al., 2015), and SPICE (Anderson et al., 2016). Because news-image captions contain a fair amount of named entities, we considered CIDEr as the most appropriate and informative metric among these five metrics. We used the MS-COCO caption evaluation tool[6] after lower-casing and removing the punctuation in captions.

Both news articles and news-image captions contain a certain amount of named entities, and these entities often carry relevant contextual knowledge. Therefore we introduced $\mathrm{Coverage_{NE}}$ to measure the coverage of named entities in generated captions. The coverage for a pair of generated and ground-truth captions is defined as recall of named entities in the captions:

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

$$\text{Coverage}_{\text{NE}} = \frac{|E_{\text{generated}} \cap E_{\text{gold}}|}{|E_{\text{gold}}|}.$$

Here, $E_{\text{generated}}$ and E_{gold} present sets of named entities in a generated and ground-truth captions, respectively. We used SpaCy[7] to identify named entities in the ground truth captions, and applied regular expressions to find exact matches in generated captions.

3.6 Results (Automatic Evaluation)

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
Multimodal Transformer (ImageNet)	18.62	10.81	6.70	4.46	8.38	**20.56**	**44.16**	**10.19**
Multimodal Transformer (Places 365)	**18.78**	**10.90**	**6.76**	**4.52**	8.42	20.41	44.04	10.15
Multimodal Transformer (ImageNet & Places 365)	18.10	10.53	6.56	4.38	8.27	20.29	43.01	10.07
Transformer (Text)	15.10	8.72	5.37	3.55	7.27	17.83	37.02	8.96
Transformer (Image) (ImageNet)	12.13	4.66	2.20	1.23	3.49	11.70	9.37	2.44
Transformer (Image) (Places 365)	11.72	4.43	2.06	1.16	3.40	11.45	8.70	2.29
Lead	14.54	7.57	4.47	2.92	**8.62**	12.91	12.79	6.96
Headline	6.86	3.16	1.52	0.81	4.69	10.31	16.80	5.46
Biten et al. (2019) (Avg + AttIns)	6.73	2.53	1.17	0.63	3.56	11.81	12.38	3.27
Biten et al. (2019) (Avg + CtxIns)	6.95	2.52	1.14	0.62	3.58	11.72	11.40	2.94
Biten et al. (2019) (TBB + AttIns)	5.22	1.74	0.72	0.36	2.95	10.80	8.43	2.62
Biten et al. (2019) (TBB + CtxIns)	5.85	2.08	0.92	0.48	3.43	11.52	10.85	2.92
Biten et al. (2019) (Wavg + AttIns)	6.31	2.34	1.05	0.56	3.61	11.72	11.92	3.18
Biten et al. (2019) (Wavg + CtxIns)	6.52	2.33	1.02	0.52	3.67	11.72	11.36	2.92

Table 2: Performance of news-image caption generation measured by the automatic evaluation metrics.

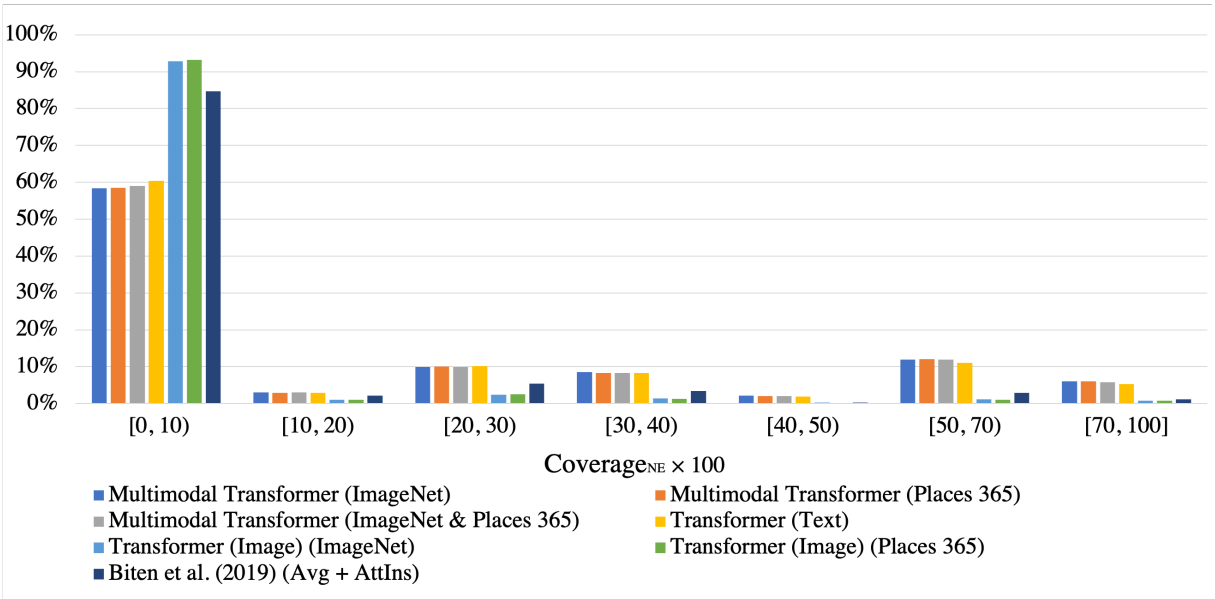

Figure 3: Distributions of $\text{Coverage}_{\text{NE}}$ scores for seven representative models.

Table 2 reports results of automatic evaluation for baseline models, variants of Transformer-based models, and the state-of-the-art model (Biten et al., 2019) on our dataset. Multimodal Transformer (ImageNet) yields the best CIDEr score. All Transformer-based models using textual features substantially outperform the state-of-the-art model (Biten et al., 2019).

The Headline baseline was roughly comparable to the state-of-the-art model (Biten et al., 2019). Just as the task of headline generation, the Lead method exhibited the performance of a strong baseline for news-image captioning. This is probably because images and captions in news articles may have the role

[7] https://spacy.io/

of indicative summaries, i.e., receiving attentions from readers and inviting readers to the articles. In other words, journalists may write an image caption as a summary of an article, expecting the image and caption to be an alternative starting point to the article.

Transformer (Text) was also a strong baseline for news-image caption generation, which outperformed Transformer (Image) models even without looking at actual images. This result indicates that news-image captions describe images and include much information mentioned in the text. The huge difference in CIDEr scores of Transformer (Text) and Transformer (Image) implies that the vision-only models suffered when it comes to named entities.

In the meantime, we could observe an improved performance in Multimodal Transformer models compared to the Transformer (Text). This result demonstrates that visual features are also useful in generating captions. Multimodal Transformer (Places 365) performed the best in BLEU, and Multimodal Transformer (ImageNet) achieved the best in ROUGE-L, CIDEr and SPICE. Mixed two types of visual features yielded a slight performance deterioration (Multimodal Transformer (ImageNet & Places 365)). Notably, performance differences are rather small among Multimodal Transformer models. In Appendix A, we will present a case study of captions generated from Multimodal Transformer models and Transformer (Text) model.

Figure 3 shows the distribution of $\mathrm{Coverage_{NE}}$ scores of each model: the x-axis of the graph presents a range of $(100 \times \mathrm{Coverage_{NE}})$ scores, and the y-axis presents the proportion of test instances for each score range. The graph indicates that the Transformer models given the article text were able to include much more correct named entities than other models. With no access to article text, it is quite natural that the Transformer (Image) models could not include correct entities. The previous state-of-the-art model (Biten et al., 2019) is somewhere between the Transformer (Image) models and the Multimodal Transformer models.

3.7 Results (Human Evaluation)

Model	Grammaticality	Faithfulness	Descriptiveness	Overlap
Multimodal Transformer (Places 365)	4.51	2.29	**1.77**	**1.60**
Transformer (Text)	**4.58**	**2.37**	1.69	1.47
Biten et al. (2019) (Avg + AttIns)	2.62	2.08	1.41	1.23

Table 3: Average scores of human evaluation for three representative models.

Because we are unsure of the appropriateness of the automatic evaluation in this task, we also conducted a human evaluation. We asked three native English speakers to evaluate generated captions from three models: Multimodal Transformer (Places 365), Transformer (Text), and Biten et al. (2019) (Att + AttIns).

We randomly chose an evaluation set with 136 images. Each instance in this evaluation set included the image, news article, ground truth caption, and generated captions from the three models. Three generated captions were presented to human subjects in random order so that they were not able to guess the quality of a caption from the appearance order. We designed four criteria for rating generated captions: grammaticality, faithfulness, descriptiveness, and overlap.

- **Grammaticality**: The caption: has no error (5), has one error (4), has two errors (3), is understandable (2), or is incomprehensible (1).

- **Faithfulness**: The caption has: no unfaithful fact (5), one unfaithful fact (4), a few (but less than 50%) unfaithful facts (3), more than 50% unfaithful facts (2), or the content that is totally unrelated to the article and image (1).

- **Descriptiveness**: The caption explains the image (3); the caption does not explain the image, but describes something related to the image (2), or the caption is totally unrelated to the image (1).

- **Overlap**: The overlap between generated and ground-truth captions, 100% overlap (5), 80% overlap (4), 50% overlap (3), 20% overlap (2), or no overlap (1).

Note that human evaluation is not easy. There is no guarantee that human evaluators are familiar with objects and scenes (e.g., people, building, location) appearing in the news. Although a news article (text) provides a hint for interpreting an image, they may find it hard to search for evidence of the image on the Internet. Therefore we always presented the ground-truth captions to human subjects to help them understand images.

Table 3 shows the average score assigned to each model and criterion. The two Transformer models outperformed the previous state-of-the-art model on the four criteria. In particular, Biten et al. (2019) (Avg + AttIns) suffered from the low score of grammaticality. However, there was no clear winner between the two Transformer models in the human evaluation. We could observe that the ranking of the overlap criterion was consistent with those of the automatic evaluation metrics used in Section 3.6. This is reasonable because the overlap criterion is a manual version of the automatic metrics. Multimodal Transformer (Places 365) has an advantage in the descriptiveness criterion, but the score difference from Transformer (Text) is small. The overall low scores also imply that news-image captioning remains challenging, especially for incorporating visual information from images.

4 Related Work

Image Captioning Recent advances in image captioning followed the encoder-decoder architecture, where the encoder extracts visual features from images, and the decoder generates the caption (Xu et al., 2015; Vinyals et al., 2015; You et al., 2016; Li et al., 2017). Notably, the attention mechanism further improved the quality of generated captions by finding implicit alignments between visual and textual features. These models achieved good performance in generating captions at a descriptive level with a consistent writing style. Different from generic image captions, news-image captions are often influenced by contextual information from the news article.

News-image captioning News-image captioning has been paid little attention in the literature. Early work (Feng and Lapata, 2013; Tariq and Foroosh, 2017) presented models with two stages: the first stage is an annotation model that suggests keywords for an image; and the second stage realizes sentences based on the extraction result of the first stage. Recent studies departed from these approaches by utilizing the deep neural network to find the implicit image-text correlations (Batra et al., 2018; Ramisa et al., 2018).

Biten et al. (2019) is the most recent method yielding the state-of-art performance on the Goodnews dataset. The method consists of two stages: in the first stage, long short-term memory (LSTM) combines textual features at sentence-level and visual features at object-level to generate a template sentence. The second stage then inserts named entities into the placeholder of the template sentence to realize captions with named entities. In contrast, our method required no template sentence and included named entities directly from article text based on the Transformer architecture.

5 Conclusion

In this paper, we presented a method for news-image captioning based on the Transformer model, which integrates text and image modalities and attends to textual features from visual features in generating a caption. The experimental results demonstrated that the proposed model could integrate visual and textual information in generating captions. Meanwhile, we also found that news-image captioning as a context-oriented image captioning task, text from news article fundamentally contributes our model in generating context-coherent captions, with the engagement of image further improves qualities.

In the future, we would like to explore a better approach for recognizing visual contents from images to improve the quality of generated captions. Besides, we are also interested in whether a model of news-image captioning is transferrable to other multimodal tasks such as multimodal translation and visual question answering.

6 Acknowledgement

We appreciate the anonymous reviewers for their constructive comments. The research results have been achieved by "Research and Development of Deep Learning Technology for Advanced Multilingual Speech Translation", the Commissioned Research of National Institute of Information and Communications Technology (NICT), Japan.

References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision (ECCV)*, pages 382–398.

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2019. A simple but tough-to-beat baseline for sentence embeddings. In *International Conference on Learning Representations (ICLR)*.

Vishwash Batra, Yulan He, and George Vogiatzis. 2018. Neural caption generation for news images. In *International Conference on Language Resources and Evaluation (LREC)*.

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

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Ninth Workshop on Statistical Machine Translation*, pages 376–380.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems*, pages 13063–13075.

Xiangyu Duan, Hongfei Yu, Mingming Yin, Min Zhang, Weihua Luo, and Yue Zhang. 2019. Contrastive attention mechanism for abstractive sentence summarization. In *2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3044–3053.

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

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

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3128–3137.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *International Conference on Learning Representations (ICLR)*.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *2018 Conference on Empirical Methods in Natural Language Processing (EMNLP): System Demonstrations*, pages 66–71.

Linghui Li, Sheng Tang, Lixi Deng, Yongdong Zhang, and Qi Tian. 2017. Image caption with global-local attention. In *Thirty-First AAAI Conference on Artificial Intelligence*, pages 4133–4139.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *42nd Annual Meeting on Association for Computational Linguistics (ACL)*, pages 605–612.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *40th Annual Meeting on Association for Computational Linguistics (ACL)*, pages 311–318.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035.

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

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *IEEE conference on computer vision and pattern recognition (CVPR)*, pages 2818–2826.

Sho Takase and Naoaki Okazaki. 2019. Positional encoding to control output sequence length. In *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 3999–4004.

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

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 *IEEE conference on computer vision and pattern recognition (CVPR)*, pages 4566–4575.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3156–3164.

Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on International Conference on Machine Learning (ICML)*, pages 2048–2057.

Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14651–4659, June.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464.

Appendix A Case Study

Figure 4 illustrates examples of captions generated by the models. For the example (a), three Multimodal Transformer models could describe the visual scene 'news conference' in the generated captions, while Transformer(Text) could only predict the name of the person.

Example (b) demonstrates a typical difficult case in the news-image caption task, journalists wrote the caption with loose correlation with visual content in the image but conveyed a message cooperating with the article storyline. Therefore, the Multimodal Transformer models generated captions that obey the image while failed to deliver real messages from journalists.

(a)

Ground Truth
yoshihiko noda japan s prime minister speaking at a news conference in tokyo on monday.

Multimodal Transformer (ImageNet & Places 365)
prime minister yoshihiko noda of japan at a news conference in tokyo on monday.

Multimodal Transformer (ImageNet)
prime minister yoshihiko noda at a news conference in tokyo on monday.

Multimodal Transformer (Places 365)
prime minister yoshihiko noda of japan at a news conference in tokyo on monday.

Transformer (Text)
prime minister yoshihiko noda of japan left and prime minister yoshihiko noda of japan in tokyo on monday.

(b)

Ground Truth
a little campari before dancing.

Multimodal Transformer (ImageNet & Places 365)
ms weinra weinra at her apartment in the west village.

Multimodal Transformer (ImageNet)
ms weinrauch in her manhattan apartment.

Multimodal Transformer (Places 365)
ms weinuch in her apartment in manhattan.

Transformer (Text)
helena weinrauch with her daughter arlene weinberg at the museum of jewish heritage in manhattan.

Figure 4: Captions generated by the Transformer models. In (a), Transformer (Text) made the correct prediction for the person in the image (the prime minister of Japan). Multimodal Transformer models injected the correct visual information (news conference) into the caption. In (b), all four models failed to generate the correct caption. Transformer (Text) predicted the correct name but wrong contextual information. Multimodal Transformer models generated captions with a different focus.