Cross-Modal Commentator: Automatic Machine Commenting Based on Cross-Modal Information

Pengcheng Yang[1,2]*, Zhihan Zhang[2]*, Fuli Luo[2], Lei Li[2], Chengyang Huang[3], Xu Sun[1,2]
[1]Deep Learning Lab, Beijing Institute of Big Data Research, Peking University
[2]MOE Key Lab of Computational Linguistics, School of EECS, Peking University
[3]Beijing University of Posts and Telecommunications
{yang_pc, zhangzhihan, luofuli, xusun}@pku.edu.cn
tobiaslee@foxmail.com, cyhuang@bupt.edu.cn

Abstract

Automatic commenting of online articles can provide additional opinions and facts to the reader, which improves user experience and engagement on social media platforms. Previous work focuses on automatic commenting based solely on textual content. However, in real-scenarios, online articles usually contain multiple modal contents. For instance, graphic news contains plenty of images in addition to text. Contents other than text are also vital because they are not only more attractive to the reader but also may provide critical information. To remedy this, we propose a new task: cross-model automatic commenting (CMAC), which aims to make comments by integrating multiple modal contents. We construct a large-scale dataset for this task and explore several representative methods. Going a step further, an effective co-attention model is presented to capture the dependency between textual and visual information. Evaluation results show that our proposed model can achieve better performance than competitive baselines. [1]

1 Introduction

Comments of online articles can provide rich supplementary information, which reduces the difficulty of understanding the article and enhances interactions between users. Therefore, achieving automatic commenting is necessary since it can contribute to improving user experience and increasing the activeness of social media platforms.

Due to the importance described above, some work (Qin et al., 2018; Lin et al., 2018; Ma et al., 2018) has explored this task. However, these efforts are all focus on automatic commenting based solely on textual content. In real-scenarios, online articles on social media usually contain multiple modal contents. Take graphic news as an example, it contains plenty of images in addition to text. Other contents except text are also vital to improving automatic commenting. These contents may contain some information that is critical for generating informative comments. In addition, compared to plain text, these contents of other modalities are more attractive to the reader, making it easily become the focus of comments.

Figure 1: An example in the constructed dataset. Red words indicate the content that is not included in the text but depicted in the images.

Toward filling this gap, we propose the task of cross-model automatic commenting (CMAC), which aims to generate comments by integrating information of multiple modalities. We construct a large-scale cross-model comments dataset, which consists of 24,134 graphic news. Each instance is composed of several news photos, news title, news body, and corresponding high-quality comments. Figure 1 visually shows a sample in the dataset.

*Equal Contribution.
[1]The dataset and code are available at https://github.com/lancopku/CMAC

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 2680–2686
Florence, Italy, July 28 - August 2, 2019.

Since the comments depend on the contents of multiple modalities, how to integrate these multimodal information becomes the focus. In fact, there exist intrinsic interactions between these input multimodal information. Various modalities can benefit from each other to obtain better representations. For instance, in the graphic news, images can help to highlight the important words in the text, while text also contributes to focusing on key regions of images. Therefore, we present a co-attention model so that the information of multiple modalities can mutually boost for better representations. Experiments show that our co-attention model can substantially outperform various baselines from different aspects.

The main contributions of this work are summarized as follows:

- We propose the task of cross-modal automatic commenting (CMAC) and construct a large-scale dataset.
- We present a novel co-attention model, which aims at capturing intrinsic interactions between multiple modal contents.
- The experiments show that our approach can achieve better performance than competitive baselines. With multiple modal information and co-attention, the generated comments are more diverse and informative.

2 Cross-Modal Comments Dataset

We introduce our constructed cross-modal comments dataset from the following aspects.

Data collecting We collect data from the photo channels of a popular Chinese news website called Netease News[2]. The crawled news cover various categories including entertainment, sports, and more. We tokenize all texts into words, using a python package Jieba[3]. To guarantee the quality of the comments, we reserve comments with the length between 5 to 30 words and remove useless symbols and dirty words. Besides, we filter out short articles with less than 10 words or 3 images in its content, while unpopular articles with less than 10 pieces of comments are also removed. Finally, we acquire a dataset with 24,134 pieces of news. Each instance contains the news title and its body, several images and a list of high-quality comments. On average, each news in the dataset contains about 39 human-written comments.

Statistic	Train	Dev	Test	Total
# News	19,162	3,521	1,451	24,134
# Comments	746,423	131,175	53,058	930,656
Avg. Images	5.81	5.78	5.81	5.80
Avg. Body	54.75	54.72	55.07	54.77
Avg. Comment	12.19	12.21	12.18	12.19

Table 1: Statistics of the dataset. **# News** and **# Comments** denote the total number of news and comments, respectively. **Avg. Images** is the average number of images per news. **Avg. Body** is the average number of words per body, and similar to **Avg. Comment**.

Evaluation	Flue.	Rele.	Info.	Overall
Score	9.2	6.7	6.4	7.6
Pearson	0.74	0.76	0.66	0.68

Table 2: Quality evaluation results of the testing set. **Flue.**, **Rele.** and **Info.** denotes fluency, relevance, and informativeness, respectively.

Data Statistics The dataset is split according to the corresponding news. The comments from the same news will appear solely in the training or testing set to avoid overfitting. In more detail, we split the data into 19,162, 3,521 and 1,451 news in the training, development, and testing sets, respectively. The corresponding number of comments is 746,423, 131,175 and 53,058, respectively. The statistics of the final dataset are presented in Table 1 and Figure 2 shows the distribution of the lengths for comments in both word-level and character-level.

Data Analysis High-quality testing set is necessary for faithful automatic evaluation. Therefore, we randomly selected 200 samples from the testing set for quality evaluation. Three annotators with linguistic background are required to score comments and readers can refer to Section 4.3 for the evaluation details. Table 2 shows the evaluation results. The average score for overall quality is 7.6, showing that the testing set is satisfactory.

3 Proposed Model

Given the texts[4] $\boldsymbol{x}$ and images $\boldsymbol{v}$ of an online article, the CMAC task aims to generate a reasonable and fluent comment $\boldsymbol{y}$. Figure 3 presents the overview of our proposed model, which is elaborated on in detail as follows.

[2] http://news.163.com/photo

[3] https://github.com/fxsjy/jieba

[4] We concatenate the title and body into a single sequence.

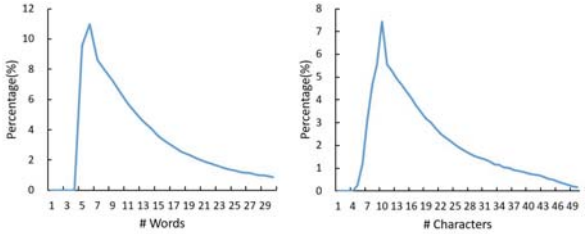

Figure 2: The distribution of lengths for comments in terms of both word-level and character-level.

3.1 Textual Encoder and Visual Encoder

The textual encoder aims to obtain representations of textual content $\boldsymbol{x}$. We implement it as a GRU model (Cho et al., 2014), which computes the hidden representation of each word as follows:

$$h_i^x = \mathrm{GRU}\big(h_{i-1}^x, e(x_i)\big) \tag{1}$$

where $e(x_i)$ refers to the embedding of the word x_i. Finally, the textual representation matrix is denoted as $\mathbf{H}^x = \{h_1^x, \cdots, h_{|\boldsymbol{x}|}^x\} \in \mathbb{R}^{|\boldsymbol{x}| \times d_1}$, where $|\boldsymbol{x}|$ is the total number of textual representations and d_1 is the dimension of h_i^x.

We apply ResNet (He et al., 2016a) as visual encoder to obtain the visual representation[5] h_i^v of the i-th image v_i. The final visual representation matrix is denoted as $\mathbf{H}^v = \{h_1^v, \cdots, h_{|\boldsymbol{v}|}^v\} \in \mathbb{R}^{|\boldsymbol{v}| \times d_2}$, where $|\boldsymbol{v}|$ is the number of visual representations and d_2 is the dimension of h_i^v.

3.2 Co-Attention Mechanism

We use co-attention mechanism to capture the intrinsic interaction between visual content and textual content. The two modal information are connected by calculating the similarity matrix $\mathbf{S} \in \mathbb{R}^{|\boldsymbol{v}| \times |\boldsymbol{x}|}$ between $\mathbf{H}^v$ and $\mathbf{H}^x$. Formally,

$$\mathbf{S} = \mathbf{H}^v \mathbf{W} (\mathbf{H}^x)^{\mathrm{T}} \tag{2}$$

where $\mathbf{W} \in \mathbb{R}^{d_2 \times d_1}$ is a trainable matrix and $\mathbf{S}_{ij}$ denotes similarity between the i-th visual representation and the j-th textual representation. $\mathbf{S}$ is normalized row-wise to produce the vision-to-text attention weights $\mathbf{A}^x$, and column-wise to produce the text-to-vision attention weights $\mathbf{A}^v$:

$$\mathbf{A}^x = \mathrm{softmax}(\mathbf{S}) \in \mathbb{R}^{|\boldsymbol{v}| \times |\boldsymbol{x}|} \tag{3}$$

$$\mathbf{A}^v = \mathrm{softmax}(\mathbf{S}^{\mathrm{T}}) \in \mathbb{R}^{|\boldsymbol{x}| \times |\boldsymbol{v}|} \tag{4}$$

where $\mathrm{softmax}(\cdot)$ means row-wise normalization. Hence we can obtain the vision-aware textual representations $\hat{\mathbf{H}}^x \in \mathbb{R}^{|\boldsymbol{v}| \times d_1}$ by a product of the attention weight $\mathbf{A}^x$ and textual representation $\mathbf{H}^x$:

$$\hat{\mathbf{H}}^x = \mathbf{A}^x \mathbf{H}^x \tag{5}$$

Similarly, the text-aware visual representations $\hat{\mathbf{H}}^v \in \mathbb{R}^{|\boldsymbol{x}| \times d_2}$ can be obtained by:

$$\hat{\mathbf{H}}^v = \mathbf{A}^v \mathbf{H}^v \tag{6}$$

Since $\mathbf{H}^x$ and $\mathbf{H}^v$ mutually guide each other's attention, these two sources of information can mutually boost for better representations.

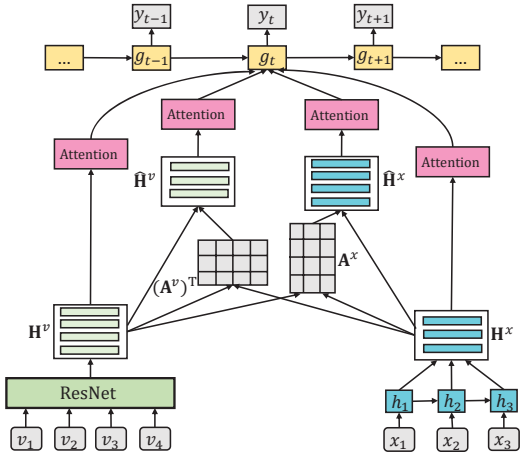

Figure 3: The overview of our proposed model.

3.3 Decoder

The decoder aims to generate the desired comment $\boldsymbol{y}$ via another GRU model. Since there exists information from multiple modalities, we equip decoder with multiple attention mechanisms. The hidden state g_{t+1} of decoder at time-step $t+1$ is computed as:

$$g_{t+1} = \mathrm{GRU}\big(g_t, [e(y_t); c_t^x; c_t^v; \hat{c}_t^x; \hat{c}_t^v]\big) \tag{7}$$

where semicolon represents vector concatenation, y_t is the word generated at time-step t and c_t^x is obtained by attending to $\mathbf{H}^x$ with g_t as query,

$$c_t^x = \mathcal{A}(g_t, \mathbf{H}^x) \tag{8}$$

where $\mathcal{A}$ refers to the attention mechanism. Readers can refer to Bahdanau et al. (2015) for the detailed approach. c_t^v, $\hat{c}_t^x$, and $\hat{c}_t^v$ are obtained in a similar manner by replacing $\mathbf{H}^x$ in Eq. (8) with $\mathbf{H}^v$, $\hat{\mathbf{H}}^x$, and $\hat{\mathbf{H}}^v$, respectively. Finally, the decoder samples a word y_{t+1} from the output probability distribution as follows:

$$y_{t+1} \sim \mathrm{softmax}(\mathbf{U} g_{t+1}) \tag{9}$$

[5] Multiple representations can be extracted from an image.

where $\mathbf{U}$ is a weight matrix. The model is trained by maximizing the log-likelihood of ground-truth $\boldsymbol{y}^* = (y_1^*, \cdots, y_n^*)$ and the loss function is:

$$\mathcal{L} = -\sum_{t=1}^{n} \log\Big(p(y_t^* | \boldsymbol{y}_{<t}^*, \boldsymbol{x}, \boldsymbol{v})\Big) \quad (10)$$

where $\boldsymbol{y}_{<t}^*$ denotes the sequence $(y_1^*, \cdots, y_{t-1}^*)$.

3.4 Extension to Transformer

We also extend our approach to Transformer (Vaswani et al., 2017). In detail, we adopt self-attention to implement the textual encoder. The representation of each word can be written as:

$$h_i^x = \text{SelfAtten}(x_i, \boldsymbol{x}) \quad (11)$$

which means that the multi-head attention component attends to the text $\boldsymbol{x}$ with the query x_i. We strongly recommend readers to refer to Vaswani et al. (2017) for the details of self-attention.

The decoder is also implemented with self-attention mechanism. More specifically, the hidden state of decoder at time-step t is calculated as:

$$g_t = \text{SelfAtten}(y_t, \boldsymbol{y}, \mathbf{H}^x, \mathbf{H}^v, \hat{\mathbf{H}}^x, \hat{\mathbf{H}}^v) \quad (12)$$

Inside the decoder, there are five multi-head attention components, using y_t as query to attend to $\boldsymbol{y}$, $\mathbf{H}^x$, $\mathbf{H}^v$, $\hat{\mathbf{H}}^x$, and $\hat{\mathbf{H}}^v$, respectively.

4 Experiments

4.1 Settings

The batch size is 64 and the vocabulary size is 15,000. The 512-dim embeddings are learned from scratch. The visual encoder is implemented as ResNet-152 (He et al., 2016a) pretrained on the ImageNet. For the Seq2Seq version of our approach, both textual encoder and decoder is a 2-layer GRU with hidden size 512. For the transformer version, we set the hidden size of multi-head attention to 512 and the hidden size of feed-forward layer to 2,048. The number of heads is set to 8, while a transformer layer consists of 6 blocks. We use Adam optimizer (Kingma and Ba, 2015) with learning rate 10^{-3} and apply dropout (Srivastava et al., 2014) to avoid over-fitting.

4.2 Baselines

We adopt the following competitive baselines:

Seq2Seq: We implement a series of baselines based on Seq2Seq. **S2S-V** (Vinyals et al., 2015) only encodes images via CNN as input. **S2S-T** (Bahdanau et al., 2015) is the standard Seq2Seq that only encodes texts as input. **S2S-VT** (Venugopalan et al., 2015) adopts two encoders to encode images and texts respectively.

Transformer: We replace the Seq2Seq in the above baselines with Transformer (Vaswani et al., 2017). The corresponding models are named **Trans-V**, **Trans-T**, and **Trans-VT**, respectively.

Models	BLEU-1	ROUGE-L	DIST-1	DIST-2
S2S-V	6.1	7.8	1348	3293
S2S-T	6.3	8.1	1771	4285
S2S-VT	6.6	8.5	1929	4437
Our (S2S)	**7.1**	**9.1**	**2279**	**4743**
Trans-V	5.9	7.6	1336	3472
Trans-T	6.4	8.3	1772	4694
Trans-VT	6.8	8.6	1891	4739
Our (Trans)	**7.7**	**9.4**	**2265**	**4941**

Table 3: Automatic evaluations of our method and baselines. **DIST-1** and **DIST-2** are the number of distinct unigrams and bigrams, respectively.

4.3 Evaluation Metrics

We adopt two kinds of evaluation methods: automatic evaluation and human evaluation.

Automatic evaluation: We use BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) to evaluate overlap between outputs and references. We also calculate the number of distinct n-grams (Li et al., 2016) in outputs to measure diversity.

Human evaluation: Three annotators score the 200 outputs of different systems from 1 to 10. The evaluation criteria are as follows. **Fluency** measures whether the comment is fluent. **Relevance** evaluates the relevance between the output and the input. **Informativeness** measures the amount of useful information contained in the output. **Overall** is a comprehensive metric. For each metric, the average Pearson correlation coefficient is greater than 0.6, indicating that the human scores are highly consistent.

4.4 Experimental Results

Table 3 and Table 4 show the results of automatic evaluation and human evaluation, respectively. We perform analysis from the following aspects.

The effectiveness of co-attention Both Table 3 and Table 4 show that our model can substantially outperform competitive baselines in all metrics.

Models	Flue.	Rele.	Info.	Overall
S2S-V	3.1	2.8	2.5	3.2
S2S-T	4.5	4.6	3.7	4.7
S2S-VT	4.6	5.1	4.3	4.9
Our (S2S)	**4.8**	**5.7**	**4.7**	**5.1**
Trans-V	2.9	2.3	2.8	2.9
Trans-T	4.3	4.8	4.4	4.6
Trans-VT	4.7	4.6	4.7	5.1
Our (Trans)	**4.9**	**5.9**	**5.0**	**5.2**

Table 4: Results of human evaluation. **Flue.**, **Rele.** and **Info.** denotes fluency, relevance, and informativeness, respectively.

For instance, the Transformer version of our approach achieves a 13% relative improvement of BLEU-1 score over Trans-VT. This illustrates that our co-attention can contribute to generating high-quality comments. The co-attention mechanism brings bidirectional interactions between visual information and textual information, so that two information sources can mutually boost for better representations, leading to improved performance.

The universality of co-attention Results show that both the Seq2Seq and Transformer version of our approach can outperform various baselines based on the same architecture. This shows that our co-attention has excellent universality, which can be applied to various model architectures.

The contribution of visual content According to Table 3 and Table 4, although the images contribute less to generating high-quality comments than texts, they still bring a positive impact on the generation. This illustrates that visual content contains additional useful information, which facilitates the generation of informative comments. Therefore, integrating multi-modal information is necessary for generating high-quality comments, which is also an important value of our work.

5 Related Work

In summary, this paper is mainly related to the following two lines of work.

Automatic article commenting. One similar task to CMAC is automatic article commenting. Qin et al. (2018) is the first to propose this task and constructs a large-scale dataset. Lin et al. (2018) proposes to retrieve information from user-generated data to facilitate the generation of comments. Furthermore, Ma et al. (2018) introduces a retrieval-based unsupervised model to perform generation from unpaired data. However, different from the article commenting that only requires extracting textual information for generation, the CMAC task involves not only the modeling of textual features but also the understanding of visual images, which poses a greater challenge to the intelligent systems.

Co-attention. We are also inspired by the related work of co-attention mechanism. Lu et al. (2016a) introduces a hierarchical co-attention model in visual question answering to jointly attend to images and questions. Xiong et al. (2017) proposes a dynamic co-attention network for the question answering task and Seo et al. (2017) presents a bi-directional attention network to acquire query-aware context representations in machine comprehension. Tay et al. (2018a) proposes a co-attention mechanism based on Hermitian products for asymmetrical text matching problems. Zhong et al. (2019) further presents a coarse-grain fine-grain co-attention network that combines information from evidence across multiple documents for question answering. In addition, the co-attention mechanism can also be applied to word sense disambiguation (Luo et al., 2018), recommended system (Tay et al., 2018b), and essay scoring (Zhang and Litman, 2018).

6 Conclusion

In this paper, we propose the task of cross-modal automatic commenting, which aims at enabling the AI agent to make comments by integrating multiple modal contents. We construct a large-scale dataset for this task and implement plenty of representative neural models. Furthermore, an effective co-attention model is presented to capture the intrinsic interaction between multiple modal contents. Experimental results show that our approach can substantially outperform various competitive baselines. Further analysis demonstrates that with multiple modal information and co-attention, the generated comments are more diverse and informative.

Acknowledgement

We thank the anonymous reviewers for their thoughtful comments. Xu Sun is the contact author of this paper.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, Conference Track Proceedings*.

David Chen and William B. Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, pages 190–200.

Deli Chen, Shuming Ma, Pengcheng Yang, and Xu Sun. 2018. Identifying high-quality chinese news comments based on multi-target text matching model. *arXiv preprint arXiv:1808.07191*.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1724–1734.

Zihang Dai, Zhilin Yang, Yiming Yang, William W Cohen, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016a. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016b. Deep residual learning for image recognition. In *2016 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.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, Conference Track Proceedings*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Proceedings, Part V*, pages 740–755.

Zhaojiang Lin, Genta Indra Winata, and Pascale Fung. 2018. Learning comment generation by leveraging user-generated data. *arXiv preprint arXiv:1810.12264*.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016a. Hierarchical question-image co-attention for visual question answering. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016*, pages 289–297.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016b. Hierarchical question-image co-attention for visual question answering. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016*, pages 289–297.

Fuli Luo, Peng Li, Jie Zhou, Pengcheng Yang, Baobao Chang, Zhifang Sui, and Xu Sun. 2019. A dual reinforcement learning framework for unsupervised text style transfer. *arXiv preprint arXiv:1905.10060*.

Fuli Luo, Tianyu Liu, Zexue He, Qiaolin Xia, Zhifang Sui, and Baobao Chang. 2018. Leveraging gloss knowledge in neural word sense disambiguation by hierarchical co-attention. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1402–1411.

Shuming Ma, Lei Cui, Furu Wei, and Xu Sun. 2018. Unsupervised machine commenting with neural variational topic model. *arXiv preprint arXiv:1809.04960*.

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 of the Association for Computational Linguistics*, pages 311–318.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2013. On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on Machine Learning*, pages 1310–1318.

Lianhui Qin, Lemao Liu, Wei Bi, Yan Wang, Xiaojiang Liu, Zhiting Hu, Hai Zhao, and Shuming Shi. 2018. Automatic article commenting: the task and dataset. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Volume 2: Short Papers*, pages 151–156.

Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. In *5th International Conference on Learning Representations, Conference Track Proceedings*.

Nitish Srivastava, Geoffrey E. 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.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014*, pages 3104–3112.

Yi Tay, Anh Tuan Luu, and Siu Cheung Hui. 2018a. Hermitian co-attention networks for text matching in asymmetrical domains. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 4425–4431.

Yi Tay, Anh Tuan Luu, and Siu Cheung Hui. 2018b. Multi-pointer co-attention networks for recommendation. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2309–2318.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, pages 6000–6010.

Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond J. Mooney, Trevor Darrell, and Kate Saenko. 2015. Sequence to sequence - video to text. In *2015 IEEE International Conference on Computer Vision*, pages 4534–4542.

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*, pages 3156–3164.

Caiming Xiong, Victor Zhong, and Richard Socher. 2017. Dynamic coattention networks for question answering. In *5th International Conference on Learning Representations, Conference Track Proceedings*.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. MSR-VTT: A large video description dataset for bridging video and language. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pages 5288–5296.

Pengcheng Yang, Xu Sun, Wei Li, Shuming Ma, Wei Wu, and Houfeng Wang. 2018. SGM: sequence generation model for multi-label classification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3915–3926.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78.

Haoran Zhang and Diane J. Litman. 2018. Co-attention based neural network for source-dependent essay scoring. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 399–409.

Yi Zhang, Jingjing Xu, Pengcheng Yang, and Xu Sun. 2018. Learning sentiment memories for sentiment modification without parallel data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1108.

Victor Zhong, Caiming Xiong, Nitish Shirish Keskar, and Richard Socher. 2019. Coarse-grain fine-grain coattention network for multi-evidence question answering. *arXiv preprint arXiv:1901.00603*.