CVPR #7245

3.3. Bag of Tricks

3.3.1 Mixing RoBERTa layers

RoBERTa consists of 24 layers of bidirectional transformer blocks. Given an input of length S, the pretrained RoBERTa encoder will return 25 sequences of embeddings, $\boldsymbol{G}_i \in \mathbb{R}^{2048 \times S}$ for $i \in \{0, 1, 2, ..., 24\}$. This includes the initial uncontextualized embeddings and the output of each of the 24 layers. Inspired by Tenney *et al.* [42], who showed that different layers in BERT represent different steps in the traditional NLP pipeline, we take a weighted sum across all layers to obtain the article embedding:

$$\boldsymbol{X}_T = \sum_{i=0}^{24} \alpha_i \boldsymbol{G}_i$$

where $\boldsymbol{x}_T \in \mathbb{R}^{1024 \times S}$ is the article embedding and α_i are learnable weights.

3.3.2 Copying with Multi-headed Attention

Inspired by pointer-generator networks [37], we introduce a copying mechanism using multi-headed attention.

3.3.3 Adaptive Softmax

The BPE vocabulary size is 50265. To make training more efficient, we use adaptive softmax [15] and divide the vocabulary into three clusters: 5K, 15K, and 25K. We tie the adaptive weights and we share the decoder input and output embeddings. We use sinusoidal positional encoding [43] to represent the position of each token.

4. Datasets

4.1. GoodNews

~~To compare to existing approaches we use~~ the GoodNews dataset, which until now was largest dataset for news image captioning [2]. Each example in the dataset is a triplet containing an article, an image, and a caption. Since only the article text, captions, and image URLs are publicly released the images need to be downloaded from the original source. Out of the 466K image URLs provided by [2], we were able to download 463K images, or 99.2% of the original dataset – the remaining are broken links.

We use this 99.2% sample of the GoodNews dataset and the train-validation-test split provided by [2]. There are 421K training, 18K validation, and 23K test captions. Note that this split was performed at the level of captions, so it is possible for a training and test caption to share the same article text (since articles have multiple images).

Table 1: Summary of news captioning datasets

	GoodNews	NYTimes800k
Number of articles	257 033	445 819
Number of images	462 642	794 044
Average article length	451	974
Average caption length	18	18
Collection start month	Jan 10	Mar 05
Collection end month	Mar 18	Sep 19
% of words that are		
– nouns	16%	16%
– pronouns	1%	1%
– proper nouns	23%	22%
– verbs	9%	9%
– adjectives	4%	4%
– named entities	27%	26%
– person~~al~~ names	9%	9%
% of captions with		
– named entities	97%	96%
– person~~al~~ names	68%	68%

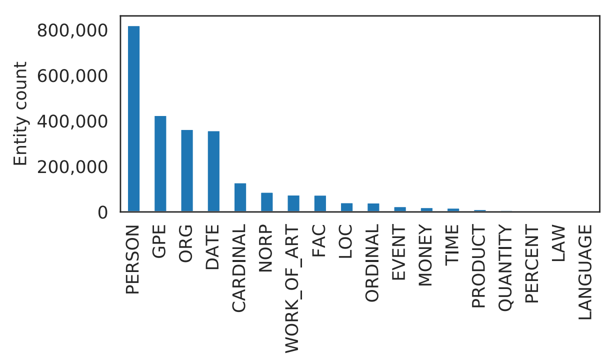

Figure 3: Entity distribution in NYTimes800k training captions. The four most common entity types are person~~al~~ names, geopolitical entities, organizations, and dates.

4.2. NYTimes800k

We constructed the NYTimes800k which is an 80% larger and more complete dataset of New York Times articles, images, and captions. The construction of this dataset was motivated by ~~the desire to clean up~~ data quality issues in the GoodNews dataset (as described below), collect a larger dataset, and include fine grained context such as the images location in the article.

We observed that many of the articles in the GoodNews dataset had been partially extracted when the generic article extractor used failed to recognise some of the HTML tags used specifically by the New York Times. Impor-

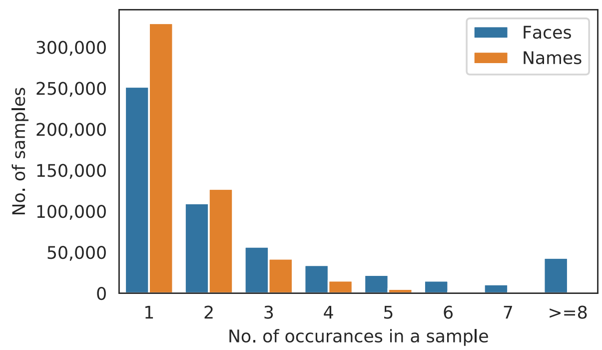

Figure 4: Co-occurrence of faces and personal names in NYTimes800k training data. The blue bars count how many images containing a certain number of faces. The orange bars count how many captions containing a certain number of personal names.

tantly, the missing text often included the first few paragraphs which frequently contain important information for captioning images. To collect the full articles for the NYTimes800k dataset we implemented a custom parser using The New York Times public API[1]. After this recollection, we observe that the average article is 963 words in comparison to GoodNews where the average length is 451 words. In addition, we found that GoodNews contains a small number of non-English articles, and captioned images from the recommendation sidebar which are not related to the main article. These were filtered out in the construction of NYTimes880k.

Table 1 presents a comparison between GoodNews and NYTimes800k. NYTimes800k exhibits several advantages:

- By increasing the collection period to the last 14 years (March 2005 – September 2019), NYTimes800k contains 80% more articles and images, thus becoming the largest news image captioning dataset.
- Using our custom parser, articles in NYTimes800k contain the full text with no missing paragraphs.
- NYTimes800k contains only English articles.
- We are careful to include only images that are part of the main article.
- Unlike GoodNews, we also collect information about where an image is located in the corresponding article. Most news articles have one image at the top that relates to the key topic. However 39% of the articles have at least one more image somewhere in the middle of text. The image placement and hence the text surrounding the image is important information for captioning as we show in our evaluations.

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

Table 2: NYTimes800k training, validation, and test splits

	Training	Validation	Test
Number of articles	434 272	3 052	8 495
Number of images	764 049	7 852	22 143
Start month	Mar 15	May 19	Jun 19
End month	Apr 19	May 19	Aug 19

Entities play an important role in the dataset, with 97% of captions containing at least one named entity. As shown in Figure 3, the most popular entity type are names of people, comprising a third of all named entities. Furthermore, 71% of training images contain at least one face and 68% of training captions mention at least one persons name. Figure 4 provides a further breakdown of the co-occurrence of faces and personal names. One important observation is that the majority of captions contain at most four names.

We split the training, validation, and test sets according to time, as shown in Table 2. For example, the test set consists of all captions and articles in the final three months of the collection period, from June to August 2019. This has two advantages over the random split used in GoodNews. Firstly, it prevents captions in the training and test sets from sharing the same context article, which allows us to evaluate how well the model can generalize to new articles. Secondly, due of the shift in the coverage of news over time, there will be events and people in the test data that have never been covered by the news before. In particular, out of the 100K proper nouns in the test captions, 4% never appear in any training captions. Half of these also never appear in any training article. Thus splitting by time allows us study how well the model can generate rare names.

5. Experiments

5.1. Training Details

In all experiments, we use Adam [19] with the weight decay fix [25] to optimize the models. We use the following parameters: $\beta_1 = 0.9, \beta_2 = 0.98, \epsilon = 10^{-6}$, and a weight decay of 10^{-5}. We clip the gradient norm at 0.1. All models see the same number of training examples, which is 6.6 million. This is equivalent to 16 epochs through GoodNews and 9 epochs through NYTimes800k. We warm up the learning rate in the first 5% of the training steps to 10^{-4}, and decays it linearly afterwards. Training is done with mixed precision to reduce the memory footprint. The full model takes 5 days to train on one Titan V GPU.

As shown in Table 3, all of our baselines are designed to

Table 3: Model complexity

	No. of Parameters
LSTM + GloVe	157M
Transformer + GloVe	148M
LSTM + weighted RoBERTa	159M
Transformer + RoBERTa	154M
+ weighted RoBERTa	154M
+ face attention	171M
+ copying	207M

have roughly the same number of trainable parameters. This allows us to attribute improvement to the individual model components rather than to simply having a bigger model.

The training pipeline is written in PyTorch [31] using the AllenNLP framework [14]. The RoBERTa model and dynamic convolution code are adapted from fairseq [29].

5.2. Results

Table 4 shows the BLEU [30], ROUGE [23], METEOR [7], and CIDEr [44] metrics on GoodNews and NYTimes800k. We show two best results from the previous state-of-the-art [2]. Note that the numbers reported here are slightly different from the original paper since we had to remove a few samples from the test set where the image is no longer available. [2] also did some post-processing on the ground-truth captions such as removing contractions and non-ASCII characters, both of which we did not do. Despite these differences, the final metrics are the same if rounded to the nearest whole number.

There is a strong correlation between of all these metrics, and in general, we mainly look at CIDEr since it uses Term Frequency Inverse Document Frequency (TF-IDF) to put more importance on less common words such as entity names. Table 5 shows the recall and precision of the named entities, personal names, and rare proper nouns.

We can make the following observations:

- Our baseline LSTM model with GloVe embeddings yields competitive results to previous the state-of-the-art [2]. This means that BPE offers a viable alternative to template-based methods.
- Models with GloVe embeddings are unable to generate rare proper nouns. This is expected since GloVe has a fixed vocabulary and if there is a unknown word in the article, the encoder will simply skip it.
- Switching from an LSTM to a transformer architecture improves the CIDEr score on NYTimes800k by 8 points, from 12 to 20. If we then use the contextualize RoBERTa embeddings instead of GloVe, CIDEr more than doubles to 44.
- Adding attention over the faces improves both the recall and precision of personal names. It has no significant effect on other entity types (see the supplementary materials for a detailed breakdown).

Table 5 also looks at the quality of the generated captions. We look at three metrics: caption length, type-token ratio (TTR), and Flesch reading ease. TTR is the ratio of the number of unique words to the total number of words in a caption. The Flesch reading ease takes into account the number of words and syllables and produces a score between 0 and 100, where higher means being easier to read.

From these metrics, we see that our generated captions are in general still shorter than real-life captions, have lower lexical diversity (lower TTR) and still use simpler language (higher Flesch reading ease).

6. Conclusion

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 2

[2] Ali Furkan Biten, Lluis Gomez, Marcal Rusinol, and Dimosthenis Karatzas. Good news, everyone! context driven entity-aware captioning for news images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1, 2, 5, 7, 9

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

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

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

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

[7] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics. 7

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

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

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

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

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

[13] Lianli Gao, Kaixuan Fan, Jingkuan Song, Xianglong Liu, Xing Xu, and Heng Tao Shen. Deliberate attention networks for image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8320–8327, 2019. 2

[14] Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia, July 2018. Association for Computational Linguistics. 7

[15] Édouard Grave, Armand Joulin, Moustapha Cissé, David Grangier, and Hervé Jégou. Efficient softmax approximation for GPUs. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1302–1310, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. 5

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 3

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

[18] Jin-Hwa Kim, Sang-Woo Lee, Donghyun Kwak, Min-Oh Heo, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Multimodal residual learning for visual qa. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 361–369. Curran Associates, Inc., 2016. 3

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

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

[21] Zhen-Zhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. *ArXiv*, abs/1909.11942. 2

[22] Jiangyun Li, Peng Yao, Longteng Guo, and Weicun Zhang. Boosted transformer for image captioning. *Applied Sciences*, 9(16):3260. 2, 3

[23] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. 7

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

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

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

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

[28] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., 2013. 3

[29] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 7

[30] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. 7

[31] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. In *NIPS Autodiff Workshop*. 7

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

[33] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2, 4

[34] Arnau Ramisa, Fei Yan, Francesc Moreno-Noguer, and Krystian Mikolajczyk. Breakingnews: Article annotation by image and text processing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40:1072–1085. 1, 2

[35] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1, 2

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

[37] Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada, July 2017. Association for Computational Linguistics. 5

[38] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725. Association for Computational Linguistics. 2, 4

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

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

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

[42] Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy, July 2019. Association for Computational Linguistics. 5

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017. 4, 5

[44] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 7

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

[46] Weixuan Wang, Zhihong Chen, and Haifeng Hu. Hierarchical attention network for image captioning. In *Proceedings*

of the AAAI Conference on Artificial Intelligence, volume 33, pages 8957–8964. 3

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

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

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

[50] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 3

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

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

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