Transform and Tell: Entity-aware News Image Captioning

Anonymous CVPR submission

Paper ID 7245

Abstract

We propose an end-to-end model to generate image captions in news articles. By combining the transformer architecture, byte-pair encoding, copying with multi-headed attention, and pretrained embeddings from three different modalities (RoBERTa for text, ResNet-152 for images, and FaceNet for faces), our system is able to describe an image with specific named entities mentioned in the article. Our model achieves a CIDEr score of 61 on the GoodNews dataset, significantly outperforming the previous state-of-art CIDEr of 13. We also introduce the NYTimes800k dataset, the largest news image captioning dataset to date. NYTimes800k is an extended version of GoodNews with higher-quality articles and metadata that allow us to study the importance of the image location within the text. On NYTimes800k, we achieve a CIDEr of 65. Pretrained models and source code are available from https://github.com/anonymized-link.

1. Introduction

The internet is home to a huge number of images, many of which lack useful captions. A growing body of work seeks to automatically generate captions that describe the objects and relationships using only visual cues extracted from the image itself [9, 45, 11, 17, 35, 27, 1, 5]. While generic image descriptions have their uses, such as for individuals with vision impairments, they are often of less benefit to the average user. To produce more useful image captions we need to go beyond generic descriptions and introduce information that cannot be gleaned directly from the image alone. Fortunately, many images have an associated context such as a news article, web page, or social media post, which give the image greater meaning than can be extracted from its pixels. To generate captions that go beyond generic description and actually add information that could not be gleaned from the image alone we must take this context into account. We focus on the news image captioning task in order to design practical methods for exploiting contextual information.

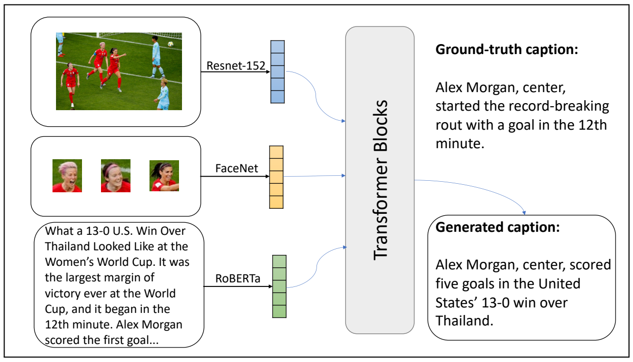

Resnet-152
FaceNet
What a 13-0 U.S. Win Over Thailand Looked Like at the Women's World Cup. It was the largest margin of victory ever at the World Cup, and it began in the 12th minute. Alex Morgan scored the first goal...
RoBERTa
Transformer Blocks
Ground-truth caption:
Alex Morgan, center, started the record-breaking rout with a goal in the 12th minute.
Generated caption:
Alex Morgan, center, scored five goals in the United States' 13-0 win over Thailand.

Figure 1: Our transformer model attends to embeddings from three different domains (image patches, faces, and article text). Using byte-pair encoding, the model can then directly produce a caption containing specific named entities without the use of templates.

News image captioning is an interesting instance of contextual captioning where news articles provide context to images.

Captions for news images, such as the example in Figure 1, typically contain details which cannot be derived from the image alone. They also frequently contain proper nouns such as names of people, places, and organizations – in many cases these proper nouns are rare (most people and places do not have many news articles written about them). A system capable of generating high quality news captions should therefor make extensive use of the provided context and be tuned for generating rare proper nouns. Existing approaches to news image captioning [41, 34, 2] rely on text extraction or template filling to deal with rare contextual terms such as names of people and organizations. This makes them relatively inflexibility and means they cannot be trained end-to-end. Moreover, existing approaches do not include specialized visual models for frequent nouns – experiments on the MSCOCO dataset have shown that pre-trained object detectors tuned for frequent nouns lead to more accurate captions [].

This motivates our novel fully end-to-end model for news image captioning that 1) combines specialised modules for incorporating and selectively attending to image features, human faces, and news article text and 2) applies a state-of-the-art sequence generation model which is able to generate rare tokens, such as proper names, even when they do not form part of the training data. Our model relies on a novel combination of sequence-to-sequence architectures, language representation learning, and vision systems.

In this paper we carefully consider the news image captioning problem and select a set of modeling tools which we combine into a novel architecture that sets a new state-of-the-art result. Our main contributions are threefold:

1. We introduce NYTimes800k, the largest news image captioning dataset to date, containing 446K articles and 794K images with captions from The New York Times spanning 14 years. NYTimes800k builds on the GoodNews dataset; but we write a custom parser to collect higher-quality articles and metadata such as the location of an image within the page.

2. We build a captioning model that combines the power of transformers, byte-pair encoding, copying via multi-headed attention, and attention over three different modalities (text, images, and faces). We show that our model achieves state-of-the-art results with a significant margin over previous methods, and in particular, it can generate names not seen during training without the use of templates.

3. We provide a detailed model analysis, deconstructing the most important modeling components and quantifying the incremental contribution that each of them makes not only to the usual metrics such as BLEU, ROUGE, METEOR, and CIDEr; but also to other linguistic measures like readability scores, caption length, and recall of rare names.

2. Related Works

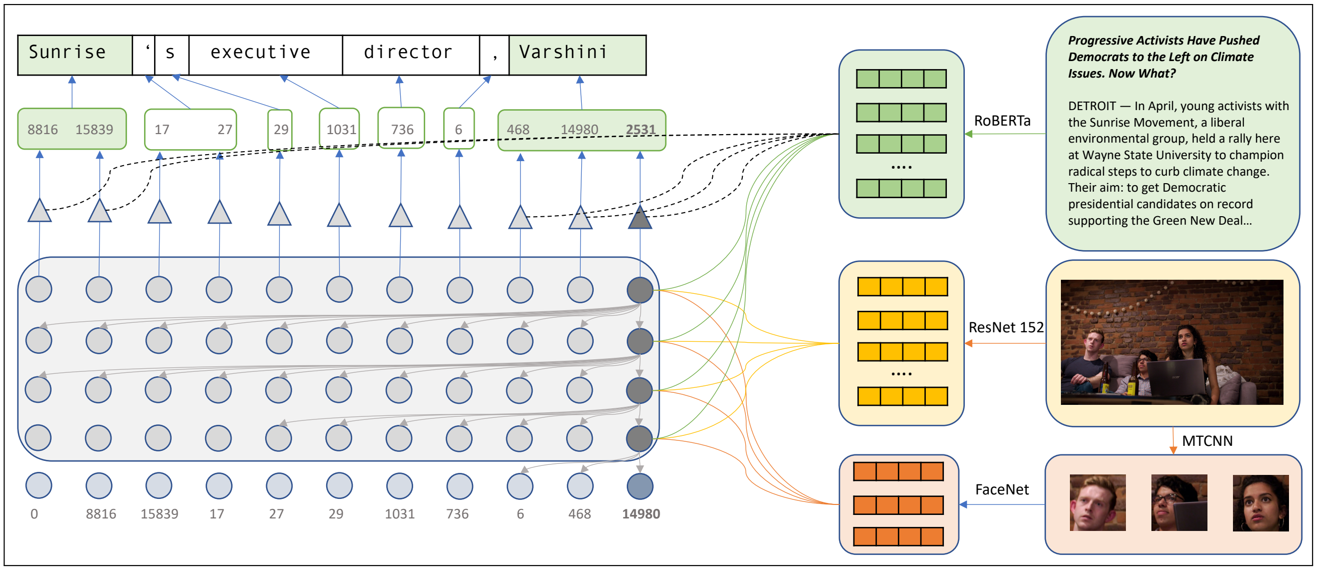
~~A large number of methods exist for generating generic image captions that describe objects and relationships using only the image as input.~~ Many of these captioning systems use some combination of a Convolutional Neural Network encoder and an RNN with a closed vocabulary as a decoder [17, 9, 45]. Attention over image patches was introduced in "Show, Attend and Tell" [48], in which the attention weights are obtained by feeding the image embeddings and the previous hidden state of the RNN through a multilayer perception. Many extensions to these models have been proposed such as giving the model the option to not attend to any image region [27], using reinforcement learning to directly optimise for the CIDEr metric [35, 13], and using a bottom-up approach to propose a region to attend to [1]. All of these systems generate restricted vocabulary generic captions without considering context external to the image.

A related task which does consider image context is news image captioning, where the image caption is generated using the article text as context. One key challenge of news image captioning is generating rare entity names, for example the names of people who do not make many media appearances. Early non-neural approaches include extractive methods that use n-gram models to combine existing phrases [12] or simply retrieving the most representative sentence [41] in the article. The neural network approach taken by Ramisa *et al.* [34] was able to generate entirely new text with an LSTM decoder that took as input a word2vec representation of the article concatenated with a CNN representation of the image. Even so this approach was unable produce names that were not seen during training.

To overcome the limitation of a fixed-size vocabulary caption templates can be used. This involves first generating a template sentence with placeholders for named entities, e.g. "PERSON speaks at BUILDING in DATE.". This template can be generated using an LSTM or other sequence generation model [2]. Afterwards, a selection algorithm picks the best candidate for each placeholder. For example, Lu *et al.* [26] built a knowledge graph for each combination of entities and select the most likely combination. Meanwhile, Biten *et al.* [2] filled the template by extracting entities from the sentence in the article which had the highest cosine similarity to the template. One key difference between our proposed model and that of previous approaches[2, 26] is that our model can generate a captions with named entities directly – without using an intermediate template.

One tool that has had seen recent successes in many natural language processing tasks is the transformer neural network. Transformers have been show to consistently outperforming Recurrent Neural Network architectures in language modeling [33], story generation [10], summarization [40], and machine translation [3]. Furthermore, transformer based models such as BERT [8], XLM [20], XLNet [49], RoBERTa [24], and ALBERT [21] have been shown to produce high level text representations suitable for transfer learning. Furthermore, using byte-pair encoding (BPE) [38] to represent uncommon words as a sequence of subword units can enable the transformer function in an open vocabulary setting.

Transformers have been shown to yield competitive results in generating generic MS COCO captions [53, 22]. Zhao *et al.* [52] have gone further and trained transformers to produce some named entities in the Conceptual Captions dataset [39]. However, this dataset provides no additional context to the image, making it a different problem to news image captioning. Moreover, the authors used web-



Progressive Activists Have Pushed Democrats to the Left on Climate Issues. Now What?

DETROIT — In April, young activists with the Sunrise Movement, a liberal environmental group, held a rally here at Wayne State University to champion radical steps to curb climate change. Their aim: to get Democratic presidential candidates on record supporting the Green New Deal...

Figure 2: On the left is the decoder with four transformer blocks, which takes as input embeddings of byte-pair tokens (e.g. 14980 represents “arsh” in “Varshini”). The grey arrows inside the decoder indicate the convolutions in each block (we only show the convolutions in the final time step). After each convolution, we also attend to three domains on the right: article text (green lines), image patches (yellow lines) and faces (orange lines). The grey triangles are switches to determine if we should copy or not. If we copy, we attend to the article embeddings (dashed black lines) and copy the token with the highest attention score. The final decoder outputs are byte-pair tokens, which are then combined to form whole words and punctuations. Copied tokens and words are shaded in green.

entity labels, extracted using Google Cloud Vision API, as inputs to the model. In our work, we do not explicitly give the model a list of entities to appear in the caption, instead our model automatically identifies relevant entities from the provided news article.

BPE offers an elegant solution to handling an open vocabulary. To date the only image captioning work that uses BPE is [52], but they did not use it for rare named entities as these were removed from the captions during preprocessing. In contrast we explicitly examine the use of BPE for generating rare names and evaluate it in comparison to template-based methods.

In addition to attending to image patches, some captioning models also attend to object regions [46] and visual concepts [50, 22, 46], both of which are derived from the image itself. When attending to more than one modality, there are various strategies on how to combine embeddings such as addition, concatenation, and multivariate residual modules (MRMs) [18]. ~~In our model we use the vector concatenation strategy and leave investigation of the more complex strategies, such as MMRs~~, to future work as they typically only yield minor performance improvements [46].

3. Transformer for News Captions

Conceptually our model can be broken into two parts: encoding and decoding. The encoding consists of a set of domain-specific encoders for producing high-level vector representations of images, faces and article text. The output of each encoder is a set of fixed size vectors that represent the input. The decoder sequentially generates captions at the sub-word level by applying multi-head attention over the sets of vectors from the encoders, and over a representation of the previously generated sub-word units. In practice there are a number details which allow us to train this large multi-faceted model on a single machine and achieve state-of-the-art performance. We have included some these details where appropriate and collected together those that do not fit elsewhere into Section 3.3.

3.1. Encoders

Our proposed model takes three types of inputs: image, faces, and article text. Each of these inputs is encoded into a set of vectors by domain specific encoders pre-trained on data from the matching domain.

Image Encoder: A high level image representation is obtained with a ResNet-152 [16] model pre-trained on ImageNet. We use the output of the final block before the pooling layer as the image representation. This is a set of

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, 6, 8, 10

[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, Belgium, Brussels, Oct. 2018. 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, 2017. 4

[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, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 2, 4

[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. 7

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

[22] Jiangyun Li, Peng Yao, Longteng Guo, and Weicun Zhang. Boosted transformer for image captioning. *Applied Sciences*, 9(16):3260, 2019. 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, 2019. 2, 4

[25] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 7

[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. 4

[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*, 2017. 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, 2014. 4

[33] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. 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, 2016. 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. 4

[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, Berlin, Germany, Aug. 2016. 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, 2019. 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, 2019. 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*, 2019. 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*, 2015. 2, 4

[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, 2019. 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, 2016. 4

[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, 3

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