Figure 2: Overall architecture of the model.

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. Model Architecture

Conceptually our model can be broken into two parts: encoding and decoding. The encoding part 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~~ potentially arbitrary, length set of fixed size vectors that represent the input. The decoder sequentially generates captions at the sub-word level by applying multi-headed 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.

3.1.1 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 49 different vectors $\boldsymbol{x}_{Ii} \in \mathbb{R}^{2048}$ where each vector corresponds to a separate image patch after the image is divided into equal size 7 by 7 patches. Using this representation $\boldsymbol{X}_I = \{\boldsymbol{x}_{Ii} \in \mathbb{R}^{2048}\}_{i=1}^{49}$ allows the decoder to attend different regions in the image—a modeling choice that has proven useful in ~~other~~ image captioning tasks [48].

3.1.2 Face Encoder

We use MTCNN [51] to detect face bounding boxes in the image. We then select the largest four faces since the majority of captions have at most four personal names (see 4.2). A vector representation of each face is obtained by passing the bounding boxes to FaceNet [36], which was pre-trained on the VGGFace2 dataset [4]. The resulting set of face vectors for each image is $\boldsymbol{x}_F = \{\boldsymbol{x}_{Fi} \in \mathbb{R}^{512}\}_{i=1}^{4}$.

Even though the faces are extracted from the image it is useful to consider them as an input domain that is separate to the image. This is because specialised models are needed to make full use of them.

3.1.3 Article Encoder

To encode the article text we use RoBERTa [24] which is a recent improvement over the popular BERT [8] model. RoBERTa is a language representation model that provides pretrained contextual embeddings for text. It consists of 24 layers of bidirectional transformer blocks.

Unlike GloVe [32] and word2vec [28] embeddings, where each word has exactly one representation, the bidirectionality and the attention mechanism in the transformer allow a word to have different vector representations depending on the surrounding context.

The largest GloVe model has a vocabulary size of 1.2 million. Although this is large, many rare names will still get mapped to the unknown token. In contrast, RoBERTa uses BPE [38, 33] which can encode any word that can be written in Unicode characters.

One limitation of RoBERTa is that the maximum length of the input sequence is 512. For GoodNews, we simply encode the first 512 tokens of the article. For NYTimes800k, since we have the image position, we concatenate the title, the first paragraph, and as many paragraphs above and below the image as we can fit, until we reach the 512 token limit. Note that since we are using BPE, a word might consist of many tokens. On average, we can only encode ...... words of the article.

The RoBERTa encoder provides gives us the set of token embeddings $\boldsymbol{X}_T = \{\boldsymbol{x}_{Ti} \in \mathbb{R}^{1024}\}_{i=1}^{S}$, where S is the number of tokens.

3.2. Decoder

The decoder is a function that estimates $p(y_t)$, the probability of the tth token in the caption, conditional on the past $\boldsymbol{y}_{<t}$ and the context embeddings $\boldsymbol{X}_I$, $\boldsymbol{X}_T$, and $\boldsymbol{X}_F$:

$$p(y_t) = \mathbb{P}(Y_t = y_t \mid \boldsymbol{y}_{<t}, \boldsymbol{X}_I, \boldsymbol{X}_T, \boldsymbol{X}_F)$$

In our architecture, the decoder consists of four transformer blocks. In each block, the conditioning on past tokens is computed using dynamic convolutions [47], and the conditioning on the contexts is computed using multi-head attention [43].

3.2.1 Dynamic Convolutions

Instead of using the standard self-attention module as in current state-of-the-art GPT-2 decoder [33], we find that dynamic convolutions [47] are more efficient to train. Suppose when decoding the tth token, we have at the ℓth block the input $\boldsymbol{z}_{\ell t} \in \mathbb{R}^{1024}$. If $\ell = 0$, then $\boldsymbol{z}_{0t}$ is the embedding of the previous token. Otherwise it is the output from the previous transformer block. Given kernel size K and 16 attention heads, for each head $h \in \{1, 2, ..., 16\}$, we first project the current and last $K-1$ steps using a feedforward layer:

$$\boldsymbol{z}'_{\ell,h,t-j} = \mathrm{GLU}(\boldsymbol{W}_{z\ell h}\, \boldsymbol{z}_{\ell,t-j} + \boldsymbol{b}_{z\ell h})$$

where $j \in \{0, 1, ..., K-1\}$, GLU is the gated linear unit activation function [6], and $\boldsymbol{z}'_{\ell,h,t-j} \in \mathbb{R}^{64}$. The output of each head's dynamic convolution is the weighted sum of these projected values:

$$\tilde{\boldsymbol{z}}_{\ell h t} = \sum_{j=0}^{K-1} \gamma_{\ell h j}\, \boldsymbol{z}'_{\ell,h,t-j}$$

where the weight $\gamma_{\ell h j}$ is a linear projection of the input, followed by a softmax over the kernel window:

$$\gamma_{\ell h j} = \mathrm{Softmax}\left(\boldsymbol{w}_{\gamma \ell h}^{T}\, \boldsymbol{z}'_{\ell,h,t-j}\right)$$

The overall output is the concatenation of all the head outputs, followed by a feedforward with a residual connection and layer normalization:

$$\begin{aligned}
\tilde{\boldsymbol{z}}_{\ell t} &= [\tilde{\boldsymbol{z}}_{\ell 1 t}, \tilde{\boldsymbol{z}}_{\ell 2 t}, ..., \tilde{\boldsymbol{z}}_{\ell 16 t}] \\
\boldsymbol{d}_{\ell t} &= \mathrm{LayerNorm}(\boldsymbol{z}_{\ell t} + \boldsymbol{W}_{\tilde{z}\ell}\, \tilde{\boldsymbol{z}}_{\ell t} + \boldsymbol{b}_{\tilde{z}\ell})
\end{aligned}$$

Note that given kernel size K, we can attend to the current time step and the last $K-1$ steps. Following ~~closely to~~ [47], our decoder has 4 transformer blocks with kernel sizes 3, 7, 15, and 31, respectively. Thus the final block output will have collected information from the last 51 tokens.

3.2.2 Multi-Head Attention

Given $\boldsymbol{d}_{\ell t} \in \mathbb{R}^{1024}$, the output of the dynamic convolution at layer ℓ, we can now attend over the image context using multi-head attention, also with 16 heads. For each head $h \in \{1, 2, ..., 16\}$, we first do a linear projection of $\boldsymbol{d}_{\ell t}$ and the image embeddings $\boldsymbol{X}_I$ into a query $\boldsymbol{q}_{I\ell h t} \in \mathbb{R}^{64}$, a set of keys $\boldsymbol{K}_{I\ell h t} = \{\boldsymbol{k}_{I\ell h t i} \in \mathbb{R}^{64}\}_{i=1}^{49}$, and the corresponding values $\boldsymbol{V}_{I\ell h t} = \{\boldsymbol{v}_{I\ell h t i} \in \mathbb{R}^{64}\}_{i=1}^{49}$:

$$\begin{aligned}
\boldsymbol{q}_{I\ell h t} &= \boldsymbol{W}_{I\ell h q}\, \boldsymbol{d}_{\ell t} \\
\boldsymbol{k}_{I\ell h i} &= \boldsymbol{W}_{I\ell h k}\, \boldsymbol{x}_{I i} \qquad \forall i \in \{1, 2, ..., 49\} \\
\boldsymbol{v}_{I\ell h i} &= \boldsymbol{W}_{I\ell h v}\, \boldsymbol{x}_{I i} \qquad \forall i \in \{1, 2, ..., 49\}
\end{aligned}$$

Then the attended image for each head is the weighted sum of the values, where the weights are obtained from the dot product between the query and key:

$$\begin{aligned}
\lambda_{I\ell h i} &= \mathrm{softmax}\left(\boldsymbol{k}_{I\ell h i}^{T} \boldsymbol{q}_{I\ell h t}\right) \\
\boldsymbol{x}'_{I\ell h t} &= \sum_{i=1}^{49} \lambda_{I\ell h i}\, \boldsymbol{v}_{I\ell h i}
\end{aligned}$$

The attention from each head is then concatenated into $\boldsymbol{x}'_{I\ell t} \in \mathbb{R}^{1024}$:

$$\boldsymbol{x}'_{I\ell t} = [\tilde{\boldsymbol{x}}_{I\ell 1 t}, \tilde{\boldsymbol{x}}_{I\ell 2 t}, ..., \tilde{\boldsymbol{x}}_{I\ell 16 t}]$$

and the overall image attention $\tilde{\boldsymbol{x}}_{I\ell t} \in \mathbb{R}^{1024}$ is obtained after adding a residual connection and layer normalization:

$$\tilde{\boldsymbol{x}}_{I\ell t} = \text{LayerNorm}(\boldsymbol{d}_{\ell t} + \boldsymbol{x}'_{I\ell t})$$

We use the same multi-head attention mechanism (with different weight matrices) to obtain the attended article $\tilde{\boldsymbol{x}}_{T\ell t}$ and the attended face $\tilde{\boldsymbol{x}}_{F\ell t}$. These three are finally concatenated and fed through a feedforward layer:

$$\begin{aligned}
\tilde{\boldsymbol{x}}_{C\ell t} &= [\tilde{\boldsymbol{x}}_{I\ell t}, \tilde{\boldsymbol{x}}_{T\ell t}, \tilde{\boldsymbol{x}}_{F\ell t}] \\
\tilde{\boldsymbol{x}}_{R\ell t} &= \boldsymbol{W}_{C\ell}\, \tilde{\boldsymbol{x}}_{C\ell t} + \boldsymbol{b}_{C\ell} \\
\tilde{\boldsymbol{x}}_{D\ell t} &= \text{ReLU}(\boldsymbol{W}_{R\ell}\, \tilde{\boldsymbol{x}}_{R\ell t} + \boldsymbol{b}_{R\ell}) \\
\boldsymbol{z}_{\ell+1,t} &= \text{LayerNorm}(\tilde{\boldsymbol{x}}_{R\ell t} + \boldsymbol{W}_{D\ell}\, \tilde{\boldsymbol{x}}_{D\ell t} + \boldsymbol{b}_{D\ell})
\end{aligned}$$

The final output $\boldsymbol{z}_{\ell+1,t} \in \mathbb{R}^{1024}$ is used as the input to the next transformer block, or if we are in the last block, it is used to compute the logits over the token vocabulary.

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, $\boldsymbol{G} = \{\boldsymbol{g}_{\ell i} \in \mathbb{R}^{2048} : i \in \{1, 2, ..., 49\}, \ell \in \{0, 1, ..., 24\}\}$. 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 $\boldsymbol{x}_{Ti}$:

$$\boldsymbol{x}_{Ti} = \sum_{\ell=0}^{24} \alpha_\ell\, \boldsymbol{g}_{\ell i}$$

where α_ℓ are learnable weights.

3.3.2 Copying with Multi-headed Attention

Inspired by pointer-generator networks [37], we introduce a copying mechanism using multi-head attention. We use the final layer output $\boldsymbol{z}_{5t}$ and the article embeddings $\boldsymbol{X}_T$ as inputs to the multi-head attention module. Unlike 3.2.2, we only need to compute the softmax weights λ_i (and not the weighted sum of the values). We interpret each λ_i as the probability of copying ith token in the article.

3.3.3 Adaptive Softmax

The decoder 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.

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%
– personal names	9%	9%
% of captions with		
– named entities	97%	96%
– personal names	68%	68%

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).

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 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

Table 4: BLEU, ROUGE, METEOR, and CIDEr metrics on GoodNews and NYTimes800k.

		BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	METEOR	CIDEr
GoodNews	Biten (Avg + CtxIns) [2]	9.04	3.66	1.71	0.89	12.2	4.37	13.1
	Biten (TBB + AttIns) [2]	8.10	3.26	1.48	0.76	12.2	4.17	12.7
	LSTM + GloVe	14.0	6.52	3.41	2.03	13.7	5.57	14.3
	Transformer + GloVe	18.3	9.49	5.45	3.43	17.0	7.52	25.7
	LSTM + weighted RoBERTa	19.2	10.5	6.28	4.04	18.0	8.32	35.4
	Transformer + RoBERTa							
	+ weighted RoBERTa	22.2	13.4	8.68	5.99	21.2	10.1	52.9
	+ face attention	22.4	13.6	8.84	6.10	21.3	10.3	53.9
	+ copying	**24.2**	**14.5**	**9.24**	**6.22**	**22.6**	**11.5**	**60.6**
NYTimes800k	LSTM + GloVe	13.4	6.00	3.05	1.76	13.2	5.36	12.2
	Transformer + GloVe	17.0	8.42	4.63	2.79	16.1	6.99	20.6
	LSTM + weighted RoBERTa	18.0	9.88	5.97	3.91	17.1	7.96	30.8
	Transformer + RoBERTa							
	+ weighted RoBERTa	20.7	12.4	8.14	5.73	19.8	9.54	44.1
	+ location-aware	21.7	13.3	8.84	6.25	21.3	10.3	52.4
	+ face attention	22.1	13.6	9.05	6.41	21.7	10.4	54.7
	+ copying	**24.3**	**15.2**	**10.0**	**7.03**	**23.7**	**12.0**	**65.3**

Table 5: Named entity, personal name, and rare proper noun recall (R) & precision (P) on GoodNews and NYTimes800k. Recall and precision are expressed as percentages. Linguistic measures on the generated captions: caption length (CL), type-token ratio (TTR), and Flesch readability ease (FRE).

		Named entities		Personal names		Rare proper nouns		CL	TTR	FRE
		R	P	R	P	R	P			
GoodNews	Ground truths	–	–	–	–	–	–	18.1	94.9	65.4
	Biten (Avg + CtxIns) [2]	6.06	8.23	6.55	9.38	–	–	9.9	92.2	78.3
	Biten (TBB + AttIns) [2]	5.64	8.87	6.98	11.9	–	–	9.1	90.7	77.6
	LSTM + GloVe	7.26	11.1	5.76	9.48	–	–	13.8	89.6	77.5
	Transformer + GloVe	10.9	14.4	10.7	14.7	–	–	15.5	88.5	73.9
	LSTM + weighted RoBERTa	13.8	17.4	16.0	20.6	–	–	15.2	89.2	74.7
	Transformer + RoBERTa					–	–			
	+ weighted RoBERTa	18.4	21.6	22.4	28.1	–	–	15.5	91.0	72.0
	+ face attention	18.7	22.1	23.2	29.2	–	–	15.5	90.7	71.9
	+ copying	**22.5**	**26.7**	**27.1**	**35.7**	–	–	15.3	90.2	69.9
NYTimes800k	Ground truths	–	–	–	–	–	–	18.4	94.6	63.9
	LSTM + GloVe	7.26	10.2	5.69	8.61	0	0	13.8	89.0	77.8
	Transformer + GloVe	10.9	13.4	9.50	13.5	0	0	15.1	88.6	73.8
	LSTM + weighted RoBERTa	15.0	17.1	18.1	22.6	15.1	15.2	14.9	90.2	72.6
	Transformer + RoBERTa									
	+ weighted RoBERTa	19.5	21.0	25.5	30.3	22.6	29.1	15.3	91.5	70.4
	+ location-aware	21.9	24.1	30.2	35.5	26.2	32.5	15.1	91.7	70.4
	+ face attention	22.3	24.5	31.3	37.1	26.6	33.6	15.2	91.6	70.5
	+ copying	**28.5**	**31.7**	**38.3**	**47.7**	**38.3**	**39.9**	15.0	91.0	68.7

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, 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. 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, 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*. 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. 4

[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, 3

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