

Exploiting Image-Text Synergy for Contextual Image Captioning

Sreyasi Nag Chowdhury*

sreyasi@mpi-inf.mpg.de

Rajarshi Bhowmik†

rajarshi.bhowmik@rutgers.edu

Hareesh Ravi†

hr268@rutgers.edu

Gerard de Melo†‡

gdm@demelo.org

Simon Razniewski*

srazniew@mpi-inf.mpg.de

Gerhard Weikum*

weikum@mpi-inf.mpg.de

*Max Planck Institute for Informatics

† Rutgers University

‡ Hasso Plattner Institute

Abstract

Modern web content – news articles, blog posts, educational resources, marketing brochures – is predominantly multimodal. A notable trait is the inclusion of media such as images placed at meaningful locations within a textual narrative. Most often, such images are accompanied by captions – either factual or stylistic (humorous, metaphorical, etc.) – making the narrative more engaging to the reader. While standalone image captioning has been extensively studied, captioning an image based on external knowledge such as its surrounding text remains under-explored. In this paper, we study this new task: given an image and an associated unstructured knowledge snippet, the goal is to generate a *contextual caption* for the image.

1 Introduction

In multimodal (image–text) documents, images are typically accompanied by captions. These may contain specific knowledge about the narrative – location, names of persons etc. – or may exhibit thematic knowledge grounding the sentimental value of the image in the narrative. The image captions explicitly or implicitly refer to the image and its surrounding text, and play a major role in engaging the reader. We call this type of captions *contextual captions*, and Figure 1 illustrates the corresponding *Contextual Image Captioning* problem.

Generating captions for standalone images (Hossain et al., 2019; Wang et al., 2020) or summarizing a piece of text (See et al., 2017; Lin and Ng, 2019) are well-studied problems. However, generating image captions accounting for contextual knowledge is a largely unexplored task and poses many challenges. Related tasks include multimodal summarization (Chen and Zhuge, 2018, 2019a) and title generation (Murao et al., 2019). Multimodal summarization usually involves segmentation and



I recently moved to Buffalo, NY and every day I am discovering how beautiful this town is. I took this pic...and I was thrilled about it! I wanted to share the pallet of colors the sunset had that evening.

Generated Contextual Captions:

- A beautiful sunset path to heaven.

- A sunset...unknown artist.

Figure 1: Our novel *Contextual Captions* capture the affective theme from a given image and its associated paragraph.

sorting of both the modalities or has specific templates along which the summary is generated (See et al., 2017). In contrast, generating contextual captions requires conditionally deciding to follow, lead or negate the knowledge offered by the context.

Inadequacy of Prior Work. Image captioning and text summarization are unimodal, and ignore information present in the dormant modality. Multimodal summarization and news image captioning (Biten et al., 2019) usually entail captions with explicit references to the context, and may be achieved with a copy mechanism (Gu et al., 2016) that can selectively copy information (e.g., named entities such as names of people, organizations, geographical locations etc.) from the surrounding text to the caption. However, most social media driven content is affective and requires implicit reasoning about the context. For example, for an image of the Grand Canyon, we might encounter captions such as “perfect for a lovely hike” or “too tired to walk”, due to the subjectivity of the task, which requires inference based on the context.

Approach and Contribution.

- We formulate the novel task of *Contextual Image Captioning* and create a new dataset from Reddit posts with images, titles and comments.
- We propose an end-to-end trained neural model for the *Contextual Image Captioning* task and comprehensively evaluate its performance using quantitative and qualitative measures.
- We study how various factors affect the generation of novel *contextual captions*.

To foster follow-up research we release the dataset and code, available at https://github.com/Sreyasi/contextual_captions.

2 Related Work

Image Captioning. Prior work on captioning conditioned only on images (Farhadi et al., 2010; Vinyals et al., 2015; Karpathy and Li, 2015; Krause et al., 2017) has been successful for descriptive captions with explicit grounding to image objects. Recently, captions with sentimental and abstract concepts have been explored (Gan et al., 2017; Chandrasekaran et al., 2018; Park et al., 2017; Liu et al., 2018; Shuster et al., 2019). Although external knowledge bases like DBpedia (factual knowledge) (Wu et al., 2018) and ConceptNet (commonsense knowledge) (Zhou et al., 2019) have been leveraged, all prior work ignores the knowledge present in the text surrounding images in social media and other domains. *Contextual Image Captioning* leverages the latter kind of knowledge.

Multimodal Summarization. Research on multimodal embeddings (Laina et al., 2019; Xia et al., 2020; Scialom et al., 2020) has facilitated studying image–text data. Summarization of multimodal documents (Chu and Kao, 2017; Zhu et al., 2018; Hessel et al., 2019) proceeds by aligning a subset of images with extracted (Chen and Zhuge, 2019a), or generated (Chen and Zhuge, 2018) text segments from the original document. In contrast, image captions in our dataset do not explicitly summarize the associated text and rather act as a short commentary that captures knowledge from both modalities.

News Image Captioning. A task similar to our problem is captioning images within news articles (Tran et al., 2020; Chen and Zhuge, 2019b). A key challenge here is to correctly identify and generate named entities (Tran et al., 2020). However, news image captions tend to be descriptive compared to the subjective nature of captions in our dataset representing common social media content.

3 Datasets

To the best of our knowledge, the only existing image–text caption datasets are from the news domain (e.g., Daily Mail Corpus) containing non-affective descriptive captions with mentions of named entities. Instead, we consider Reddit, which offers a rich source of multimodal data. Out of the image-related subreddits, */r/pics* is particularly suitable for our problem because of the nature of posts. Firstly, the posts do not contain expert jargon, unlike other subreddits like */r/photographs*. Secondly, the image captions are mostly affective and not drab descriptions. Lastly, post frequency is high, presenting a big dataset.

Data Scraping. We scrape the subreddit */r/pics* to collect 250,000 posts over the span of a year. For each post, we grab the image, the post title, and 1-10 comments. We consider the post title as ground truth caption since it is written by the image poster, ensuring a consistent and coherent intent. The comments are concatenated, preserving their tree structure, to serve as the unstructured knowledge associated with the image. Inappropriate posts that do not adhere to community standards and were flagged by moderators are removed.

Data Characteristics. The collected images do not adhere to any particular subject or theme. The paragraphs are ~59.2 words long, and the captions are ~10.6 words long on an average.

In some posts, captions and comments may contain different named entities (NE), making prediction of the ground truth NE difficult. For example, the caption “My friend and I are en route to the Grand Canyon” may be accompanied with the comment “Try to hike down to the Colorado. Also visit Zion National Park!” The NEs in the paragraph (Colorado, Zion) do not match that in the caption (Grand Canyon). Owing to this characteristic, we study two distinct variants of the dataset – one containing NEs and the other without NEs. We denote these variants as $+NE$ and NE , respectively.

The comments sometimes exhibit topic drift, e.g., a comment on the Grand Canyon post may be “I remember my last trip to India...we had spicy food!”. Hence, we also study variants with ensured context overlap – one common word (ignoring stop words) between caption and comments. These variants are suffixed *overlap* – e.g. $+NE-overlap$.

We report experimental results on all of these variants, adopting a 30,000/8,000/8,000 train/val/test split for each of them.

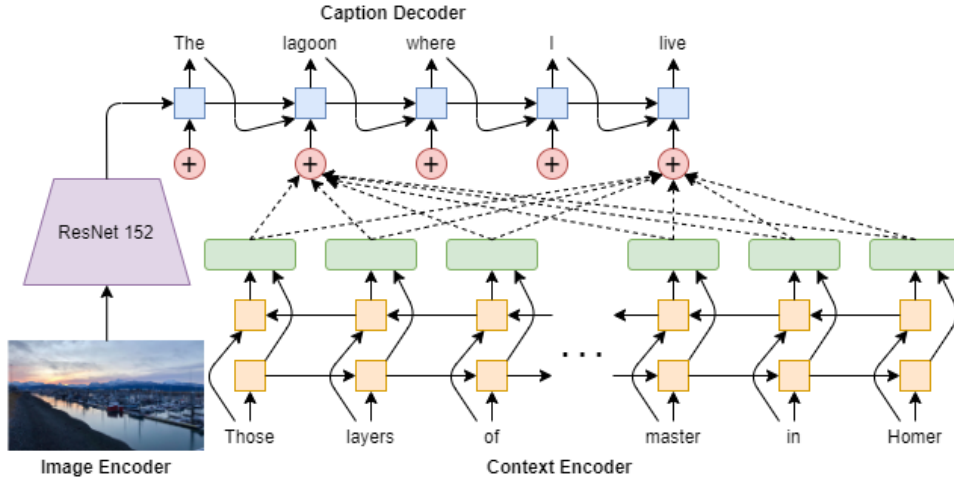


Figure 2: A schematic diagram of our contextual captioning model

4 Contextual Captioning Model

We refer to the text context associated with each image as ‘paragraph’. This offers external knowledge which may be absent in the image modality alone. Figure 2 shows our proposed model architecture. Given an input image I and an associated input paragraph $P = \{w_1^p, \dots, w_M^p\}$ of length M , our model (an encoder–decoder architecture) generates a caption $C = \{w_1^c, \dots, w_N^c\}$ of length N . For image encoding, we use features extracted from a pre-trained ResNet152 (He et al., 2016) model.

To encode the input paragraph, we deploy a bidirectional LSTM (BiLSTM). The outputs of the BiLSTM, denoted as $\mathbf{G} = \{\mathbf{g}_1, \dots, \mathbf{g}_M\}$, where $\mathbf{g}_i = \text{BiLSTM}(\mathbf{x}_i, \mathbf{g}_{i-1}) \forall i \in \{1, \dots, M\}$, is the encoded representation of the input paragraph. \mathbf{x}_i is the vector embedding of the word w_i^p .

We deploy a unidirectional LSTM for sequential decoding of the caption C that leverages both the encoded image and paragraph representations. The image embedding is provided as an input to the decoder LSTM at timestep $t = 1$. In all subsequent timesteps, the decoder input is the embedding \mathbf{y}_{t-1} of the previous token w_{t-1}^c . The decoder state at each timestep t is obtained as $\mathbf{h}_t = \text{LSTM}(\mathbf{y}_{t-1}, \mathbf{h}_{t-1})$. To incorporate contextual information from the input paragraph, we concatenate an attention-weighted sum of the encoder states, denoted as $\tilde{\mathbf{G}}_t$, to the current state \mathbf{h}_t .

At each decoder time step $t \in \{2, \dots, N\}$, the attention weights α^t over the encoder states depend on the current decoder state \mathbf{h}_t and the encoder states \mathbf{G} . Formally,

$$\tilde{\mathbf{G}}_t = \sum_{i=1}^M \alpha_i^t \mathbf{g}_i \quad (1)$$

$$\alpha_i^t = \frac{\mathbf{v}^\top (\mathbf{W}_g \mathbf{g}_i + \mathbf{W}_h \mathbf{h}_t + \mathbf{b})}{\sum_{i'=1}^M \mathbf{v}^\top (\mathbf{W}_g \mathbf{g}_{i'} + \mathbf{W}_h \mathbf{h}_t + \mathbf{b})} \quad (2)$$

Finally, we pass the concatenated output through two dense layers with a non-linear activation layer (e.g. ReLU) placed in between. The output logits are then passed through a Softmax function to obtain the output distribution $p(\cdot)$ over the vocabulary. We train our model end-to-end by minimizing the negative log-likelihood, i.e., $\theta^* = \text{argmin}_\theta -\log p(C | I, P; \theta)$. Note that we obtain the input embeddings, \mathbf{x}_i , and \mathbf{y}_t , of the encoder and decoder, respectively, from the embedding layer of a pretrained BERT_{BASE} model.

The model’s objective is to learn the optimal parameters θ^* to maximize the log-likelihood $\log p(C|I, P; \theta)$. Therefore, we train our model end-to-end by minimizing the negative log-likelihood defined as:

$$\mathcal{L}(\theta) = \sum_{t=1}^N -\log p(w_t^c | w_1^c, \dots, w_{t-1}^c, I, P; \theta) \quad (3)$$

5 Experiments and Results

5.1 Quantitative Evaluation

Metrics. We use the MSCOCO (Lin et al., 2014) automatic caption evaluation tool¹ to quantitatively evaluate our proposed model variants using the BLEU-1, ROUGE-L, CIDEr, and SPICE

¹<https://github.com/tylin/coco-caption>

| | BLEU-1 | ROUGE-L | CIDEr | SPICE | SemSim |
|-------------|--------------|-------------|-------------|-------------|-------------|
| +NE | | | | | |
| Image-only | 9.72 | 8.42 | 0.42 | 0.18 | 0.72 |
| Text-only | 8.71 | 7.85 | 0.68 | 0.29 | 0.73 |
| Contextual | 7.94 | 7.82 | 0.50 | 0.17 | 0.71 |
| +NE-overlap | | | | | |
| Image-only | 8.64 | 7.84 | 0.50 | 0.19 | 0.73 |
| Text-only | 8.34 | 7.48 | 0.53 | 0.20 | 0.73 |
| Contextual | 10.13 | 9.57 | 0.84 | 0.31 | 0.75 |
| NE | | | | | |
| Image-only | 5.96 | 6.42 | 0.37 | 0.14 | 0.71 |
| Text-only | 5.36 | 5.29 | 0.30 | 0.16 | 0.68 |
| Contextual | 6.37 | 6.93 | 0.45 | 0.19 | 0.72 |
| NE-overlap | | | | | |
| Image-only | 7.80 | 7.50 | 0.38 | 0.16 | 0.76 |
| Text-only | 6.87 | 6.54 | 0.61 | 0.36 | 0.72 |
| Contextual | 9.30 | 9.68 | 0.78 | 0.50 | 0.77 |

Table 1: Quantitative Evaluation of baselines and Contextual Captioning on different data splits

metrics. In addition, we also report scores for semantic similarity between ground truth (c_{gt}) and generated (c_{gen}) captions: $SemSim(c_{gt}, c_{gen}) = \cos(\mathbf{v}_{c_{gt}}, \mathbf{v}_{c_{gen}})$, where $\mathbf{v}_{c_{gt}}$ and $\mathbf{v}_{c_{gen}}$ are the mean vectors of constituent words in the respective captions from 300-dimensional GloVe embeddings.

Baselines. To the best of our knowledge, there is no existing work that studies contextual image captioning. Therefore, we present two baselines that can also be regarded as ablated versions of our model: *Image-only* and *Text-only* captioning.

Results. In Table 1, we report scores² for the baselines and our model variants. Recall from Section 3 that our models are based on various data splits: +NE, NE, and their respective *overlap* variants. We observe that for the +NE split, contextual captions are not better than the unimodal baselines on n-gram overlap based scores. This can be attributed to the nature of the dataset: NEs in the paragraph differ from those in ground truth captions. Since contextual captions draw inference from the paragraphs, the predicted NEs differ from ground truth captions as well, leading to lower scores for n-gram overlap based metrics. For the NE splits as well as both the *overlap* splits, contextual captions fare better than the baselines.

The observed low scores for BLEU-1, ROUGE-L, CIDEr and SPICE hint towards the subjectivity of the task. N-gram overlap based metrics do not

²The BLEU-1 and ROUGE-L scores are multiplied by 100, and CIDEr and SPICE scores are multiplied by 10 following the standard practice.



| Image |  |  |
|-----------------------|--|---|
| Paragraph | Made the hike to Franklin Falls and while waiting for some other people to clear my shot, I noticed how good the light looked hitting the rocks. | I was driving down the mountain... popped out my camera to snag this shot. It's beautiful right now... there wasn't nearly as much snow as last year. |
| Text-Contextual | <i>Rush hour on the nature coast.</i> | <i>I love the snow mountains. Come in the countryside often.</i> |
| Image-Contextual only | <i>The view from the top of the cosmopolitan.</i> | <i>The view from the top of the moon.</i> |
| Text-Image only | <i>Pretty cool sunset.</i> | <i>Rain ready for a local bar.</i> |

Figure 3: Linguistic richness of *Contextual Captions* in contrast to those generated from only image or only text.

accommodate varied interpretations and linguistic diversity. Figure 3 exemplifies how image-only captions for different images are often similar, while contextual captions are linguistically richer.

High average SemSim scores of contextual captions are indicative of their thematic similarity with the ground truth. Note that the splits with enforced similarity (*-overlap*) between paragraph and caption fare better on SemSim, leading to the conjecture that with a cleaner dataset, it would be possible to generate very relevant contextual captions.

5.2 Qualitative Evaluation

Setup. The scope of this evaluation is to analyze the different splits of our dataset. A user study was set up on the crowd-sourcing platform Appen³. 250 test samples were studied. For each sample, users were shown the image and its associated paragraph, and were asked to rate 6 captions (4 contextual captions and 2 baselines from Table 1) on a scale from 1 (irrelevant) to 5 (highly relevant).

Observations. We observe that for 80% of samples (201 out of 250), at least one of the 4 contextual captioning models is rated strictly higher than both baselines, and for 95% of samples they are at least as good as both baselines. A variant-wise analysis of this is shown in Table 2.

In 75% of samples, contextual captions were rated highest among all 6 captions. The variant-wise analysis of the same is shown in Table 3.

We identify three categories of samples:

³<https://appen.com>, formerly named Figure8.

Table 2: Percentage of samples where contextual captions are rated as good as or better than baselines.

| | Image-only | | Text-only | |
|-----------------------------------|-------------|-------------|-------------|-------------|
| | \geq | $>$ | \geq | $>$ |
| +NE | 71.6 | 42.4 | 74.4 | 38.4 |
| +NE-overlap | 69.6 | 42.8 | 74.0 | 44.4 |
| $\mathcal{N}\mathcal{E}$ | 70.0 | 45.2 | 73.6 | 41.2 |
| $\mathcal{N}\mathcal{E}$ -overlap | 76.0 | 48.4 | 81.2 | 49.6 |

Table 3: Percentage of samples rated highest per model.

| +NE | +NE-overlap | $\mathcal{N}\mathcal{E}$ | $\mathcal{N}\mathcal{E}$ -overlap | Image-only | Text-only |
|------|-------------|--------------------------|-----------------------------------|------------|-----------|
| 17.8 | 15.8 | 15.9 | 25.7 | 15.0 | 9.9 |

- *Significant*: samples where at least one of the 6 variants generate a caption with rating ≥ 3 . These constitute 46% of samples (115/250).
- *Insignificant*: samples on which all 6 variants obtain a rating < 3 . Here, paragraphs show substantial randomness and offer little context for the image. It appears impossible to generate good contextual captions for such samples.
- *Bad-base*: samples which are insignificant (rating < 3) with respect to both baselines. These constitute 80% of samples (201/250).

For 86% of *Significant* samples (99/115), contextual captions were rated higher than the baselines. A detailed analysis is given in Table 4.

Table 4: Percentage of *Significant* samples where contextual captions are rated as good as or better than baselines.

| | Image-only | | Text-only | |
|-----------------------------------|-------------|-------------|-------------|-------------|
| | \geq | $>$ | \geq | $>$ |
| +NE | 66.1 | 55.7 | 67.8 | 47.0 |
| +NE-overlap | 64.4 | 53.9 | 69.6 | 54.0 |
| $\mathcal{N}\mathcal{E}$ | 60.9 | 48.7 | 64.4 | 43.5 |
| $\mathcal{N}\mathcal{E}$ -overlap | 72.2 | 59.1 | 78.3 | 58.3 |

The ratings of 33% (67 of 201) of *Bad-base* samples were made significant, i.e., improved to strictly ≥ 3 , by the best contextual captioning model variant. In other words, contextual captioning generates superior captions for samples with inferior baseline captions.

$\mathcal{N}\mathcal{E}$ -overlap emerged as the best-suited contextual captioning variant in both quantitative and qualitative evaluations.

Factorial Experiment. We conduct another study taking the form of a $2 \times 2 \times 2$ full factorial experiment based on three factors – presence of NEs, caption-paragraph overlap, and use of pre-trained BERT token embeddings. We study the effect of these factors with a user study with all factor combinations. The effect of each of the factors can be seen in Figure 4. Using BERT token embeddings is by far the most effective in enhancing caption qual-

ity. It is interesting to note that presence of NEs (including its interaction with other factors) has a negative effect – captions without NEs are rated higher by human evaluators. Caption-paragraph overlap splits are also rated higher, which indicates that high inter-modality content overlap is necessary for generating good contextual captions.

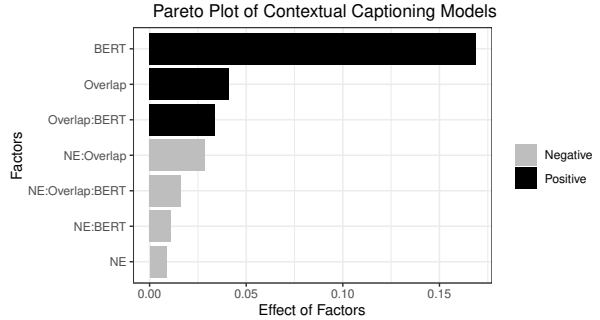


Figure 4: Effect of various factors in Contextual Captioning.

5.3 Discussion

Named Entities in Captions. Our user study shows that contextual captions with named entities (NE) are less preferred by humans. We conjecture that a lack of strong cues from the paragraphs lead to incorrectly generated NEs. Future work should also explore copy mechanisms to copy NEs from paragraphs to captions.

Caption Quality. We observe that the baseline captions do not show linguistic diversity (Figure 3). “The view from...”, “My friend and I...” etc. are common templates learned by the models. We conjecture that training the model on samples containing coherent paragraphs that have high content overlap with the image would yield nicer captions. We partially emulate this in our *-overlap* splits, which indeed show better model performance.

6 Conclusion

We propose the novel task of *Contextual Image Captioning* that exploits complementary knowledge from different modalities in multimodal contents. To facilitate a thorough study of this task, we curate a new dataset comprising $\sim 250,000$ multimodal Reddit posts. We provide an analysis of the dataset along with experimental results to identify interesting factors that determine the quality of generated contextual captions. We hope that our work will kindle and support follow-up research on this under-explored task, with downstream applications such as content authoring tools and multimodal conversational agents.

References

- Ali Furkan Biten, Lluís Gómez, Marçal Rusiñol, and Dimosthenis Karatzas. 2019. acute; good news, everyone! context driven entity-aware captioning for news images. In *CVPR*.
- Arjun Chandrasekaran, Devi Parikh, and Mohit Bansal. 2018. Punny captions: Witty wordplay in image descriptions. In *NAACL-HLT*.
- Jingqiang Chen and Hai Zhuge. 2018. Abstractive text-image summarization using multi-modal attentional hierarchical RNN. In *EMNLP*.
- Jingqiang Chen and Hai Zhuge. 2019a. Extractive summarization of documents with images based on multi-modal RNN. *Future Gener. Comput. Syst.*
- Jingqiang Chen and Hai Zhuge. 2019b. News image captioning based on text summarization using image as query. In *SKG*.
- Wei-Ta Chu and Ming-Chih Kao. 2017. Blog article summarization with image-text alignment techniques. In *ISM*.
- Ali Farhadi, Seyyed Mohammad Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David A. Forsyth. 2010. Every picture tells a story: Generating sentences from images. In *ECCV*.
- Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. 2017. Stylenet: Generating attractive visual captions with styles. In *CVPR*.
- Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O. K. Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *ACL*.
- K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep residual learning for image recognition. In *CVPR*.
- Jack Hessel, Lillian Lee, and David Mimno. 2019. Unsupervised discovery of multimodal links in multi-image, multi-sentence documents. In *EMNLP/IJCNLP*.
- MD Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. 2019. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CSUR)*, 51(6):118.
- Andrej Karpathy and Fei-Fei Li. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*.
- Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. 2017. A hierarchical approach for generating descriptive image paragraphs. In *CVPR*.
- Iro Laina, Christian Rupprecht, and Nassir Navab. 2019. Towards unsupervised image captioning with shared multimodal embeddings. In *ICCV*.
- Hui Lin and Vincent Ng. 2019. Abstractive summarization: A survey of the state of the art. In *AAAI*.
- 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 *ECCV*.
- Bei Liu, Jianlong Fu, Makoto P. Kato, and Masatoshi Yoshikawa. 2018. Beyond narrative description: Generating poetry from images by multi-adversarial training. In *ACM Multimedia*.
- Kazuma Muraio, Ken Kobayashi, Hayato Kobayashi, Taichi Yatsuka, Takeshi Masuyama, Tatsuru Higurashi, and Yoshimune Tabuchi. 2019. A case study on neural headline generation for editing support. In *NAACL-HLT*.
- Cesc Chunseong Park, Byeongchang Kim, and Gunhee Kim. 2017. Attend to you: Personalized image captioning with context sequence memory networks. In *CVPR*.
- Thomas Scialom, Patrick Bordes, Paul-Alexis Dray, Jacopo Staiano, and Patrick Gallinari. 2020. BERT can see out of the box: On the cross-modal transferability of text representations. *CoRR*.
- Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *ACL*.
- Kurt Shuster, Samuel Humeau, Hexiang Hu, Antoine Bordes, and Jason Weston. 2019. Engaging image captioning via personality. In *CVPR*.
- Alasdair Tran, Alexander Mathews, and Lexing Xie. 2020. Transform and Tell: Entity-Aware News Image Captioning. In *CVPR*.
- Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *CVPR*.
- Haoran Wang, Yue Zhang, and Xiaosheng Yu. 2020. An overview of image caption generation methods. *Comp. Int. and Neurosc.*, 2020:3062706:1–3062706:13.
- Qi Wu, Chunhua Shen, Peng Wang, Anthony R. Dick, and Anton van den Hengel. 2018. Image captioning and visual question answering based on attributes and external knowledge. *IEEE Trans. Pattern Anal. Mach. Intell.*
- Qiaolin Xia, Haoyang Huang, Nan Duan, Dongdong Zhang, Lei Ji, Zhifang Sui, Edward Cui, Taroon Bharti, Xin Liu, and Ming Zhou. 2020. XGPT: cross-modal generative pre-training for image captioning. *CoRR*.
- Yimin Zhou, Yiwei Sun, and Vasant G. Honavar. 2019. Improving image captioning by leveraging knowledge graphs. In *WACV*.
- Junnan Zhu, Haoran Li, Tianshang Liu, Yu Zhou, Jijun Zhang, and Chengqing Zong. 2018. MSMO: multimodal summarization with multimodal output. In *EMNLP*.

A Appendix

A.1 Dataset Details

- Total number of samples: 242,767
- Samples with named entities (NE) in caption: 137,732 (56.82%)
- Samples with no NE in caption: 104,653 (43.18%)

We ensure a context overlap between paragraph and caption with the following splits:

- *+NE* samples with one common word between paragraph and caption: 50,730 (20.93%). These are named *+NE-overlap* in Table 1.
- *+NE* samples with two common words between paragraph and caption: 23,283 (9.61%).
- *NE* samples with one common word between paragraph and caption: 38,301 (15.80%). These are named *NE-overlap* in Table 1.
- *NE* samples with two common words between paragraph and caption: 15,070 (6.22%)

We use SpaCy to detect named entities in captions. SpaCy detects 18 kinds of named entities⁴. TIME, MONEY, PERCENT, and LANGUAGE were not considered since they include common conversational phrases like “day before yesterday”, “my two cents”, “an English breakfast” etc. Examples of captions with NEs: “Just the (Earth/LOC) letting off some steam (Iceland/GPE)”, “The (first/CARDINAL) Chipotle , opened in (Denver/GPE) in (1993/DATE).” Examples of captions without NEs: “Texture of the paint on a skull I painted.”, “My girlfriend and I handle social situations differently.”

In future work, the NE types could be leveraged to learn positional relationships in sentences.

A.2 Experimental Setup

Our architecture is developed in PyTorch. The number of samples in all train/val/test splits is 30,000/8000/8000. Each model is trained for 20 epochs, with a batch size of 16. On a Tesla V100-PCIE-16 GB GPU, training 1 epoch taken 8 min.

⁴<https://spacy.io/api/annotation#named-entities>

For each model variant, the best validation model is used for testing. We experiment with models using pre-trained BERT token embeddings, as well as learning token embeddings from scratch (with a vocabulary size of 100,000). We observe that BERT token embeddings have a positive effect on the quality of captions (Figure 4), and hence consider this configuration as default.

A.3 Quantitative Evaluation

Validation Performances of Test Models. We train each model for 20 epochs and chose the best validation model for testing. In Table 5 we report the validation losses of our reported test models.

Table 5: Validation loss of the reported test models.

| Models | +NE | +NE-overlap | <i>NE</i> | <i>NE</i> -overlap |
|------------|------|-------------|-----------|--------------------|
| Image-only | 0.89 | 0.70 | 1.41 | 1.05 |
| Text-only | 1.39 | 1.46 | 1.38 | 1.27 |
| Contextual | 1.29 | 1.28 | 1.14 | 1.13 |

A.4 Examples

Table 5 shows a few good and bad examples of contextual captions.





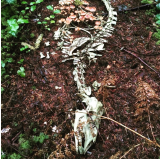


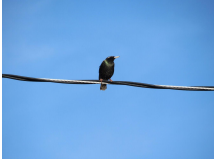
| | | | | |
|---------------------|---|--|--|--|
| Image |  |  |  |  |
| Paragraph | Shes pretty. Sorry for your loss. If only dogs had a longer lifespan than humans! If only we picked the one that would bury us. . . . That aside, I see you made each others life wonderful. | Thats a dope shot, nice. Thats cool Looks like the Rouge River. Arent those tunnels better known as bridges ? Its not really a bridge its not used for people to get across. | Lenticular clouds are cool. . . Super cool! Looks like a giant tree. . . These are called lenticular clouds. This must be the most enticing pic of a mountain I ever saw. | I hope it has LED lights too for full 90s effect. . . Happy new year bro! That view is majestic. Doesn't look like that thing even plays CDs. You're right. |
| Contextual Captions | <i>My only friend passed away last year. He passed away from cancer.</i> | <i>A beautiful stream I encountered on holiday.</i> | <i>Some of these clouds are having fun as we get out of water.</i> | <i>I took this picture of my friends living room in the mountains where the view is breathtaking.</i> |
| Image-only Captions | <i>My friend is a new friend. My son's dog.</i> | <i>A picture I took at a morning in my hometown.</i> | <i>A picture I took in my morning. The view from my hotel.</i> | <i>A picture i took in my local park.</i> |
| Text-Only Captions | <i>Rest easy, hear surgery. Cancer, and essa. He died last year.</i> | <i>Milky way over the clouds worth it however.</i> | <i>A collage from the top of a cliff bench at midnight.</i> | <i>Usually a replica of the most beautiful cities in the world before going to Inverness, France</i> |
| Image |  |  |  |  |
| Paragraph | You absolute legend you. Thanks so much! I saw it and thought the same thing! I tried r/skeletons just gave me spooky bone memes. I'll check fossils though! One of those students are headed to krypton. | there is nothing NSFW about this Beijing agent says what. . . US says HK police can handle this. . . HK ppl says please stop violence. . . I say i am just a HK citizen. . . Cop used to be an adventurer. | The Jefferson Memorial will always be my favorite DC monument. I've lived in the area for fifteen years and *finally* saw this beauty for myself in person. It was absolutely worth the hellish traffic! | Bird on a Wire. I saw it once on a Leonard Cohen album. . . Definitely a starling. The Counting Crow Power Line Black Bird. Put a bird on it Definitely a European Starling, it's the most common bird in the world. |
| Contextual Captions | <i>I drew this and I thought it looked cool. I want to be a good artist.</i> | <i>This is a real unedited picture ever taken.</i> | <i>A beautiful sunset ship swallowed sea lion waterfalls.</i> | <i>This bird looks like a warrior bird dead bird bird, bird bird dead</i> |
| Image-only Captions | <i>My first time on a year ago, and I just got a photo of this photo.</i> | <i>My first ever attempt at a photo of a year ago.</i> | <i>The view from my house in the morning.</i> | <i>My friend drew this in 1995 today, the delicious</i> |
| Text-Only Captions | <i>Danny Devito ink drawing lights up optical.</i> | <i>Hong Kong protesters fired a letter to the protest police threatening by police. Pepper spray</i> | <i>Tragedy in Pittsburgh, pa 18 years final night the highway.</i> | <i>Hummingbird from the moon landing. I think this is beautiful.</i> |

Figure 5: Good examples (top row) and bad examples (bottom row) of *Contextual Captions*.