Focus! Relevant and Sufficient Context Selection for News Image Captioning

Mingyang Zhou[1]   **Grace Luo**[3]   **Anna Rohrbach**[3]   **Zhou Yu**[2]

[1]University of California, Davis   [2]Columbia University

[3]University of California, Berkeley

minzhou@ucdavis.edu, zy2461@columbia.edu, {graceluo, anna.rohrbach}@berkeley.edu

Abstract

News Image Captioning requires describing an image by leveraging additional context from a news article. Previous works only coarsely leverage the article to extract the necessary context, which makes it challenging for models to identify relevant events and named entities. In our paper, we first demonstrate that by combining more fine-grained context that captures the key named entities (obtained via an oracle) and the global context that summarizes the news, we can dramatically improve the model's ability to generate accurate news captions. This begs the question, how to automatically extract such key entities from an image? We propose to use the pre-trained vision and language retrieval model CLIP to localize the visually grounded entities in the news article and then capture the non-visual entities via an open relation extraction model. Our experiments demonstrate that by simply selecting a better context from the article, we can significantly improve the performance of existing models and achieve new state-of-the-art performance on multiple benchmarks.

1   Introduction

News Image Captioning is an extension of the standard image captioning task, where, besides the image, one needs to leverage longer context in the form of the news article. This is important in order to capture the global news story context to properly discuss the image. At the same time, the specific entities (e.g. people) in the image are also often referenced in the article.

Coincidentally, generating the named entities (names, locations, dates, etc), is one of the most critical challenges in news image captioning. We analyze the captions and the corresponding articles on three popular benchmarks, GoodNews (Biten et al., 2019), NY800KTimes (Tran et al., 2020), VisualNews (Liu et al., 2020), and find that 58.7% – 74.5% of the named entities present in captions

also appear in the articles. This shows that the key information of a news caption, namely the named entities, can often be directly derived from the associated news article. However, due to the length and abundance of other information in the full articles, it is challenging to uncover the key entities. Inspired by this, we explore how to select relevant yet sufficient context from the news article to assist with the news image caption generation.

To analyze what the right type of context is, we start with an oracle-based study where we demonstrate that combining key local context and global context leads to the best result. Next, we move to designing a multimodal retrieval method to handle automatic extraction of key local context (key named entities). Our method consists of two stages, first we retrieve the visually grounded entities, then we discover the non-visual ones via open relation extraction. Figure 1 illustrates our approach. Our full method outperforms the vanilla baseline on three benchmarks, and achieves the new state-of-the-art results on two of them.

2   Related Work

Automatic image captioning has achieved tremendous success over the past few years, where many methods(Vinyals et al., 2015; Johnson et al., 2016; Karpathy and Li, 2015) are developed to accurately describe the visual objects and their relationships in the image. Recently, news image captioning has started to get increased attention with the focus on producing narration of news images with richer human-like information derived from the news article. The research on news image captioning is inspired by several recently collected benchmark datasets including BreakingNews (Ramisa et al., 2018), GoodNews (Biten et al., 2019), NY800KTimes (Tran et al., 2020), and VisualNews (Liu et al., 2020). Earlier works such as Ramisa et al. (2018) and Biten et al. (2019) propose a two-stage pipeline, where they first generate tem-

Global Context

Serena Williams and Andy Murray in Wimbledon Mixed Doubles: It's Happening

WIMBLEDON, England — After days of copy remarks about the possibility, it has been confirmed that **Andy Murray** and **Serena Williams** will be a team in the mixed doubles event at **Wimbledon**.

Key Local Context

The two will sign in for the 48-team competition ahead of the 11 a.m. deadline on Wednesday morning, Matt Gentry, a spokesman for **Murray**, said on **Tuesday**.

Open relation extraction

Murray, 32, is playing his third tournament since hip replacement surgery in January, and he is focusing on doubles to ease into his comeback.

Visually grounded entities

Caption: **Andy Murray** practicing at the **All England Club** on **Tuesday**. He and **Serena Williams** have agreed to play together in **Wimbledon's** mixed doubles.
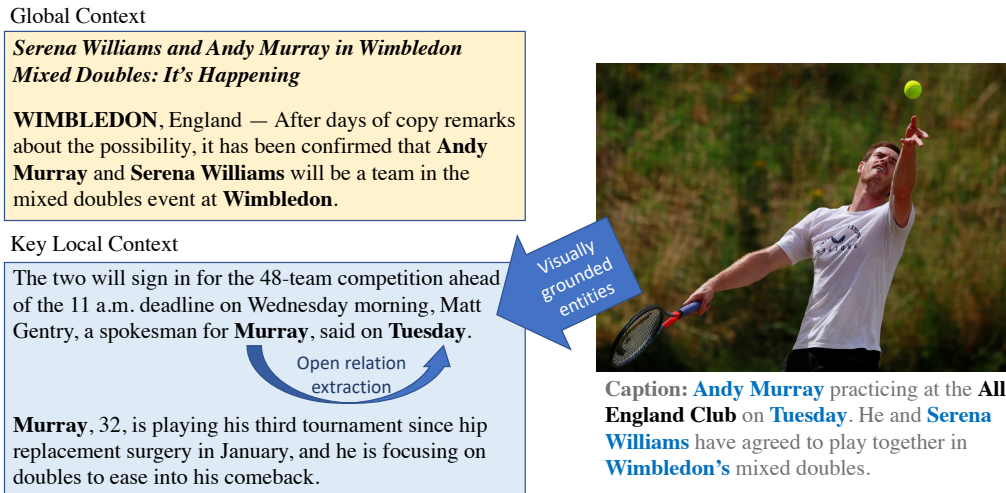
Figure 1: We propose to use a combination of key local context and global context extracted from the news article. Both contain useful complementary information. While automatically extracting the key local context, we (a) detect visually grounded entities (e.g. Murray) and (b) discover non-visual entities via open relation extraction (e.g. Tuesday). Still, some caption entities (e.g. All England Club) may be absent in the article.

plate captions and then insert named entities into corresponding positions. Hu et al. (2020) propose a hierarchical article encoding mechanism, where the model first retrieves the most relevant sentences and then attends to the appropriate words in the retrieved context with images involved in both steps. Later works, like Liu et al. (2020) and Tran et al. (2020) introduce end-to-end models that directly encode articles and images with pre-trained transformer architectures (BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019)), then generate the caption directly. The end-to-end method is better than the template-based approach as it models the rich contextual information in named entities using byte-pair encoding (BPE) (Sennrich et al., 2016). A recent work by Yang et al. (2021) proposes to follow journalistic guidelines when generating news image captions, i.e. each caption must contain the *who*, *when*, *where*, and *how* entities.

There is rarely any effort spent on context selection to help news caption generation in previous work. The LSTM-based method such as Ramisa et al. (2018) and Biten et al. (2019) coarsely encode the full article either as a general embedding vector or a set of sentence embedding vectors. Later, the transformer-based approach such as Tran et al. (2020) and Liu et al. (2020) directly encode the article at the token-level and achieved much better performance than the LSTM-based approach. However, they usually need to truncate the article to meet the encoding length limitation of the trans-

former architecture, which inevitably leads to the loss of useful information. Hu et al. (2020) is one of few works that addresses context selection for news image caption, where they train a cross-modal retrieval model VSE++ (Faghri et al., 2018) to find the sentences from the article that are most relevant to the image. However, the retrieved relevant sentences from VSE++ are limited to only capture visually grounded information. We argue that other valuable information, i.e. non-visual named entities, should be included as the relevant context to help news image caption generation.

3 Oracle Study: Optimal News Context Selection

We posit that there are two critical types of context in the news article that we need to extract to help the model generate accurate news image captions: (1) **The key local context**: a snippet of the news article that contains the named entities mentioned in the caption. (2) **The global context** that summarizes the news story. We propose a strategy to extract these two types of context using an oracle (ground-truth caption) and validate their contribution by training a model using them as input. We compare against the models trained on the news article with alternative schemes of context selection. Next, we introduce the details of the datasets, captioning model, and evaluation metrics that we use for our study to validate our hypothesis on optimal news context selection.

Context	BLEU-4	ROUGE	CIDER	Named Entities	
				P	R
Original	6.0	21.4	53.8	22.2	18.7
Oracle Key Local (Par)	6.9	23.0	62.7	29.1	23.8
Oracle Key Local (Sent)	7.1	23.5	66.4	**31.9**	**25.6**
Oracle Key Local (Sent) + Global	**7.2**	**24.2**	**67.4**	30.0	24.5

Table 1: Evaluation results on GoodNews with different strategies for selecting the article context. The Original context here means the first 500 words from the article following (Tran et al., 2020).

Context	BLEU-4	ROUGE	CIDER	Named Entities	
				P	R
Original	6.3	21.7	54.4	24.6	22.2
Oracle Key Local (Par)	9.3	26.4	81.4	38.7	34.7
Oracle Key Local (Sent)	9.7	27.1	83.9	**41.9**	**37.2**
Oracle Key Local (Sent) + Global	**10.3**	**27.6**	**84.5**	39.8	35.5

Table 2: Evaluation results on NYTimes800K with different strategies for selecting the article context. The Original context here means the nearest 512 tokens surrounding the image following (Tran et al., 2020).

3.1 Datasets

We use two well-known news image captioning datasets: GoodNews (Biten et al., 2019) and NY-Times800k (Tran et al., 2020). We follow the data split setting in Tran et al. (2020) with 421K training, 18K validation, and 23K testing for the GoodNews and 763K training, 8K validation, and 22K test for the NYTimes800k.

3.2 News Captioning Model

We conduct our experiment with *Transform and Tell* (Tran et al., 2020) (*Tell* for short), which is a state-of-the-art transformer-based captioning model. It encodes the article with a pre-trained RoBERTa model (Liu et al., 2019) and extracts the representation of an image via a set of visual encoders including ResNet152 (He et al., 2016), YOLOv3 object detector (Redmon and Farhadi, 2018), and MTCNN face detector (Zhang et al., 2016) to represent different visual concepts. Then the extracted multi-modal feature is passed to a transformer decoder to generate the caption.

Due to the maximum length limitation of encoded text handled by RoBERTa, the full article is truncated before feeding into the language encoder. Tran et al. (2020) apply two different strategies to truncate the article. For GoodNews, they simply choose the first 500 words. For NYTimes800k, as the image position in the original news article is available, they select the 512 tokens "surrounding" the image, which may provide the model with more image-relevant context.

3.3 Evaluation Metrics

Following previous literature (Tran et al., 2020; Liu et al., 2020; Yang et al., 2021), we use BLEU-4 (Papineni et al., 2002), ROUGE (Lin, 2004), and CIDEr scores (Vedantam et al., 2015) to measure the similarity between the generated caption and the referenced ground-truth captions. Among the three metrics, CIDEr is the most suitable one for news image captioning evaluation as it focuses more on unusual words generation. In addition, we evaluate the precision and recall on named entities to verify whether the key information is covered in the generated caption. Specifically, we check the exact string matches between the named entities detected in both the ground truth captions and the generated captions using SpaCy [1].

3.4 Oracle Optimal Context Selection

We construct the key local context for a given image, assuming access to its ground truth caption.

Given a set of named entities present in the caption, we define the key local context as the set of sentences (or paragraphs) that contain these named entities The selected sentences (paragraphs) are concatenated following the original order in the news article. We specifically compare the key information at both the sentence-level and paragraph-level to measure the impact of the different degrees

---

[1]https://spacy.io/

of condensation. For global context, we concatenate the article title and the first paragraph to summarize the high-level news story. When we combine the two types of context, the global context is always put before the local key information context, following the presentation in the original paragraph. The content in the local context that overlaps with the global context will be omitted. We cap the length of the combination of the global context and the local key information context to 500 words to stay within the maximum sequence length of the transformer architecture.

3.5 Results and Discussion

The results of training Tell with different contexts on GoodNews and NYTimes800k are summarized in Tables 1 and 2. The Original context corresponds to the respective scheme used by Tran et al. (2020) on each dataset (as described above). It is clear that when we feed the model more relevant context that contains the key information, the performance of the model is dramatically improved on all the reference-based metrics. Meanwhile, the precision and recall of the named entities in the generated caption are significantly increased by over 5% in both GoodNews and NYTimes800k. We also observe that when the key information is condensed to an even more concise level (from paragraph to sentence), the performance of the model can be further improved. This indicates that the model can learn more effectively from the condensed context that focuses on the critical information.

When we combine the global context and the key local context, we find that we can achieve an improvement on the reference metrics but observe a drop on named entity accuracy in the generated captions. Adding global context inevitably adds more words which leads to the introduction of potentially irrelevant entities. However, it provides a valuable source of information to create captions that cover not only the named entities found in the image but also convey the key topic of the article. Therefore, we deem the combination of the global and local context as sufficient and relevant knowledge for the news captioning model.

4 Automatic Key Local Context Detection

In the previous section, we identified an optimal strategy for selecting news context, which relied on the named entities found in the ground truth captions. In practice, automatically determining

the key named entities remains an open challenge. To address it, we propose a simple multi-modal retrieval pipeline to extract the key named entities from the news article. The pipeline consists of two stages: 1. we identify visually grounded named entities, such as people's names and geographic locations, via cross-modal retrieval; 2. we discover the key named entities that are not visually grounded, e.g. time, by exploring factual relations between the entities in the article and the detected visually grounded entities.

spaCy Type	Description	Components
PERSON	People, including fictional	WHO
NORP	Political groups	WHO
ORG	Companies, agencies, etc	WHO
DATE	Dates or periods	WHEN
TIME	Times smaller than a day	WHEN
FAC	Buildings, airports, highways	WHERE
GPE	Countries, cities, states	WHERE
LOC	Locations, mountains, waters	WHERE
PRODUCT	Objects, vehicles, foods	MISC
EVENT	Named wars, sports events	MISC
ART	Titles of books, songs	MISC
LAW	Laws	MISC
LAN	Any named language	MISC
PERCENT	Percentage, including "%"	MISC
MONEY	Monetary values	MISC
QUANTITY	Measurements	MISC
ORDINAL	"first", "second", etc	MISC
CARDINAL	Numerals	MISC

Table 3: Mapping from spaCy named entities types to the four named entities components defined in (Yang et al., 2021).

Visually Grounded Named Entities Retrieval According to Yang et al. (2021), the named entities in the news captions can be generally divided into four groups: WHO, WHEN, WHERE, and MISC. The mapping between the SpaCy named entity types and each category can be found in Table 3. We define named entities belonging to WHO and WHERE as visually grounded named entities, as they are often related to specific image regions. For example, people's names can be associated to the faces and names of cities can be inferred from the landmarks shown in the image. It is also evident by the high ratio of appearance of these two entity types in the news captions (Yang et al., 2021). WHO appears in more than 93% and WHERE appears in more than 50% of the news captions in GoodNews and NYTimes800k.

To localize the visually grounded named entities in the news article, we first find the relevant sentences from the entire news article that are se-

mantically close to the image. Our approach is to leverage a large pre-trained cross-modal retrieval model CLIP (Radford et al., 2021) trained on over 400 million image-text pairs from the Internet. The sentence relevance is measured by the cosine similarity with the image in the learned embedding space of CLIP. As there is a large domain discrepancy between news sentences and the training data of CLIP, we first need to fine-tune CLIP on the news image captioning dataset. The CLIP model is fine-tuned with a contrastive loss to distinguish the positive image-sentence pairs from the negative pairs. For the positive sentences we use the news image captions since there are no labels for which news article sentences are relevant to each image, and the writing style of the captions is close to that of the article sentences. Besides using captions describing other images as negative samples, we also define hard negatives as news article sentences that do not cover any non-stop words in the news caption. During image captioning, we use the trained CLIP model to rank news sentences against the image and pick the top 2 as the most relevant sentences. The visually grounded named entities are then retrieved as the WHO and WHERE entities contained in the retrieved relevant sentences. If the top 2 retrieved sentences do not contain any visually grounded entity, we keep searching the rest of the retrieved sentences descending by similarity score to the news image until at least a visually grounded entity is captured.

Non-visually Grounded Named Entities Retrieval We define the named entities belonging to the rest of the categories: WHEN and MISC as non-visually grounded name entities. Such entities are not semantically related to the image and are more challenging to localize via cross-modal retrieval. However, they are often related to the visually grounded named entities in the news text. For example, in the sentence *Ali Kashani-Rafye, started selling bastani (Persian for ice cream) in 1980 at his grocery.*, the TIME entity *1980* is associated with the person's name *Ali Kashani-Rafye* with the action *selling bastani*. The task of extracting such relations between named entities is known as open relation extraction. We leverage an open relation extraction method OpenNRE (Han et al., 2019) to detect all the relations between each pair of named entities in every sentence of the news. We filter out the relations with a detection confidence score lower than 0.7. After filtering, if there is a relation

extracted between a non-visually grounded named entity and a detected visually grounded named entity from the cross-modal retrieval, we use such non-visually grounded named entities during final context extraction.

News Context Selection Once we get the full list of the detected named entities, we follow the optimal strategy derived in Sec 3 to select the news context. We combine the local-focused context guided by the detected named entities and the global context that is consisted of the title and the first paragraph to format the relevant news context.

5 Experiment

In this section, we demonstrate the impact of context section for news image captioning in the automatic setting. We first introduce the baselines and the evaluation datasets. Then we discuss our findings from the experimental results.

5.1 Datasets

We evaluate the performance of different methods on three well-known news image caption datasets, including GoodNews (Biten et al., 2019), NYTimes800k (Tran et al., 2020), and VisualNews (Liu et al., 2020). The details of GoodNews and NYTimes800k are introduced in Section 3. Unlike the other two datasets that are only collected from the New York Times, VisualNews is sourced from multiple news agencies, including The Guardian, BBC, USA Today, and The Washington Post. We follow (Liu et al., 2020) to split the data into 400K for training, 40k for validation, and 40k for testing.

5.2 Compared Models

VisualNews Captioner   (Liu et al., 2020) is a transformer (Vaswani et al., 2017) based approach with several augmentations to enhance the named entity generation. They add a list of entities from the news article as additional input to guide the named entity generation in the captions. Meanwhile, they also propose to decode the out-of-vocabulary named entities as the entity type token instead of an unknown token. Then when an entity type token is decoded, they simply retrieve the most frequent named entity of the same type in the news article as the final generation.

JoGANIC   (Yang et al., 2021) proposes to generate news caption following journalistic principles

Model	BLEU-4	ROUGE	CIDER	Named Entities	
				P	R
VisualNews	6.1	21.6	55.4	22.9	19.3
JoGANIC	**6.8**	**23.0**	**61.2**	**26.9**	**22.1**
Tell (Original)	6.0	21.4	53.8	22.2	18.7
Tell (Ours)	6.3	22.4	60.3	24.2	20.9

Table 4: Comparison between Tell trained with our context selection and other SoTA methods on GoodNews.

Model	BLEU-4	ROUGE	CIDER	Named Entities	
				P	R
VisualNews	6.4	21.9	56.1	24.8	22.3
JoGANIC	6.8	22.8	59.4	28.6	24.5
Tell (Original)	6.3	21.7	54.4	24.6	22.2
Tell (Ours)	**7.0**	**22.9**	**63.6**	**29.8**	**25.9**

Table 5: Comparison between Tell trained with our context selection and other SoTA methods on NYTimes800k.

Model	BLEU-4	ROUGE	CIDER	Named Entities	
				P	R
VisualNews	5.3	17.9	50.5	19.7	17.6
Tell (Original)	9.6	22.8	83.8	23.7	19.2
Tell (Ours)	**11.6**	**25.0**	**107.6**	**26.2**	**21.2**

Table 6: Comparison between Tell trained with our context selection and other SoTA methods on VisualNews.

where the caption must contain components such as *who*, *when*, *where*, and *how*. They propose to first generate the most likely caption components and then use the component-specific decoding method to generate the named entities that follow the guidance of the predicted components.

Tell (Original) is the Transform and Tell model (Tran et al., 2020) trained on the truncated news articles. The truncation strategies for GoodNews and NYTimes800k were introduced in Section 3.1. For Visual News, following GoodNews, we also truncate the full article to 500 words.

Tell (Ours) is the Transform and Tell model (Tran et al., 2020) trained on the context from our proposed automatic selection approach as described in Sec 4.

Context	BLEU-4	CIDER	Named Entities	
			P	R
LSTM (Original)	2.0	13.9	10.7	7.1
LSTM (Ours)	**2.1**	**15.4**	**11.4**	**7.6**

Table 7: Evaluation of an LSTM-based Captioner with our proposed context on NY800K.

5.3 Results

The results for the three benchmarks are summarized in Tables 4, 5, and 6. We observe that the Tell model trained on our proposed automatically selected context consistently outperforms the Tell model trained on the original context on all metrics in all three datasets. Meanwhile, Tell trained on our selected context also achieves the new state-of-the-art performance on NYTimes800k and Visual News. This demonstrates that even when the key named entities are automatically retrieved from the article, our proposed context selection strategy still helps the model learn to generate captions with better quality. We believe the main reason for the better performance is that the selected context covers more key named entities than the original context.

We also notice that the Tell model trained with our selected context does not outperform JoGANIC on the GoodNews dataset. We think this is due to the lower coverage ratio of the caption-relevant named entities within the news articles compared to the other two datasets (58% vs. 70%). This seems to limit somewhat the benefit of focusing the key context on the named entities, although we still improve over Tell (Original). Note, that we do outperform JoGANIC on the NYTimes800k

Context	BLEU-4	ROUGE	CIDER	Named Entities P	R
Tell (Original)	6.0	21.8	57.8	22.5	19.1
Tell (Only Visual NE)	5.7	21.5	54.4	22.8	19.2
Tell (Ours)	**6.3**	**22.4**	**60.3**	**24.2**	**20.9**

Table 8: Evaluation of Tell trained on the context guided by different sets of detected named entities on GoodNews including Visually Grounded Named Entities (Visual NE) and all the extracted named entities (our full approach).

Context	BLEU-4	ROUGE	CIDER	Named Entities P	R
Tell (Original)	6.3	21.7	54.4	24.6	22.2
Tell (Only Visual NE)	6.8	21.8	59.4	28.4	23.2
Tell (Ours)	**7.0**	**22.9**	**63.6**	**29.8**	**25.9**

Table 9: Evaluation of Tell trained on the context guided by different sets of detected named entities on NYTimes800k including Visually Grounded Named Entities (Visual NE) and all the extracted named entities (our full approach).

Context	BLEU-4	ROUGE	CIDER	Named Entities P	R
Tell (Original)	9.6	22.8	83.8	23.7	19.2
Tell (Only Visual NE)	10.9	23.9	96.4	24.5	20.7
Tell (Ours)	**11.6**	**25.0**	**107.6**	**26.2**	**21.2**

Table 10: Evaluation of Tell trained on the context guided by different sets of detected named entities on Visual News, including Visually Grounded Named Entities (Visual NE) and all the extracted named entities (our full approach).

dataset.

We also evaluate the usefulness of our proposed context selection strategy on different news captioning methods other than the transformer-based model Tell. We train an LSTM based news image captioning model[2] on our selected context and the original context with NYTimes800k. The results are summarized in Table 7. It is clear that the LSTM captioner trained on the selected context is still consistently better, which demonstrates the generalization of our proposed context selection strategy to other news captioning approaches.

5.4 Ablation on the Detected Named Entities

We validate the contribution of the different named entities obtained by our proposed multi-modal retrieval pipeline. We compare using context guided only by the detected visually grounded named entities and that of using the full set of detected named entities. When we select the context with these sets of entities, we follow the optimal strategy introduced in Sec 3 to include both the key local

context and the global context. We perform the ablation study on all three datasets, and the results are summarized in Tables 8, 9, 10.

We observe that when we only extract the visually grounded named entities with the fine-tuned CLIP model, we can already select a better context to train the model than the original context on NYTimes800k and Visual News. However, the performance on GoodNews is still worse than that trained on the original context. This mixed effect can likely be explained by the significantly different average article length in the GoodNews vs. NYTimes800k and Visual News dataset. Since GoodNews often contains very short articles that can almost entirely be fit within the first 500 words, simply selecting context guided by the detected visually grounded named entities will lead to incomplete key information coverage. On the other hand, in NYTimes800k and Visual News, as the full articles contain much more than 500 words (on average above 900 words), when we truncate the context to 500 words models suffer from a great loss of useful information. Therefore, even when

---
[2]LSTM+GLOVE baseline in Transform and Tell (Tran et al., 2020)

Gt caption:
Brandon Bostian, the engineer involved in the 2015 Amtrak derailment in Philadelphia that left eight people dead and about 200 injured.

Generated Caption with Selected Context:
Brandon Bostian, a Philadelphia engineer who was driving a train that derailed in May 2015.

Generated Caption with Original Context:
The Amtrak train that derailed in Philadelphia killed eight people.

Second Judge Dismisses Criminal Charges in Philadelphia Train Wreck That Killed 8

For the second time in two years, a Pennsylvania judge has dismissed charges against an Amtrak engineer who was driving a speeding train that derailed in Philadelphia in 2015, killing eight people. The engineer, Brandon Bostian, 36, faced more than 200 charges in the May 2015 derailment, including one count of causing catastrophe, a felony, and several counts of involuntary manslaughter and reckless endangerment, both misdemeanors. Just before the derailment, Mr. Bostian had accelerated the train to 106 miles per hour as it entered a curved section of track with a 50 m.p.h. In dismissing the charges on Tuesday, Judge Barbara McDermott of the Philadelphia Court of Common Pleas sided with Mr. Bostian's lawyer in ruling that the mistakes he made did not constitute a crime. The Philadelphia district attorney declined to charge Mr. Bostian in May 2017, saying he did not believe there was enough evidence to prove Mr. Bostian consciously disregarded a "substantial and unjustifiable risk." The families filed the private complaint against Mr. Bostian in Philadelphia Municipal Court. In September 2017, a Philadelphia Municipal Court judge dismissed the charges, saying the episode appeared to be an accident and not the result of criminal negligence. Mr. McMonagle said Tuesday's dismissal came after he filed motions — citing new decisions in separate criminal recklessness cases — that supported the argument that Mr. Bostian should not be charged. In a 2016 report, the National Transportation Safety Board said Mr. Bostian accelerated the train after "he lost his situational awareness because his attention was diverted to an emergency situation with a nearby Southeastern Pennsylvania Transportation Authority (SEPTA) train that had made an emergency stop after being struck by a projectile," the report said. The board said Mr. Bostian accelerated because he thought it was at a different section of the route. Thomas R. Kline and Robert Mongeluzzi, lawyers representing victims' families, said in an emailed statement that they disagreed with Tuesday's ruling: "Our clients are hopeful that the ruling will be reversed on appeal, and that ultimately there will be public accountability of Mr. Bostian, whose recklessness caused the death of eight individuals and mayhem to the lives of hundreds of others."

(a)

Gt caption:
Heather Russinko, one of three named plaintiffs in a lawsuit against New Jersey's health department, dipping homemade cake pops in white chocolate.

Generated Caption with Selected Context:
Heather Russinko, a home baker, is one of three plaintiffs in a lawsuit against the state's Department of Health.

Generated Caption with Original Context:
Ms. Russinko's cake pop, left, and a cake pop, right.

Home Cooking for Profit? Sure, Just Not in New Jersey

With just a little white chocolate and some sprinkles, Heather Russinko can make a wedding gown in under seven minutes. Give her five minutes more, and she can dress a groom, too. Three buttons, a bow tie, and a tuxedo swell over a round white chest. "I want to be able to say, 'O.K., Jared, you can go to college. "There's this rogue law standing in my way and preventing me from earning an income," said Ms. Russinko, one of three named plaintiffs in a lawsuit against the state's Department of Health. New Jersey's sanitary code, like most states', is derived from federal food laws based on a 1906 act; these codes have long excluded home kitchens from the definition of retail food establishments. For almost a decade, the New Jersey Home Bakers Association, which has more than 350 members including Ms. Russinko, has been lobbying elected officials for the right to sell homemade goods. "I support the fact that they want to pursue this type of entrepreneurship, but it has to be done in a manner that is safe." In Texas, which passed its first such law in 2011, new legislation that takes effect in September will significantly loosen cottage food restrictions. She is the 43rd home baker to be approved — she can tell, because it's the number on her license — and she mostly makes cupcakes, taking home a few hundred dollars a month. If there were a health risk, they ask, why would they be allowed to donate food to bake sales? "If there are going to be, let's say, 20 home bakers that are created through this legislation and they're selling their product to the public, it's likely that if they didn't exist, those public persons would have gone to a bakery," he said. Martha Rabello, another named plaintiff in the New Jersey suit, said she tried to do business the legal way: She rented a commercial kitchen, but the cost proved prohibitive, and she closed down shop after just a few months. "If you can start from home, you have a better chance of this business surviving," she said.

(b)

Figure 2: Examples of caption generation. For each example, we display news image (top left block) and the corresponding captions (top right block) from the NYTimes800k test set. We compare the captions generated by the model trained on the original context and the model trained on our selected context to the ground truth caption. We also display the selected context used by our proposed strategy (bottom block). The key information in the generated captions, ground truth caption, and the selected context are highlighted with yellow bars.

we simply detect the visually grounded named entities, we achieve a higher named entity coverage ratio than that of the Original Context. When we include the non-visually named entities detected via OpenNRE, we achieve additional improvement and outperform the model trained with the original context for all benchmarks. This demonstrates that non-visually named entities do add essential information for the model in order to generate captions with better quality.

5.5 Qualitative Generation Results

We also show qualitative examples that demonstrate the effectiveness of our proposed context selection strategy. In Fig 2, we show that the generated caption from the model trained on our selected context has covered more accurate key information than that trained on the original context. On the top example, the engineer's name "*Brandon Bostian*" is captured by our multi-modal retrieval pipeline, and then our open relation extraction model successfully captures the relation between the TIME entity "*2015*" with the detected name. These key pieces of information are correctly captured by our model that uses selected context, while they are missed by the model that uses the original unfiltered context. Meanwhile, the LOCATION named entity "Philadelphia" is captured by the global context (the title) and is also generated by the model trained on our selected context. Similarly, on the bottom, the caption generated from the model trained on our selected context contains the correct person entity ("Heather Russinko") and key event ("one of three plaintiffs") discussed in the article.

6 Conclusion

Our paper explores news context selection to improve the performance of existing models on news caption generation. Our proposed strategy to select relevant context is two-fold: (1) Include the key local context that focuses on the key information in the news caption, namely the named entities. (2) Include the global context, such as titles and the first paragraph, that summarizes the news story. Both parts are validated in an oracle setting with a strong news captioning model, Transform-and-Tell (Tell) (Tran et al., 2020), across multiple benchmarks. We further study key local context selection in a practical (automatic) setting, where we first detect a set of visual named entities with a multi-modal retrieval pipeline, find related non-visual named entities via relation extraction, then use these named entities to pick relevant news article sentences. The performance of Tell is consistently improved using our proposed context selection, and it achieves the new state-of-the-art on two benchmarks. We hope that our findings will inspire future work to develop models that even more effectively leverage the relevant context in the article to improve caption generation quality. Meanwhile, our proposed method can also be potentially extended to select a relevant context from external sources to cover the key named entities missing from the original news article.

Test Set	BLEU-4	CIDER	Named Entities P	R
All	6.3	60.3	24.2	20.9
High NE Cover	**7.3**	**65.1**	**31.2**	**28.2**

Table 11: Evaluation of Tell with our proposed context on the full Good News test set vs. the subset where the article contains more than 70% of the ground truth caption's named entities.

7 Limitations, Ethics, and Broader Impacts

Our proposed context selection method performs a hard filtering strategy to keep just the context that focuses on the key named entities and short global context that summarizes the story, which inevitably might lead to losing some potentially useful information in the article. A more optimal strategy could be to develop a soft filtering technique where higher weights are assigned to the more relevant context and lower weights to the rest of the context during the encoding stage.

We also find that our approach is more useful when there is a high ratio of covered named entities from the caption in the article. We compare the performance of Tell using our proposed context on the full test set vs. the subset of GoodNews where more than 70% of caption-relevant named entities are mentioned in the article. The result is summarized in Table 11. We find that the performance on this subset is much better than on the whole test set, where on average only articles only contain 59% of the caption-relevant named entities.

Finally, since we leverage the large-scale pretrained CLIP (Radford et al., 2021) model, we might transfer any biases (including gender or racial biases) that CLIP has learned into our context selection process. We recommend exercising caution when adopting our approach in practice.

Acknowledgements

References

Ali Furkan Biten, Lluis Gomez, Marcal Rusinol, and Dimosthenis Karatzas. 2019. Good news, everyone! context driven entity-aware captioning for news images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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. Association for Computational Linguistics.

Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. Vse++: Improving visual-semantic embeddings with hard negatives.

Xu Han, Tianyu Gao, Yuan Yao, Deming Ye, Zhiyuan Liu, and Maosong Sun. 2019. OpenNRE: An open and extensible toolkit for neural relation extraction. In *Proceedings of EMNLP-IJCNLP: System Demonstrations*, pages 169–174.

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

Anwen Hu, Shizhe Chen, and Qin Jin. 2020. ICECAP: information concentrated entity-aware image captioning. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, pages 4217–4225. ACM.

Justin Johnson, Andrej Karpathy, and Li Fei-Fei. 2016. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Andrej Karpathy and Fei-Fei Li. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, pages 3128–3137. IEEE Computer Society.

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

Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. 2020. Visualnews : Benchmark and challenges in entity-aware image captioning.

Y. Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020.

Arnau Ramisa, Fei Yan, Francesc Moreno-Noguer, and Krystian Mikolajczyk. 2018. Breakingnews: Article annotation by image and text processing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(5):1072–1085.

Joseph Redmon and Ali Farhadi. 2018. Yolov3: An incremental improvement. *CoRR*, abs/1804.02767.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Alasdair Tran, Alexander Mathews, and Lexing Xie. 2020. Transform and tell: Entity-aware news image captioning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *CVPR*, pages 3156–3164. IEEE Computer Society.

Xuewen Yang, Svebor Karaman, Joel Tetreault, and Alejandro Jaimes. 2021. Journalistic guidelines aware news image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5162–5175, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503.

A   Context Selection vs CLIP Retrieval

Previously, (Hu et al., 2020) demonstrate that cross-modal text retrieval model like VSE++ (Faghri et al., 2018) can localize image relevant news context to help improve news caption generation quality. To compare our context strategy with the image-to-text retrieval context selection method, we use the fine-tuned CLIP model as a stronger image-to-text retrieval context and select the optimal context as the top-k sentences that has the highest similarity to the image based on CLIP's prediction. Following (Hu et al., 2020), we iteratively sample k from 1 to 20 and find that the optimal value of k is 10 as it leads to the best CIDER score of Tell on the NY800K dataset. The comparison between the context selected by our proposed strategy and that retrieved by CLIP is summarised in Table 12. We observe that Tell trained on our selected context is consistently better than that is trained on CLIP retrieved news context. We find that this improvement comes from entities that cannot be captured by cross-modal retrieval, such as the non-visual named entities. CLIP-retrieved context also does not include the global context, which missed the high-level summary of the major story of the news.

Context	BLEU-4	CIDER	Named Entities	
			P	R
CLIP Retrieved	6.4	57.5	25.7	22.7
Our	**7.0**	**63.6**	**29.8**	**25.9**

Table 12: Evaluation of Tell with our proposed context against Tell trained on CLIP retrieved top k sentences from news articles.