Exploring the Impact of Vision Features in News Image Captioning

Junzhe Zhang and **Xiaojun Wan**
Wangxuan Institute of Computer Technology, Peking University
junzhezhang@stu.pku.edu.cn
wanxiaojun@pku.edu.cn

Abstract

The task of news image captioning aims to generate a detailed caption which describes the specific information of an image in a news article. However, we find that recent state-of-art models can achieve competitive performance even without vision features. To resolve the impact of vision features in the news image captioning task, we conduct extensive experiments with mainstream models based on encoder-decoder framework. From our exploration, we find 1) vision features do contribute to the generation of news image captions; 2) vision features can assist models to better generate entities of captions when the entity information is sufficient in the input textual context of the given article; 3) Regions of specific objects in images contribute to the generation of related entities in captions.[1]

1 Introduction

Image captioning is a multi-modal task which has developed a lot in recent years (Wang et al., 2020; He et al., 2020; Sammani and Melas-Kyriazi, 2020). Image captioning models can generate captions which accurately describe the generic object categories and object relations in images on image captioning datasets such as COCO(Lin et al., 2014; Chen et al., 2015) or Flickr(Hodosh et al., 2013). However, generic image captioning datasets above often contain less details in captions such as names, places or other specific entity information which are common in captions of news images.

News image captioning(Lu et al., 2018; Biten et al., 2019; Tran et al., 2020) aims to generate more specific descriptions of images by providing rich contextual information in associated news articles. Specifically, with a news image and the corresponding article given, models need to generate a caption which not only describes the whole image generally but also contains the specific information such as names or places of objects in the image. There has made significant progress with the introduction of transformer-based end-to-end captioning methods (Tran et al., 2020; Liu et al., 2021; Yang et al., 2021).

As a multi-modality task, news image captioning models usually generate captions with both textual and vision features. Note that a series of explorations (Elliott, 2018; Caglayan et al., 2019; Wu et al., 2021; Li et al., 2022) have been made for other multi-modality tasks like Multimodal Machine Translation(MMT) to resolve the impact of vision features. It is natural and important to explore the role of vision features in the news image captioning task where all specific textual information is in news articles. We come up with these questions: **RQ1: Do vision features help or not?** and **RQ2: How do vision features help?** Particularly, **RQ2** can be decomposed into two subquestions: **RQ2-1: Which part in captions do vision features help to generate?** and **RQ2-2: Which part in images helps the generation?**

In order to answer the above questions, we conduct a series of experiments using the most successful news image captioning models in recent studies on two main news image caption datasets GoodNews(Biten et al., 2019) and NYTimes800k(Tran et al., 2020). We first evaluate models under incongruent vision features to preliminarily determine whether the model is sensitive to the vision features. Then we modify the vision features and textual features respectively to explore the specific contribution of vision features to the captioning process. We cover the specific type of image regions to explore the relationship between the generation of entities and their related regions of images. Following Caglayan et al. (2019)'s work, we also conduct probing tasks to find how vision features affect caption generation under insufficient input textual features. Our main conclusions are:

[1] Our code is available at https://github.com/reroze/Explore_Vision_impact_NIC

Findings of the Association for Computational Linguistics: ACL 2023, pages 12923–12936
July 9-14, 2023 ©2023 Association for Computational Linguistics

- Vision features do contribute to generating news image captions. (Answer to **RQ1**)

- Vision features can better improve model's ability to generate specific entities which appear in textual context. The scarcity of textual entity information will seriously damage the impact of vision features. (Answer to **RQ2-1**)

- The specific regions of objects in images can help models generate the related entity information in captions more accurately when the corresponding entity information is sufficient in textual context. (Answer to **RQ2-2**)

2 Exploration Method

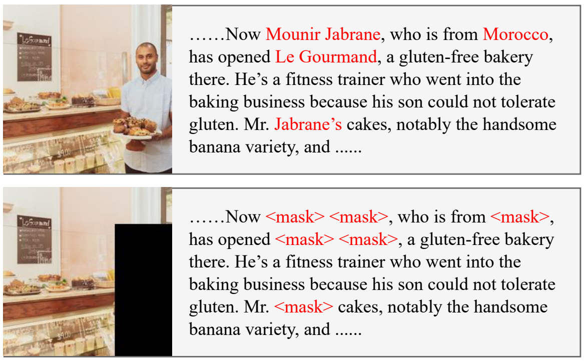
We introduce a series of methods to analyze whether and how vision features help, respectively. We start with the definition of Incongruent Decoding and then introduce the Degradation and Cover methods for textual and vision features respectively.

2.1 Incongruent Decoding

Incongruent decoding is widely considered as a method to evaluate whether visual information plays a role in MMT tasks (Elliott, 2018; Caglayan et al., 2019). We evaluate the performance of models under incongruent multi-modal information while keeping the congruence of multi-modal features during the training stage. Specifically, we replace the input image randomly with another image in the same dataset during the validation period. Generally, this method will decrease the performance if the model is sensitive to the visual modality information.

2.2 Article Degradation

According to previous work (Caglayan et al., 2019; Li et al., 2022), we analyze the contribution of vision features to generating concrete entity information in captions by providing entity mask(EM) experiment. Following the previous configuration in Tran et al. (2020); Yang et al. (2021), we obtain the entities in articles and captions by SpaCy[2] and then mask tokens of most entities with <mask> as shown in Figure 1. Nearly 14% of all words are masked in news articles of training and test data. To exclude the influence of the degradation method itself, we also mask words which are not entities randomly with the same rate. Through article degradation experiments, we can analyze the way images help generating the specific entities: generating them from scratch or assist models to select the right entities appearing in textual context.



……Now Mounir Jabrane, who is from Morocco, has opened Le Gourmand, a gluten-free bakery there. He's a fitness trainer who went into the baking business because his son could not tolerate gluten. Mr. Jabrane's cakes, notably the handsome banana variety, and ……

……Now <mask> <mask>, who is from <mask>, has opened <mask> <mask>, a gluten-free bakery there. He's a fitness trainer who went into the baking business because his son could not tolerate gluten. Mr. <mask> cakes, notably the handsome banana variety, and ……

Figure 1: The original multi-modal information (upper) and the modified multi-modal information in Article Degradation and Image Cover experiments (below)

2.3 Image Cover

In order to further study the mechanism of visual modality in news image captioning, we explore the relationship between specific regions of images and the generation of corresponding entities in captions. Since 80% of images are detected with regions of people, and the average proportion of regions accounts for approximately half of all images, we choose regions of people and people's names to investigate the relationship during our image cover experiments. According to (Tran et al., 2020), we first recognize people regions of images with YOLOv3 (Redmon and Farhadi, 2018) and then cover them with blank regions, as shown in Figure 1, on both training and test sets.

Statistics	GoodNews	NYTimes800k
Recognition of People		
- Recognition Rate	80%	80%
- Proportion of Regions	49%	45%
Matches of Entities		
- In News Articles	39%	65%
- In Limited Context	36%	51%

Table 1: Analysis for recognition of regions with people and matches of Named entities on datasets GoodNews and NYTimes800k with SpaCy and YOLOv3.

[2] We use SpaCy which achieve pos tagging accuracy of 97.0% and NER F-score of 86.6% on the OntoNotes corpus.

3 Experiments Setup

3.1 Datasets

We conduct our experiments on two public news captioning datasets: GoodNews and NYTimes800k which are both collected from The New York Times API[3]. We follow the data split by the original authors. Goodnews contains 421K, 18K, and 23K instances in training, validation, and test sets, respectively. Compared to GoodNews, NYTimes800k is a larger and more complete dataset that consists of 763K training, 8K validation, and 22K test instances. According to (Tran et al., 2020)'s research, in GoodNews and NYTimes800k, the entities account for 27% and 26% of the total words in captions, respectively. The average article length is 451 in GoodNews, and 974 in NYTimes800k.

We further analyze the occurrence rate of entities in captions. Specifically, we first apply SpaCy to recognize the entities in articles and captions, then exactly compare their strings to obtain the match rate. From Table 1, we see only 39% of entities in captions match entities in articles on GoodNews. Though NYTimes800k contains longer and more sufficient articles, there are still 35% of entities which do not appear in the articles. Besides, we also employ YOLOv3 to detect regions of people in news images. On GoodNews and NYTimes800k, all regions of people account for 49% and 45% of the whole image area respectively, and 80% of images are detected with regions of people on both datasets. Following (Tran et al., 2020), we choose first 500 words as limited input context in GoodNews, and choose location-aware paragraphs until they contain more than 512 sub-words as limited input context in NYTimes800k.

3.2 Multi-Modality Features

Vision Features We use the ResNet-152(He et al., 2016) model pre-trained on ImageNet to obtain the representations of images. Following the settings in Tran et al. (2020), we use the same image pre-process pipeline and then use the output before the average pooling layer of ResNet-152 to obtain our vision features.

Textual Features According to Tran et al. (2020), we obtain the textual features with the RoBERTa-large model (Liu et al., 2019), which is a pretrained model including 24 layers of bidirectional transformer blocks and encodes text to contextual embeddings. We use the weighted RoBERTa technique that obtains final textual features by using a weighted sum of the output from each transformer layer and initial uncontextualized embeddings. On Goodnews dataset, which does not contain image location information, we use the first 512 sub-words obtained by BPE(Byte Pair Encoding) technique in RoBERTa. On NYTimes800k dataset, with the image location provided, we use the location-aware 512 sub-words tokens as textual input features for both textual-modality-only and multi-modality models.

3.3 Models

Tell (Tran et al., 2020) is an open-source SOTA model which uses RoBERTa to encode input articles and ResNet-152 to extract features from input images. Tell uses a transformer decoder to generate a caption with dynamic convolutions (Wu et al., 2019) to attend to the generated tokens and multi-head attention (Vaswani et al., 2017) to attend to the multi-modality features. Tell also equips the weighted RoBERTa technique and extracts location-aware text. We also implement Tell with these techniques when only textual features are provided. Tell with weighted RoBERTa and location-aware technique is referred as Tell(full) in our experiments.

JoGANIC (Yang et al., 2021) is a transformer-based model with component template guidance following the journalistic guidelines. Oracle template vectors are generated from captions according to the high-level component class which contains several component classes(Who, When etc.) based on journalistic guidelines. JoGANIC uses a hybrid transformer decoder to generate captions with the template vector predicted by a multi-layer perceptron. JoGANIC also applies extra Named Entity Embedding(NEE) and Multi-Span Text Reading(MSTR) method to obtain better representations of entities and read longer articles. In our experiments, we implement JoGANIC with the same multi-modality features obtained in Section 3.2.

Tell(EG+SA) (Liu et al., 2021) encodes named entities as an independent textual input with article encoder, and uses AoA(Attention On attention) module which is an extension of self-attention mechanism to attend to the generated

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

	Model	Vision Feature	BLEU-4	ROUGE	CIDEr	Named entities		People's names	
						P	R	P	R
GoodNews	Tell	-	4.60	18.60	40.90	19.30	16.10	24.40	18.70
	Tell(full)	-	5.14	19.06	44.79	19.90	17.07	25.78	20.59
		ResNet	5.92	21.00	51.97	21.70	18.61	28.38	22.56
	JoGANIC*	-	5.09	19.07	44.10	19.67	16.72	25.79	20.06
		ResNet	5.69	20.69	50.42	20.98	18.05	26.22	21.66
	Tell(EG+SA)	-	5.17	19.18	46.33	19.71	17.20	25.69	20.74
		ResNet	6.05	21.18	53.10	21.48	18.70	27.97	22.92
NYTimes800k	Tell	-	4.26	17.30	33.90	17.80	16.30	23.60	19.70
	Tell(full)	-	5.33	19.19	44.71	21.51	19.99	30.62	27.02
		ResNet	6.21	21.36	53.20	23.95	21.73	35.84	30.17
	JoGANIC*	-	5.29	18.99	43.98	21.34	19.72	31.00	26.10
		ResNet	6.30	21.39	53.91	24.24	22.20	35.57	30.23
	Tell(EG+SA)	-	5.22	19.27	45.44	21.79	20.15	30.93	27.34
		ResNet	6.20	21.51	53.59	23.81	21.87	35.50	30.47

Table 2: Results of models on GoodNews and NYTimes800k, we implement Tell(full), JoGANIC and Tell(EG+SA) with or without vision features with standard news image captioning data. We report BLEU-4, ROUGE, CIDEr, precision(P) and recall(R) of Named entities, and People's names. We implement the JoGANIC model based on the framework of Tell since no official code for JoGANIC has been released. We directly use the results of Tell model from Tran et al. (2020).

tokens. Inspired by (Liu et al., 2021), we also implement Tell(full) with entity guidance(EG) technique and replace Dynamic Convolution with self-attention(SA) mechanism, which is referred as Tell(EG+SA) in our experiments.

We conduct experiments on extra models in Appendix A.

3.4 Implementation Details

We follow (Tran et al., 2020) and (Yang et al., 2021) to conduct our experiments with all models.

The hidden size of input textual features and vision features are 1024 and 2048. Tell(full) includes 4 transformer decoder layers and JoGANIC includes 8 transformer decoder layers. The number of heads is 16 in multi-modality attention. Specifically, we implement Tell(full) model without face and object encoder to make sure the multi-modality features are the same for all models.

Our implementation is based on PyTorch(Paszke et al., 2017) and AllenNLP framework(Gardner et al., 2018), and our training settings are the same as (Tran et al., 2020). We apply fairseq(Ott et al., 2019) to accomplish RoBERTa model and dynamic convolution. We use a batch size of 16 and train all models for 16 epochs for GoodNews and 9 epochs for NYTimes800k. For training, we use Adam optimizer(Kingma and Ba, 2015) with $\beta_1 = 0.9, \beta_2 = 0.98, \epsilon = 10^{-6}$. We apply mixed precision to train models on a single 3080Ti GPU for 3 to 4 days on both datasets.

We use the following evaluation metrics: BLEU-4 (Papineni et al., 2002), ROUGE (Lin, 2004) and CIDEr (Vedantam et al., 2015) obtained by COCO caption evaluation toolkit[4]. According to previous research (Tran et al., 2020; Liu et al., 2021), CIDEr is the most suitable metric for the news image captioning task. We use SpaCy to identify named entities in both generated captions and ground-truth captions, and then obtain the precision and recall rate by matching the texts of entities exactly. We select entities with PERSON label to obtain the score of people's names.

4 Result & Analysis

4.1 Helpfulness of Vision Features

We first implement news image captioning models on standard news image captioning data. Table 2 shows the performance of multi-modality models and their text-only versions, respectively. From the table, we can see models using multi-modality features achieve better results over all automatic metrics than relying only on textual features.

[4] https://github.com/tylin/coco-caption

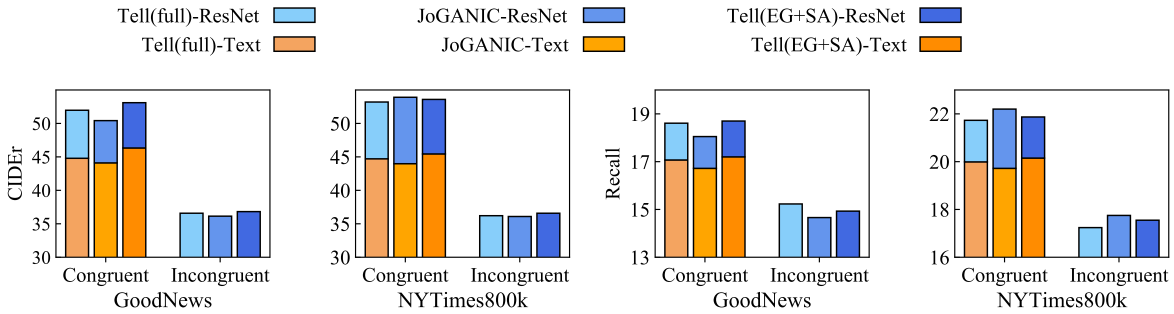


Figure 2: CIDEr and Recall of Named entities with Congruent and Incongruent multi-modality features, all models are trained with congruent multi-modality features.

Systems	Vision Feature	CIDEr	Named entities R
Tell(full)	Rd-ResNet	44.35	20.06
Blind	ResNet	44.85	20.13
Blank	ResNet	44.82	19.84
Incongruent	ResNet	36.21	17.24

Table 3: CIDEr and Recall of Named entities with Tell(full) under other training settings on NYTimes800k. The Blind system uses incongruent vision features both to train and test models.

Figure 2 shows that both CIDEr and Recall of named entities exhibit a significant drop for all models with incongruent vision features on GoodNews and NYTimes800k. The result from the incongruent decoding experiment indicates that models are sensitive to vision features while generating captions.

We also train and test Tell(full) with incongruent vision features or vision features from blank images to exclude the influence of parameter size. Besides, we use vision features from random-initialized ResNet(Rd-ResNet) for training and testing to determine whether vision features contribute by reducing overfitting. From Table 3, we can see under Blind, Blank, and random-initialized ResNet training settings, Tell(full) obtains CIDEr and Recall of Named entities close to text-only models. This result indicates that vision features contribute to the generation instead of just providing more parameters or preventing model training from overfitting.

4.2 Visually Sensitive Parts of Captions

We first conduct Article Degradation experiments to indicate that vision features can hardly assist models in generating entities when the entities in textual context are masked. Since there are many entities in captions which are also not contained in the standard textual context, we further analyze the Recall of entities in and out of textual context.

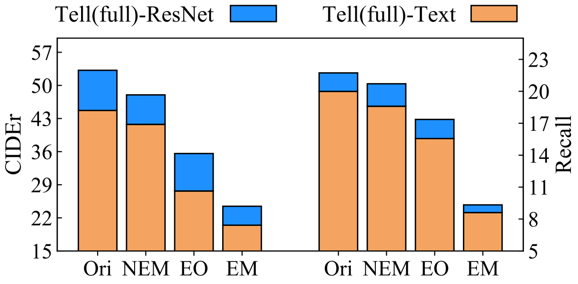


Figure 3: CIDEr and Recall of Named entities of four different Article Degradation methods with Tell(full) Model on NYTimes800k. NEM for masking non-entity words randomly, EO for input entity sequence only.

Table 4 summarizes the result of models in Article Degradation Experiments. For Tell(EG+SA), we also mask the independent entity sequences. Compared to table 2, we can see that in Article Degradation experiments, models with text-only features or multi-modality features both perform badly on automatic metrics and the generation of entities. What's more, masking the entity information also shrinks the improvement of multi-modality models compared to text-only models.

We randomly mask the words except for entities in context with the same mask rate (called NEM) to exclude the influence introduced by the degradation method itself. To explore the importance of entities in context, we also conduct the degradation experiment with only entity sequences provided as textual context information(called EO). In Figure 3, Ori means original articles, and EM means the entity mask article degradation method. Models are trained and tested both with the corresponding insufficient textual context. The result shows that the performance of the model declines little while masking other non-entity words, and the im-

Systems		GoodNews			NYTimes800k		
Model	Vision Feature	BLEU-4	CIDEr	Named entities R	BLEU-4	CIDEr	Named entities R
Tell(full)	-	2.87	22.30	8.21	2.86	20.45	8.61
Tell(full)	ResNet	3.44	26.46	8.93	3.37	24.45	9.33
JoGANIC	-	2.98	23.15	8.61	2.91	21.15	9.01
JoGANIC	ResNet	3.68	27.82	9.34	3.57	26.19	9.74
Tell(EG+SA)	-	2.84	22.52	8.17	2.84	20.92	8.65
Tell(EG+SA)	ResNet	3.48	26.53	9.04	3.36	25.24	9.17

Table 4: The Results of Article Degradation Experiments on GoodNews and NYTimes800k datasets. We report BLEU-4, CIDEr, and the Recall(R) of Named entities.

	Model	Vision Feature	Named entities In	Out
GoodNews	Tell(full)	-	36.26	6.26
		ResNet	38.98	**7.13**
	JoGANIC	-	35.49	6.14
		ResNet	37.86	6.88
	Tell(EG+SA)	-	36.52	6.31
		ResNet	**39.60**	6.91
NYTimes800k	Tell(full)	-	35.43	3.91
		ResNet	38.22	4.56
	JoGANIC	-	34.97	3.83
		ResNet	**38.83**	**4.88**
	Tell(EG+SA)	-	35.81	3.84
		ResNet	38.51	4.54

Table 5: Recall of Named entities which are in or out of the input textual context with standard multi-modality features on GoodNews and NYTimes800k.

provement of vision features shrinks when no entity information is provided by textual context.

Considering that many entities are not contained in the textual context, we further analyze the impact of vision features on entities in and out of the textual context. Table 5 indicates that vision features can better improve models to generate entities contained in the textual context and contribute little to generating entities out of textual context. The result indicates that the textual context insufficiency exists without any degradation method and limits the contribution of vision features since the entities which are not contained in textual context are less sensitive to vision features.

4.3 Impact of Image Regions on Entity Generation

We divide the test dataset into two subsets by whether the image is detected with regions of people by YOLOv3. We then conduct our image cover experiment with three models on the subsets and analyze the Recall of people's names. Table 6 shows that models with vision features can better generate people's names on the subset where images contain regions of people, compared to the text-only models. On the other subset where images do not contain the regions of people, the performance of most multi-modality models is lower than that of text-only models. After the corresponding regions are covered, models achieve a lower Recall of people's names on both subsets. After the discussion in Section 4.2, we further analyze people's names that do and do not occur in the input context on the subset where images contain people regions. Table 7 shows that Recall of people's names obtains more significant improvement on people's names which also appear in context. These results indicate that the vision features extracted from regions of people contribute to the generation of people's names in captions.

However, with regions of people in images covered, models can still obtain a higher Recall of people's names compared to text-only models. The difference between using original and covered images further shrinks for all people's names because of the insufficient entity information in articles. A possible reason is that models can infer the existence of people from the covered images since we mask all regions with blank, and knowing the occurrence of people may be enough for models to generate better people's names in some samples. We analyze Tell(full) with the original and covered images on NYTimes800k dataset. Table 8 indicates that vision features extracted from covered images still contribute a lot to samples where text-only Tell(full) can hardly produce any people's names. We leave the detailed mechanism as future work.

	Model	With Regions of People			Without Regions of People		
		Text	Origin	Cover	Text	Origin	Cover
GoodNews	Tell(full)	21.14	23.29	22.68	14.40	14.31	13.67
	JoGANIC	20.57	22.45	21.84	14.36	12.75	12.56
	Tell(EG+SA)	21.30	**23.75**	22.51	**14.40**	13.58	13.33
NYTimes800k	Tell(full)	28.17	31.77	30.23	16.71	15.85	15.71
	JoGANIC	27.29	31.75	30.13	15.47	16.52	15.66
	Tell(EG+SA)	28.47	**32.09**	30.29	**17.14**	15.95	15.71

Table 6: Recall of People's names for models applying textual features and multi-modality features with original or covered images on two subsets from GoodNews and NYTimes800k divided by whether images contain regions of people.

Models	Text	Origin	Cover
GoodNews			
Tell(full)	43.58	47.40(↑ 3.82)	46.50(↑ 2.92)
JoGANIC	42.00	45.84(↑ 3.84)	44.43(↑ 2.43)
Tell(EG+SA)	43.79	**48.86**(↑ **5.07**)	46.53(↑ 2.74)
NYTimes800k			
Tell(full)	44.26	49.58(↑ 5.32)	47.14(↑ 2.88)
JoGANIC	42.64	49.38(↑ **6.74**)	47.01(↑ 4.37)
Tell(EG+SA)	44.58	**50.22**(↑ 5.64)	47.50(↑ 2.92)

Table 7: Recall of People's names contained in input textual context with image cover experiment on the subset where the image is detected with regions of people from GoodNews and NYTimes800k.

Vision Feature	#Names generated by Text-only Model		
	0	1	2+
-	0	49.59	**53.53**
covered	26.19	52.09	48.32
origin	**33.48**	**54.03**	49.63

Table 8: Recall of People's names contained in input textual context on three subsets of the test dataset. All subsets are extracted from NYTimes800k test dataset with regions of people contained in images and divided by the number of generated People's names using text-only Tell(full). The captions in corresponding subsets contain 14%, 49%, and 37% of all People's names in the three subsets, respectively.

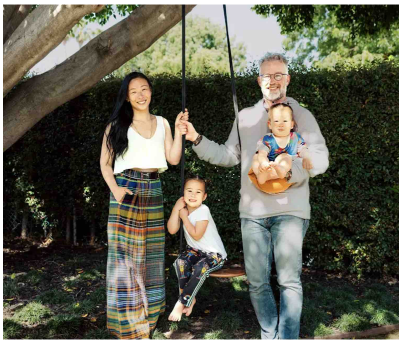
4.4 Case study

At last, we choose an example from NYTimes800k to specifically analyze the impact of vision features under diverse multi-modality features. We use Tell(full) to generate captions with text-only or multi-modality features extracted from different exploration methods, as shown in Figure 4. We see that Tell(full) with original multi-modality features successfully generates all people's names and achieves the highest Recall of all entities. Specifically, Tell(full) with images where regions of people are covered performs well on the generation of other entities except for people's names "Zachary Lara" and "Sonya Yu". That indicates the specific regions are helpful for the generation of related entities. Besides, when the entity information is degraded from the input context, Tell(full) can not generate the correct entities with or without vision features.

5 Related Work

Image captioning has improved significantly in recent years, and early models (Karpathy and Fei-Fei, 2015; Vinyals et al., 2015; Donahue et al., 2015) exploited convolutional neural network and recurrent neural network to encode images and decode captions, respectively. Xu et al. (2015) attended to different image patches when generating different tokens and Wang et al. (2019) applied the attention mechanism to regions of the corresponding object.

News image captioning takes the news article-image pairs as input and generates image captions rich in named entity information, making it a challenging task. News image captioning models generate specific entity information by applying the textual information from the associated articles. There are two main categories of news image captioning models: (1) template-based methods (Lu et al., 2018; Biten et al., 2019) which first generate the intermediate templates with placeholders for entities and then extract the specific entity information from associated articles. (2) end-to-end methods (Hu et al., 2020; Tran et al., 2020; Liu et al., 2021) which generate the whole caption directly in one step. Tran et al. (2020) applied pre-trained model RoBERTa as text encoder and transformer decoder to generate the captions with byte-pair-encoding(BPE)(Sennrich et al., 2016). Liu et al. (2021) utilized transformer architecture with ex-



(1) Zachary Lara and Sonya Yu (shown with their daughter, Evelyn, and son, August) bought and renovated a 1920s Spanish Colonial-style house in Los Angeles as their getaway.	(2) Zachary Lara and Sonya Yu with their children, Evelyn, 4, and August, 1.	(3) The couple's home in the Sunset Square neighborhood was a little more than a year ago.
(4) The couple's two children, Evelyn, 4, and August, 1, in the Sunset Square neighborhood of Los Angeles.	(5) The couple with their children, from left, Kaitlyn, 4, and Kaitlyn Yu, 4, and their daughter, Chia.	(6) The living room of the house, which was designed by the architect David and the architect John C. Liu.

An Indoor-Outdoor Escape in Los Angeles
When most urbanites begin looking for a second home, they dream of finding a pastoral landscape or beachfront place as a refuge from hectic city life. But Zachary Lara and Sonya Yu, who live in San Francisco, wanted something else: a foothold in a second city where they could find more creative energy. "It's something that we really crave," said Mr. Lara, 44, a technology consultant who also works in real estate development with Ms. Yu. But with plans to start a family - they now have two children, Evelyn, 4, and August, 1 - they weren't immune to the appeal of outdoor space and a less vertiginous lifestyle. "In San Francisco, our house is on one of the steepest hills in the city. It's four stories, and there are lots of stairs," Ms. Yu said. For their getaway, "I wanted a single story where we didn't have to go up and down stairs all day," she said. "And we wanted a pool for the kids." After a year of hunting for a Spanish Colonial-style home that checked those boxes, they were thrilled to find an online listing for an ideal-looking 1920s house in the Sunset Square neighborhood ......

Figure 4: An example of news image caption generation. (1) is the ground-truth caption and others are generated by Tell(full) with: (2) standard multi-modality features. (3) textual features only. (4) original article with the covered image. (5) degraded article with original image. (6) degraded article only. We highlight the entities in captions in red.

tra methods like Entity-Guide and Visual-selective layer to obtain better textual and vision features. The early template-based method ended up with relatively low performance due to generating captions in the linguistically limited context in the second step. Yang et al. (2021) generated captions directly with predicted template guide vectors based on transformer architecture as well and achieved competitive performance. Zhou et al. (2022) proposed a method to select more relevant and sufficient context with multi-modal pre-trained model CLIP(Radford et al., 2021) and relation extraction model to obtain better performance.

News image caption task is a multi-modal task with textual and vision features as input. Some of the previous models can obtain competitive performance even with single-textual input only, which also happened in other multi-modal tasks like MMT(Multimodal Machine Translation).Elliott (2018) found that models with random unrelated images can also obtain competitive results, Caglayan et al. (2019); Li et al. (2022) conducted probing mask experiments and pointed out that vision features contribute more when the input text is insufficient. Wu et al. (2021) found that vision features assist MMT models through regularization training. We refer to some of these methods to analyze the contribution of vision features in news image captioning task as well.

6 Conclusions

In this work, we design and conduct extensive experiments to explore the impact of vision features on news image captioning models. First, we determine that vision features do contribute to generating news captions. Second, from our degradation experiment, we find that vision features can help models obtain a better generation of entities that appear both in textual context and captions. The low inclusion of entities in textual context will limit the improvement obtained from applying vision features. At last, we show that specific regions of images help models with the better generation of related entities in captions. However, images with regions covered can also generate the corresponding entities better. We believe it is important for future research to improve the ability of models to generate entities out of textual context and make better use of the specific information of different regions in images.

Limitations

Our experiments are conducted on transformer-based models with the same multi-modality features. Considering the importance of entity information in textual context and specific regions of images, it is also important to investigate whether the performance of the model promotes with different methods of extracting multi-modality features. We use the bpe technique to encode all entities in input articles which may separate the whole entity word into several sub-word tokens and may affect the impact of vision features. There are still a lot of entities of captions that don't appear in articles from datasets on our experiments. Conclusions will be more powerful if we can conduct experiments on other datasets which contain more sufficient information of captions.

Acknowledgements

This work was supported by National Key R&D Program of China (2021YFF0901502), National Science Foundation of China (No. 62161160339), State Key Laboratory of Media Convergence Production Technology and Systems and Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology). We appreciate the anonymous reviewers for their helpful comments. Xiaojun Wan is the corresponding author.

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)*.

Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. 2019. Probing the need for visual context in multimodal machine translation. 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 4159–4170, Minneapolis, Minnesota. Association for Computational Linguistics.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. 2015. Microsoft COCO captions: Data collection and evaluation server. *CoRR*, abs/1504.00325.

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

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Desmond Elliott. 2018. Adversarial evaluation of multimodal machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2974–2978, Brussels, Belgium. Association for Computational Linguistics.

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

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

Sen He, Wentong Liao, Hamed R. Tavakoli, Michael Yang, Bodo Rosenhahn, and Nicolas Pugeault. 2020. Image captioning through image transformer. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899.

Anwen Hu, Shizhe Chen, and Qin Jin. 2020. Icecap: Information concentrated entity-aware image captioning. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4217–4225.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR (Poster)*.

Bei Li, Chuanhao Lv, Zefan Zhou, Tao Zhou, Tong Xiao, Anxiang Ma, and JingBo Zhu. 2022. On vision features in multimodal machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6327–6337, Dublin, Ireland. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham. Springer International Publishing.

Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. 2021. Visual news: Benchmark and challenges in news image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6761–6771, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. 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. Association for Computational Linguistics.

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.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch. In *NIPS 2017 Workshop on Autodiff*.

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. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Joseph Redmon and Ali Farhadi. 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.

Fawaz Sammani and Luke Melas-Kyriazi. 2020. Show, edit and tell: A framework for editing image captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

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 (Volume 1: Long Papers)*, pages 1715–1725.

Alasdair Tran, Alexander Mathews, and Lexing Xie. 2020. Transform and tell: Entity-aware news image captioning. In *Proceedings of the 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Weixuan Wang, Zhihong Chen, and Haifeng Hu. 2019. Hierarchical attention network for image captioning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):8957–8964.

Zeyu Wang, Berthy Feng, Karthik Narasimhan, and Olga Russakovsky. 2020. Towards unique and informative captioning of images. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII*, page 629–644, Berlin, Heidelberg. Springer-Verlag.

Felix Wu, Angela Fan, Alexei Baevski, Yann Dauphin, and Michael Auli. 2019. Pay less attention with lightweight and dynamic convolutions. In *International Conference on Learning Representations*.

Zhiyong Wu, Lingpeng Kong, Wei Bi, Xiang Li, and Ben Kao. 2021. Good for misconceived reasons: An empirical revisiting on the need for visual context in multimodal machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6153–6166, Online. Association for Computational Linguistics.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of*

Machine Learning Research, pages 2048–2057, Lille, France. PMLR.

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.

Mingyang Zhou, Grace Luo, Anna Rohrbach, and Zhou Yu. 2022. Focus! relevant and sufficient context selection for news image captioning.

A Appendix

We conduct extra experiments based on other models and vision features on NYTimes800k dataset. We implement LSTM-based model with RoBERTa and Transformer model with GloVe following the setting from (Tran et al., 2020) with location-aware textual context. We also test Tell(full) with vision features extracted from ViT-L/16 (Dosovitskiy et al., 2021). The performance on original data is shown in table 9. Their performance in the Article Degradation experiment is in table 10. Table 11 shows the result of the generation of Named entities by extra models, and table 12 shows the result of extra models in the Image Cover experiment. The results of extra experiments follow our conclusions.

Model	Vision Feature	BLEU-4	ROUGE	CIDEr	Named entities		People's names	
					P	R	P	R
Tell(full)	-	5.33	19.19	44.71	21.51	19.99	30.62	27.02
	ViT	6.37	21.60	54.34	24.09	21.92	35.81	30.27
LSTM	-	3.82	16.52	29.52	17.47	15.41	23.22	19.13
	ResNet	4.17	17.63	33.68	18.31	15.91	24.19	19.55
Transformer	-	2.17	14.25	17.46	13.21	10.16	12.73	8.93
	ResNet	2.74	16.13	21.13	13.98	11.05	14.45	10.22

Table 9: Results of extra models on NYTimes800k, we implement LSTM-based and Transformer-glove models with or without vision features with standard news image captioning data. We report BLEU-4, ROUGE, CIDEr, precision(P), and recall(R) of Named entities and People's names. We also implement Tell(full) with vision features extracted from VIT-L/16.

Model	Vision Feature	BLEU-4	ROUGE	CIDEr	Named entities		People's names	
					P	R	P	R
Tell(full)	-	2.86	14.90	20.45	10.46	8.61	9.58	7.18
	ViT	3.50	16.90	25.59	11.75	9.47	11.28	8.38
LSTM	-	2.03	13.34	13.32	8.43	6.37	7.25	4.98
	ResNet	2.26	14.20	15.40	8.78	6.67	7.24	5.12
Transformer	-	1.62	12.73	11.47	7.81	5.81	6.02	4.15
	ResNet	2.06	14.34	14.19	8.52	6.45	7.26	4.82

Table 10: Results of extra models in Article Degradation experiment on NYTimes800k. We report BLEU-4, ROUGE, CIDEr, precision(P), and recall(R) of Named entities and People's names.

Model	Vision Feature	Named entities	
		In	Out
Tell(full)	-	35.43	3.91
	ViT	**38.44**	**4.72**
LSTM	-	27.76	2.53
	ResNet	28.47	2.83
Transformer	-	17.93	2.07
	ResNet	19.17	2.60

Table 11: Recall of Named entities which are in or out of the input textual context by extra models with standard multi-modality features on NYTimes800k.

Model	Vision Feature	With Regions of People			Without Regions of People		
		Text	Origin	Cover	Text	Origin	Cover
Tell(full)	ViT	28.17	**31.94**	29.93	**16.71**	15.33	14.80
LSTM	ResNet	20.03	20.86	20.35	11.04	7.81	8.71
Transformer	ResNet	9.36	10.98	10.17	5.14	3.43	3.43

Table 12: Recall of People's names for extra models applying textual features and multi-modality features with original or covered images on two subsets from NYTimes800k divided by whether images contain regions of people.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitations

☑ A2. Did you discuss any potential risks of your work?
Limitations

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

3

☑ B1. Did you cite the creators of artifacts you used?
3

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
3

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
2,3

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
The specific information of names is important in our task and can not be anonymized.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
3

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
3

C ☑ Did you run computational experiments?

3

□ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
No response.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
3

☒ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
we conduct our experiment for a single run because of large training consume and conduct parallel experiment on many models

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
3

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.