Entity-Aware Multimodal Alignment Framework for News Image Captioning

Anonymous ACL submission

Abstract

News image captioning task is a variant of image captioning task which requires model to generate a more informative caption with news image and the associated news article. Multimodal Large Language models have developed rapidly in recent years and is promising in news image captioning task. However, according to our experiments, common MLLMs are not good at generating the entities in zero-shot setting. Their abilities to deal with the entities information are still limited after simply finetuned on news image captioning dataset. To obtain a more powerful model to handle the multimodal entity information, we design two multimodal entity-aware alignment tasks and an alignment framework to align the model and generate the news image captions. Our method achieves better results than previous state-of-the-art models in CIDEr score (72.33 -> 86.29) on GoodNews dataset and (70.83 -> 85.61) on Nytimes800k dataset.

1 Introduction

News image captioning, a variant of image captioning task that aims to generate the image caption which contains more specific information, has developed a lot as one of the standard multimodal tasks in recent years(Zhang et al., 2022; Rajakumar Kalarani et al., 2023; Qu et al., 2023). In news image captioning task, given news image and associated news article, models are required to generate an informative caption full of details such as names of people or news event. Compared with general image captioning task, news image captioning task requires model to not only understand the overall information of image, but also apply the knowledge of model to extract related specific information from associated news article.

Large Language model(LLM) has made a huge progress in recent years. By differently incorporating vision modality into LLM on multimodal dataset, Multimodal Large Language Models(MLLMs) also develop rapidly. There are a lot of MLLMs such as LLaVA-v1.5(Liu et al., 2023a), InstructBLIP(Dai et al., 2023) and others. The multimodal ability of MLLMs and knowledge in pretrained LLM will make a contribution to news image captioning task. However, there are many challenges on generating news image captions with MLLMs. First, few of them have been tested for news image captioning task, and their ability for news image captioning remains unknown though they are good at general image captioning. Second, most common MLLMs are pretrained on large scale multimodal datasets which are far from the domain of news image captions that are more rich in entities. Third, a news article is very long and contains many images, so it is challenging for MLLMs to find the related information in such long text input and generate the news image captions.

To solve these challenges, we first conduct plenty experiments with MLLMs in zero-shot setting and find that common MLLMs are not good at dealing with multimodal entity information. Some MLLMs even tend not to generate entities. We simply finetune the two common MLLMs, InstructBLIP and LLaVA on news image caption domain and find that they can then generate good captions. However, their ability to generate the entity information can be further improved by Entity-Aware Alignment. We believe that MLLMs are more powerful models and have the potential to generate high-quality captions of news images, so simply finetuning on news image captioning task is perhaps not enough for MLLMs to learn the ability to handle the entity recognition and generation since news image captioning dataset is more diverse that commonly used pretraining multimodal dataset. To further promote the ability to handle the entity information with MLLMs, we design two Entity-Aware alignment tasks and come up with an Entity-Aware

Multimodal Alignment Framework for news image captioning task. We explicitly utilize the aligned MLLM to refine the textual input, which achieves better performance and is more memory efficient than taking whole news article as textual input. Our method not only obtains the best result on all automatic metrics but also achieves the best performance on entities generation.

Our contributions are:

- We conduct experiments with MLLMs with various settings, and find that existed MLLMs are not good at dealing with entity information in zero-shot setting. The ability of generating entities remains insufficient after simply finetuned on news image captioning dataset and the performance of MLLMs are sensitive to the sufficient and concise textual input information.

- We design two entity-aware multimodal alignment tasks and an alignment framework to align MLLMs on fine-grained entity information level. Our aligned model obtains better results than previous methods.

- We design a method to explicitly use aligned MLLM to refine the textual input context on news image captioning task. We find that the improvement of simply expanding the textual input is not large while keeping the textual information related and concise can further improve the performance.

Code and data will be released to the community.

2 Related Works

2.1 News Image Caption

News image captioning(Biten et al., 2019; Tran et al., 2020) is a variant task of the general image captioning task which takes news images and associated news articles as input to generate informative captions of news images. In early researches, models(Lu et al., 2018) and (Biten et al., 2019)(Avg+CtxIns) often generate templates which don't contain the specific entities, and then extract entities from news articles.

However, template-based methods often constrain the representation of descriptions. New proposed end-to-end methods(Tran et al., 2020; Yang et al., 2021; Zhang et al., 2022; Rajakumar Kalarani et al., 2023; Qu et al., 2023) generate the final caption in one single step and obtain higher performance. Tell(Tran et al., 2020) applys RoBERTa and ResNet-152 to encode textual and visual information and uses transformer decoder to generate captions. JoGANIC (Yang et al., 2021) is a transformer based model with component template guidance following the journalistic guidelines. NewsMEP (Zhang et al., 2022) applies CLIP (Radford et al., 2021) to translate the image with open-domain multimodal knowledge and BART to encode textual and vision information simultaneously. Rajakumar Kalarani et al. (2023)(UVLP_OFA) pretrains OFA model on three vision-language tasks, and finetunes OFA on news image captioning datasets. Qu et al. (2023)(VA-BART) also utilizes BART architecture and design distinct modules for different vision inputs.

2.2 Multimodal Large Language Models and Alignment

Multimodal large language models have developed rapidly in recent years. BLIP-2(Li et al., 2023b) applys Q-Former architecture to transform image input into LLMs input tokens. LLaVA(Liu et al., 2023b) and LLaVA-v1.5(Liu et al., 2023a) utilize linear layers or perceptrons to map the vision features into the inputs of LLMs. Through instruction tuning on BLIP2, InstructBLIP(Dai et al., 2023) gains the ability to follow the instruction on different tasks. There are many other MLLMs such as mPLUG-Owl(Ye et al., 2023), Otter(Li et al., 2023a) and Qwen-VL (Bai et al., 2023). Among all MLLMs, GPT-4V(OpenAI et al., 2023) is the most powerful one now. We select some of these MLLMs on our research.

There are numerous studies on multimodal alignment tasks recent years. The most traditional task is image-text matching in CLIP(Radford et al., 2021). There are also fine-grained alignment tasks such as Region-Tag and Region-Noun alignment tasks in (Zhou et al., 2022b). MLLMs such as(Li et al., 2023b; Ye et al., 2023) often align textual features and visual features on multimodal tasks with language modeling loss.

3 Method

3.1 Overview

MLLMs (Multimodal Large Language Models) are highly effective in various multimodal tasks. In our approach, we utilize InstructBLIP, one of the most widely recognized MLLMs, as the

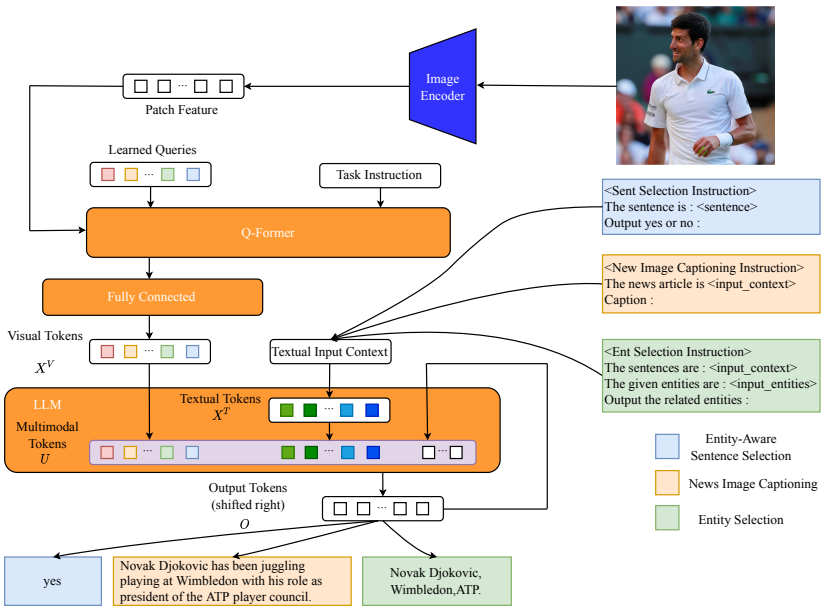

Figure 1: An overview of our model on Entity-Aware Sentence Selection, Entity Selection and News Image Captioning tasks.

foundational architecture. Figure 1 provides an overview of our model, highlighting its integration with Entity-Aware Sentence Selection, Entity Selection, and News Image Captioning tasks. We apply language modeling loss as our optimization objective on all tasks. For each instance in dataset I, the input image is embedded to visual soft tokens $X^V = \{x_1^V, x_2^V, ..., x_{n_v}^V\}$ after vision encoder and alignment modules, while the textual input context tokens T are embedded to $X^T = \{x_1^T, x_2^T, ..., x_{n_t}^T\}$ by LLM. During alignment, we combine these two parts and target tokens $O = \{o_1, o_2, ..., o_{n_o}\}$ into the multimodal input tokens $U = [X^V; X^T; O] = \{u_1, u_2, ..., u_{n_u}\}$ for the LLM, where $n_u = n_v + n_t + n_o$. Therefore, the multimodal language modeling loss can be written as

$$\mathcal{L}_{\text{LM}}(u) = -\frac{1}{n_o}\sum_{i=1}^{n_o} \log \mathbf{P}(o_i|o_1, ..., o_{i-1}) \quad (1)$$

where the conditional probability $\mathbf{P}$ is calculated by neural network with mutimodal features. In our method, we fully finetune the weights of Q-Former, Fully Connected Layer and LLM on our Entity-Aware alignment tasks.

3.2 Entity-Aware Multimodal Alignment

Entity-Aware Sentence Selection Previous research often involves selecting sentences related to the caption from the provided news article (Zhou et al., 2022a). A common approach is to employ an image-text matching model, such as CLIP, to extract relevant sentences based on news images. However, this method incurs extra training costs on CLIP because CLIP is trained on open-domain datasets, which differ from captions that are rich in entities.

Interestingly, image-text matching task is also commonly used to align vision and textual features in multimodal alignment techniques. Therefore, we attempt to create a specialized image-text matching task within the news image captioning task to align vision features with entity-aware textual features. Furthermore, given that MLLMs are capable of following instructions and managing multiple tasks, selecting sentences related to the caption can be seamlessly integrated, which also reduce the cost of training a sentence selection model separately.

Previous research(Zhou et al., 2022a) indicates that entity information is crucial in news image captioning tasks. Consequently, traditional image-text matching tasks may not be suitable for entity-aware tasks such as news image captioning. Therefore we construct an Entity-aware Sentence Selection as one of our Entity-Aware Multimodal Alignment tasks.

We utilize sentences from the news article as the text source. Given an instance with the news image V, news article $T = \{s_1, s_2, ..., s_m\}$ consisting of m sentences and the news image caption

C, we first recognize the visual entities $E^C = \{e_1^C, e_2^C, ..., e_{n_{E^C}}^C\}$ of the caption. The visual entities do not contain entities invisible in the news image such as 'DATE'.

Next, we select sentences from the news article that contain the same visual entities as the caption, considering them together with the corresponding caption itself as positive samples. Negative samples are constructed from the remaining sentences of the article. Each input instance comprises soft visual tokens X^V extracted from news images and textual tokens X^{s_i} embedded from s_i. The output O_{si} is labeled as

$$O_{si} = \begin{cases} yes, \ \exists e_j^C \in E^C, s.t. e_j^C \ \in s_i \\ no, \ otherwise \end{cases} \tag{2}$$

Specifically, one single group for sentence selection task contains three sentences which contain the most visual entities of caption together with one target caption as positive samples. Other four sentences are randomly selected from the sentences which do not contain any visual entity of caption as negative samples.

On dataset I, the final loss of Entity-Aware Sentence Selection task is

$$\mathcal{L}_{\text{LM}}^{\text{SENT}} = \mathbb{E}_{([V;s_i;O_{si}])\sim I} \mathcal{L}_{\text{LM}}(X^V; X^{s_i}; X^{O_{si}}]) \tag{3}$$

X^V, X^{s_i}, $X^{O_{si}}$ are visual soft tokens, textual input tokens and output tokens of the image, sentence and output label in each instance on Entity-Aware Sentence Selection task respectively.

In contrast, traditional image-text matching methods typically randomly select captions corresponding to other images as negative samples. However, in our approach, we extract sentences from the same article that do not contain the visual entities as hard negative samples. These sentences often exhibit a more similar distribution and cover the same topic, only lacking the key entity information. We believe this strategy provides the model with a deeper understanding of entities.

Entity Selection Compared with the Sentence Selection task, Entity Selection involves a more fine-grained alignment process. In the Entity Selection task, we select entities that occur in both the input text and the caption, following the order of appearance in the caption as the output sequence. The model takes the news image, associated input context, and recognized entities from the input context as input and predicts the output entities accordingly.

Specifically, for each instance of news image captioning dataset, we recognize the entities $E^{T'}$ from input context T' and E^C from the news image caption. The output is

$$O_e = \{e_i, e_i \in E^C \cap E^{T'}\} \tag{4}$$

We combine the input context T' and recognized entities $E^{T'}$ into the final input context T^*. In dataset I, the final loss of Entity Selection Task is

$$\mathcal{L}_{\text{LM}}^{\text{ENT}} = \mathbb{E}_{([V;T^*;O_e])\sim I} \mathcal{L}_{\text{LM}}([X^V; X^{T^*}; X^{O_e}]) \tag{5}$$

Similarly, X^V, X^{T^*}, X^{O_e} are visual soft tokens, textual input tokens and output tokens of the image, combined input context and output entities in each instance on the Entity Selection task, respectively.

It is worth mentioning that, unlike the Sentence Selection task, we retain both visual and non-visual entities in the Entity Selection task, inspired by Zhou et al. (2022a). Since models can access both the news image information and the background context information in the news article, we believe they have enough contextual information to assist in selecting the non-visual entities. These non-visual entities are often related to the visual entities in the news article and are also crucial for caption generation.

3.3 Model Training and Inference

Our alignment training tasks consist Entity-Aware Sentence Selection task, Entity Selection task and News Image Captioning task. We train models on all three tasks simultaneously.

Specifically, following BLIP2 (Li et al., 2023b), we organize the data into mini-groups, each containing 2 instances of the News Image Captioning task, 1 instance of the Entity-Aware Sentence Selection task, and 1 instance of the Entity Selection task. This ensures that all alignment tasks are trained in every batch. By doing so, we can balance the data flow across all alignment tasks and the loss for each task during the training stage. All three tasks use the multimodal language modeling loss defined in Equation1, and the final loss is formulated as

$$\mathcal{L} = \alpha \mathcal{L}_{\text{LM}}^{\text{SENT}} + \beta \mathcal{L}_{\text{LM}}^{\text{ENT}} + \gamma \mathcal{L}_{\text{LM}}^{\text{CAP}} \tag{6}$$

The $\mathcal{L}_{\text{LM}}^{\text{SENT}}$, $\mathcal{L}_{\text{LM}}^{\text{ENT}}$ and $\mathcal{L}_{\text{LM}}^{\text{CAP}}$ are the loss of Entity-Aware Sentence Selection task, Entity Selection task and News Image Captioning task, respectively.

Dataset	MLLM	BLEU-4	ROUGE	CIDEr	Named entities	
					P	R
GoodNews	BLIP-2+GPT-3*	2.06	**13.22**	**17.12**	/	/
	LLaVA-v1.5	1.98	12.59	13.49	**19.47**	6.92
	InstructBLIP	**2.52**	10.78	14.82	13.34	**8.78**
NYTimes800k	LLaVA-v1.5	1.99	**12.34**	12.32	**19.36**	7.59
	InstructBLIP	**2.52**	11.01	**12.74**	12.79	**11.21**

Table 1: Performance of models with MLLM technique in zero-shot settings. The result of BLIP-2+GPT-3* is cited from (Rajakumar Kalarani et al., 2023). LLaVA-v1.5 and InstructBLIP are tested with 7B checkpoints.

With various instructions, our aligned model can efficiently handle different tasks, contributing to better performance. When generating captions, we leverage our model to extract related sentences from the entire article, utilizing the training from our Sentence Selection task. Specifically, we first use our aligned model, following the instruction format from the Sentence Selection task, to acquire multimodal Entity-Aware related sentences. We then integrate these sentences with the default textual input context, maintaining the same order as they appear in the original news article. We eliminate redundant sentences and truncate the combined sentences to adhere to a 600-word limitation. Since the NYTimes800k dataset contains longer input articles with more unrelated information per instance compared with the GoodNews dataset, we limit the number of newly selected sentences to 4. By selecting related sentences, our method can incorporate more relevant textual information without significantly increasing the GPU memory overhead, unlike simply expanding the textual input.

4 Experiment Setup

4.1 Datasets & Metrics

We conduct experiments on two public large-scale News Image Captioning datasets: GoodNews(Biten et al., 2019), which contains 421K, 18K, 23K samples for train, validation and test datasets, and NYTimes800k(Tran et al., 2020) which contains 763K, 8K, 22K samples for train, validation and test datasets from The New York Times[1]. We follow the validation/test data split as previous researches and randomly select 100k samples for train dataset on both datasets. According to (Tran et al., 2020), over 96% news image captions contain entity information and approximately 26% words of news image captions are entities in two datasets. The average length of news articles and news image captions in GoodNews dataset are 451 and 18. In NYTimes800k, the average length of news articles and captions are 974 and 18.

We use BLEU-4 (Papineni et al., 2002), ROUGE (Lin, 2004) and CIDEr (Vedantam et al., 2015) as automatic evaluation metrics. Following previous research (Tran et al., 2020), we calculate these metrics with the COCO caption evaluation toolkit[2]. According to previous research (Tran et al., 2020; Zhou et al., 2022a; Zhang and Wan, 2023), CIDEr is the most important metric as it puts more attention to uncommon words which are often specific named entities. We also evaluate the Precision and Recall of named entities. We first use SpaCy to recognize the entities in generated captions and gold captions and count the number of exact match which requires the same text and label on entities.

4.2 Implementation Details

When finetuning InstructBLIP and LLaVA-v1.5, following the official finetune scripts, we finetune Q-Former and Fully Connected Layer in InstructBLIP and fully finetune LLaVA-v1.5 on checkpoints of 7B. We use the same prompt on news image captioning task in Figure 1 for training and testing LLaVA-v1.5 and InstructBLIP.

When aligning our model with two designed Entity-Aware alignment tasks, we set α, β and γ in Equation (6) as 0.25, 0.75 and 1 respectively in our model. More details can be found in appendix A.

5 Experimental Results

5.1 Results of Zero-shot MLLMs

LLaVA-v1.5(Liu et al., 2023a) and InstructBLIP(Dai et al., 2023) are two common MLLMs. We first test InstructBLIP and LLaVA-v1.5 in zero-shot setting and present results in Table 1.

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

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

	Model	BLEU-4	ROUGE	CIDEr	Named entities	
					P	R
GoodNews	Avg+CtxIns	0.89	12.20	13.10	8.23	6.06
	Tell	6.05	21.40	53.80	22.20	18.70
	JoGANIC	6.83	23.05	61.22	26.87	22.05
	NewsMEP	8.30	23.17	63.99	23.43	23.24
	UVLP_OFA	7.14	24.30	72.33	24.37	20.09
	VA-BART	7.91	22.61	69.11	24.38	23.56
	LLaVA-v1.5 (finetuned)	7.04	24.37	73.52	**27.38**	22.07
	InstructBLIP (finetuned)	9.53	25.61	78.03	25.89	27.33
	Ours	**10.07**	**26.78**	**86.29**	26.96	**28.48**
NYTimes800k	Tell	6.30	21.70	54.40	24.60	22.20
	JoGANIC	6.79	22.80	59.42	28.63	24.49
	NewsMEP	9.57	23.62	65.85	26.61	28.57
	UVLP_OFA	7.54	23.28	66.41	28.11	23.25
	VA-BART	8.81	23.02	70.83	27.57	27.16
	LLaVA-v1.5 (finetuned)	6.06	22.80	62.41	28.18	22.43
	InstructBLIP (finetuned)	10.05	25.45	75.95	27.39	30.37
	Ours	**10.94**	**27.00**	**85.61**	**29.32**	**31.85**

Table 2: Experimental results on GoodNews and NYTimes800k datasets compared with other models. P and R are Precision and Recall of generated Named entities. LLaVA-v1.5 and InstructBLIP are trained with their official scripts on pretrained 7B checkpoints. We directly cite the results of models which are not MLLMs from researches (Rajakumar Kalarani et al., 2023; Qu et al., 2023)

Through the results in Table 1, we can prove our previous assumption that MLLMs trained on general multimodal data can not generate informative news image captions. What's more, the Recall of named entities are much lower than the Precision of named entities on LLaVA, which means LLaVA tends not to generate entities information in zero-shot setting. The average number of generated entities by LLaVA-v1.5 are 1.22 and 1.24 on GoodNews and NYTimes800k datasets, while the average number of entities on references are 3.43 and 3.19 on two datasets, respectively.

5.2 Comparison Results of Different Models

Table 2 shows the comparison results of our proposed model with entity aware alighment and the baselines models, including the InstructBLIP and LLaVA-v1.5 models finetuned on news image captioning dataset.

According to the results in Table 2, simply supervised finetuning MLLMs on GoodNews and NYTimes800k dataset obtains fine performance on both datasets. After trained on news image captioning dataset, the performance of InstructBLIP and LLaVA-v1.5 improves a lot from zero-shot setting in Table 1. The ability of MLLMs to generate the entities also gains much improvement. Thus we believe MLLMs have the ability to handle the multimodal Entity-Aware tasks only if we train and align them properly with multimodal data rich in entity information. We observe that finetuned InstructBLIP obtains better results than LLaVA-v1.5. We think this is because InstructBLIP apply Q-Former architecture to compress vision features into 32 tokens, which can take more textual input tokens. However, though MLLMs gain obvious improvement on automatic evaluation metrics, the improvement on the generation of entities is not significant, which leaves space for further improvement.

We can also see that our proposed model with entity-aware alignment achieves the best performance on all metrics. Compared with previous SOTA (UVLP_OFA (Rajakumar Kalarani et al., 2023) in GoodNews and VA-BART (Rajakumar Kalarani et al., 2023) in NYTimes800k), our model gains 13.96 (from 72.33 to 86.29) and 14.78 (from 70.83 to 85.61) in CIDEr on GoodNews and NYTimes800k datasets, correspondingly. Compared with finetuned MLLMs, our method still obtains 8.26 (from 78.03 to 86.29) and 9.66 (from 75.95 to 85.61) in CIDEr on GoodNews and NY-

Model	BLEU-4	ROUGE	CIDEr	Named entities	
				P	R
InstructBLIP	10.05	25.45	75.95	27.39	30.37
InstructBLIP+Fully_sft	10.88	26.60	82.70	29.22	31.32
InstructBLIP+Fully_sft+Alignment	10.80	26.80	84.45	29.15	31.45
Ours	**10.94**	**27.00**	**85.61**	**29.32**	**31.85**
InstructBLIP+Fully_sft+Full_Input	10.76	26.18	83.17	27.71	30.03
InstructBLIP+Fully_sft+Oracle_Input	**11.24**	**27.70**	**87.56**	**33.42**	**34.80**

Table 3: Ablation study results on the NYTimes800k dataset. InstructBLIP denotes finetuning Q-Fromer and Fully Connected Layer while keeping the LLM freezed. We refer fully finetuned InstructBLIP which finetunes Q-Former, Fully Connected Layer and LLM on training stage as InstructBLIP+Fully_sft. InstructBLIP+Fully_sft+Alignment means .

Times800k datasets. Besides, our method not only achieves the best result on automatic evaluation metrics, but also on entity generation. Our method can obtain both high Precision and Recall of Entity generation at the same time.

5.3 Ablation Study

Ablation on Alignment Framework We conduct ablation experiments on NYTimes800k dataset to study the improvement of the different components of our model. Except following the default setting to finetune InstructBLIP, we also try to fully finetune the LLM weights which is referred to as (+Fully_sft). We also test our aligned model without explicitly applying the Sentence Selection ability, which is (+Fully_sft+Multi_Alignment).

From results on Ablation Study in Table 3, we find that fully finetuning LLM is important for InstructBLIP to generate news image captions. We empirically suppose that handling entity information such as names of people with multimodal information is a very difficult task and different from other common multimodal data, making only tuning the Q-Former and Fully Connected Layer not enough to learn the task. Further, our aligned model can obtain better results without explicitly applying the ability of Entity-Aware Sentence Selection (from 82.70 to 84.45 in CIDEr), which highlights the importance of entity-aware alignment. With explicitly utilizing the ability to select the multimodal entity-aware sentences to refine the textual input context, our method further improves the performance from 84.45 to 85.61 in CIDEr and obtains better generation of entities as well.

Ablation on Input Text We present the result with Oracle Selected Input which is combined by sentences contains the caption entities from whole news articles within 500 words. We named this result as (+Fully_sft+Oracle_Input). We also conduct experiments with full news articles input both in training and testing stages. We set the model_max_tokens to 2048 on experiment with full article input. The result of ablation study with full article input is referred to as (+Fully_sft+Full_Input). All results of ablation study are presented in Table 3. Applying more textual input information into MLLMs may be a simple method to increase the performance since MLLMs have the ability to deal with long input texts. However, we conduct experiments with training and testing both with full articles input on NYTimes800k dataset, but the performance just rises from 82.70 into 83.17, which is not that large while longer textual input makes the time of training and testing nearly doubled. Besides, we conduct an extra experiment by testing the model with oracle selected sentences as textual input, which averagely contain 14% words of whole news articles. The oracle sentences are selected from the news articles and they contain the same entities as the caption. Only a small amount of oracle textual input can achieve a competitive performance. So, compared with simply expanding the length of textual input, selecting related sentences is a better solution.

5.4 Results of Entity Generation

According to (Zhang and Wan, 2023), the generation of entities which occur both in textual input context and captions can better evaluate the ability of models on entity information. We further analyze the Recall of Entities in and out of input textual context. The Recall of entities of caption in the textual input context is more important in news image captioning task, because generating entities

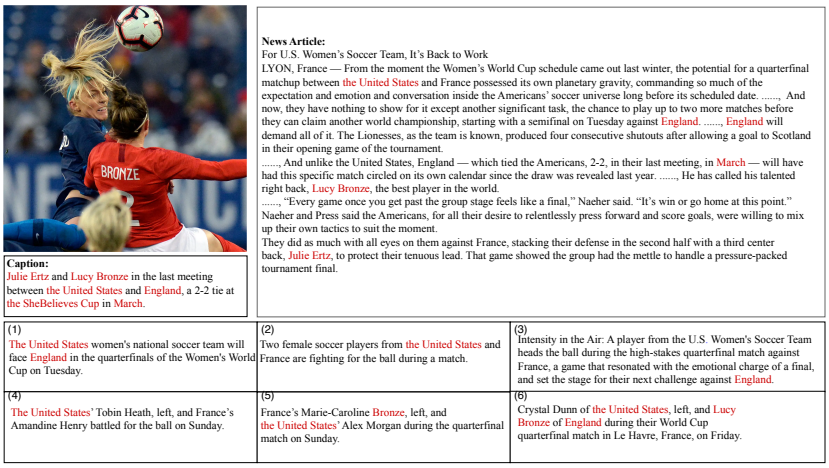

Figure 2: An example of news image caption generation. We present the news image, news article and the news image caption of one sample from NYTimes800k dataset. The news image caption is respectively, generated in: (1)InstructBLIP in zero-shot setting. (2)LLaVA-v1.5 in zero-shot setting. (3)GPT-4V in zero-shot setting. (4)InstructBLIP finetuned on NYTimes800k. (5)LLaVA-v1.5 Fully finetuned on NYTimes800k. (6)Our method. We highlight Entities occurred in reference in red.

which are not in the textual input context is not only difficult (Zhao et al., 2019) but could also let model generate hallucinated responses frequently. Without guidance of directly related textual information, MLLMs may have hallucinations and generate wrong entities which are not related to the news at all according to Zhang and Wan (2023).

All results are listed in Table 4, MLLMs obtain higher Recall especially for entities of caption occurring in the textual input context. Our method obtains the highest Recall of entities appearing in textual input context, which is consistency with the result in Table 2.

Model	Named entities	
	In	Out
Tell*	38.22	4.56
JoGANIC*	38.83	4.88
LLaVA-v1.5	45.55	8.49
InstructBLIP	48.81	**11.23**
Ours	**50.06**	10.25

Table 4: The Recall of Entities in and out of textual input context, all models except Ours task loc_context as textual input, Ours takes mixed context as textual input. We cite the results of Tell and JoGANIC from (Zhang and Wan, 2023)

5.5 Case Study

We randomly select a sample from NYTimes800k dataset, the generated captions of different models are presented on Figure 2. From the generations, we can see that MLLMs in zero-shot setting can only generate the common entities information such as "The United States" or "England" with news image and associated text given. Even the most powerful MLLM GPT-4V(OpenAI et al., 2023) can not generate the specific entities which are related to the news image and occurred in news articles such as "Julie Ertz" and "Lucy Bronze". After finetuned on NYTimes800k dataset, InstructBLIP still cannot generate the names of people in the news image. LLaVA-v1.5 generates the "Bronze" successfully but ignores "England". Our method generates both common entities and names of people occurred in the news image caption and news article, further proving the effectiveness of our method in entity recognition.

6 Conclusion

In this work, we first point out that MLLMs trained on common mutlimodal data are not good at generating entities on news image captioning task. Simply finetuning MLLMs are not enough to let MLLMs learn to operate the multimodal information of entities. We come up with two Entity-Aware multimodal alignment tasks and a new framework to align the MLLM and then use the aligned MLLM to refine the textual input context, since the generated captions are sensitive to related and concise input texts. Our method not only achieves the best performance on automatic evaluation metrics, but also owns the best ability to generate the entities on news image captioning task.

7 Limitations

We conduct our experiments mainly on two common MLLMs. The conclusion will be more general if we test more emerging MLLMs as well. Our method requires big GPU memory both on training and testing stage, and are slower than previous models. With the Optimization technique such as flash-attention provided, the speed of our method will become faster. The data of our alignment tasks are extracted only from news image captioning tasks. The knowledge of news is updating rapidly. Our method still need to align the model on data which contains new updated knowledge to obtain the best performance.

8 Ethics Statement

The generation of our method is highly related to the pretrained multimodal checkpoints and the alignment task data. They may both introduce bias and hallucination into the generated captions which may cause inaccuracy news report. We recommend to manual check the results before actually using in the real situation.

References

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

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)*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. 2023a. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

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.

OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela

Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. Gpt-4 technical report.

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.

Tingyu Qu, Tinne Tuytelaars, and Marie-Francine Moens. 2023. Visually-aware context modeling for news image captioning. *arXiv preprint arXiv:2308.08325*.

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.

Abisek Rajakumar Kalarani, Pushpak Bhattacharyya, Niyati Chhaya, and Sumit Shekhar. 2023. "let's not quote out of context": Unified vision-language pretraining for context assisted image captioning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 695–706, Toronto, Canada. Association for Computational Linguistics.

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)*.

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.

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.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Jingjing Zhang, Shancheng Fang, Zhendong Mao, Zhiwei Zhang, and Yongdong Zhang. 2022. Fine-tuning with multi-modal entity prompts for news image captioning. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 4365–4373, New York, NY, USA. Association for Computing Machinery.

Junzhe Zhang and Xiaojun Wan. 2023. Exploring the impact of vision features in news image captioning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12923–12936, Toronto, Canada. Association for Computational Linguistics.

Sanqiang Zhao, Piyush Sharma, Tomer Levinboim, and Radu Soricut. 2019. Informative image captioning with external sources of information. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6485–6494, Florence, Italy. Association for Computational Linguistics.

Mingyang Zhou, Grace Luo, Anna Rohrbach, and Zhou Yu. 2022a. Focus! relevant and sufficient context selection for news image captioning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6078–6088, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mingyang Zhou, Licheng Yu, Amanpreet Singh, Mengjiao Wang, Zhou Yu, and Ning Zhang. 2022b. Unsupervised vision-and-language pre-training via

retrieval-based multi-granular alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16485–16494.

A Training and Inference Details

The image resolution is fixed at 224 * 224 on all experiments. We restricted the multimodal input tokens length to 1024 and the length of output new tokens to 50. We use the official scripts for InstructBLIP and LLaVA-v1.5 training with their default hyper-parameter configuration which both set initial learning rate of $2e$ -5.

We use a beam size of 5 in inference stage in on InstructBLIP architecture and follow default decoding configuration for the rest MLLMs. We use mixed bp16 precision training to accelerate the training progress. Our model takes 2-3 days to train on four Nvidia A40 GPUs.