Show, Interpret and Tell: Entity-Aware Contextualised Image Captioning in Wikipedia

Khanh Nguyen, Ali Furkan Biten, Andres Mafla, Lluis Gomez, Dimosthenis Karatzas

Computer Vision Center, Universitat Autònoma de Barcelona, Barcelona, Spain
{knguyen, abiten, andres.mafla, lgomez, dimos}@cvc.uab.es

Abstract

Humans exploit prior knowledge to describe images, and are able to adapt their explanation to specific contextual information, even to the extent of inventing plausible explanations when contextual information and images do not match. In this work, we propose the novel task of captioning Wikipedia images by integrating contextual knowledge. Specifically, we produce models that jointly reason over Wikipedia articles, Wikimedia images and their associated descriptions to produce *contextualized* captions. Particularly, a similar Wikimedia image can be used to illustrate different articles, and the produced caption needs to be adapted to a specific context, therefore allowing us to explore the limits of a model to adjust captions to different contextual information. A particular challenging task in this domain is dealing with out-of-dictionary words and Named Entities. To address this, we propose a pre-training objective, Masked Named Entity Modeling (MNEM), and show that this pretext task yields an improvement compared to baseline models. Furthermore, we verify that a model pre-trained with the MNEM objective in Wikipedia generalizes well to a News Captioning dataset. Additionally, we define two different test splits according to the difficulty of the captioning task. We offer insights on the role and the importance of each modality and highlight the limitations of our model.

Introduction

Scene understanding involves composing a story that explains the perceptual observations. This ability, to draw from previous experience to explain what is happening, is termed image interpretation (Lake et al. 2017). Image interpretation is one of the hallmarks of linguistic intelligence (Gardner 2011) where contextual information is drawn upon to compose an explanation of the depicted events (Terman 1916). We identify Wikipedia as an excellent probing ground to further the advance in image interpretation and integrate contextual information into captioning models.

Wikipedia is an invaluable source of world knowledge information. Each page consists of different information sources comprised of title, sections, images and their captions. Each author of Wikipedia writes an article and illustrates it with images that are selected from Wikimedia. Each image has a corresponding description that includes very detailed information, including Named Entities (NEs). In order to write a good caption for the image while remaining relevant to the article, the writers need to take into account the section they wrote, the image they selected and its corresponding description to produce a contextualized caption, a task referred to as *Wikipedia captioning*. In this work, we propose Wikipedia Captioning as a real use case scenario to alleviate the work load of Wikipedia writers while allowing them to better illustrate their articles. This task offers the unique opportunity to potentially aid thousands of writers directly and millions of readers indirectly on a daily basis.

The importance of the context is paramount given the possibility of having several matching and distinct captions for a single image. Naturally, this introduces numerous challenges. On one hand, the context (description and section) needs to be encoded and information selectively drawn from it while being guided by the visual content. On the other hand, explicit contextual information, typically found in the form of NEs such as proper names, prices, locations, dates, etc., needs to be properly injected in the generated captions in natural language. This explicit contextual-dependant information is typically out-of-dictionary or at best it is underrepresented in the statistics of the dictionary used. To better prepare a model that deals with Named Entities, we propose a pre-training strategy in which we extend Masked Language Modeling (MLM) (Devlin et al. 2018) to Masked Named Entity Modeling (MNEM). We show that MNEM is more effective at selecting the correct NE given an article while reducing the language prior produced by NEs.

Wikipedia Captioning is closely related to News Captioning (Biten et al. 2019; Feng and Lapata 2013; Ramisa et al. 2018). In the case of News Captioning, a single source of context is available in which the news article is roughly equivalent to a Wikipedia article, and no separate (article independent) description for the image is available. Yet in Wikipedia Captioning, images are associated to specific sections of the article, plus specific article-independent descriptions for each image are also given. The two tasks would be equivalent if we ignore the descriptions and we equate a Wikipedia section with the whole news article. Given the similarities between the two tasks, we examine whether the performance can be improved by transferring from Wikipedia to News captioning. We discover that the knowledge acquired by the model on Wikipedia captioning is transferable to News captioning, achieving better results over the baseline. Moreover, we demonstrate that MNEM is effective as well on News captioning, outperforming a base model in terms of image captioning metrics.

Furthermore, we uncover several key insights on the Wikipedia captioning task and on the capabilities of the proposed methodology. Firstly, since such models are prone to take shortcuts (Geirhos et al. 2020), we discover a bypass strategy that our model is taking, a copying process of NEs from descriptions to captions. Additionally, humans have the ability to come up with a caption for any image given a specific context (article, section, description, etc.). Contrary to open-world story generation (Martin et al. 2018) which is to come up with stories for any topic without prior knowledge, context-based story-telling is to generate captions with context acting as a prior knowledge for any given image. Trying to emulate this specific skill, we examine if our model can generate a contextualized caption given mismatched but potentially relevant images. Despite our model's state-of-the-art performance according to captioning metrics and injecting the most relevant NEs, we reveal that current models are still far off from endowing this skill. To summarize, the key contributions of our work are:

- We introduce the Wikipedia Captioning as a real use case scenario to benefit writers and readers on a daily basis.
- We propose an architecture-agnostic pretext task, namely Masked Named Entity Masking, that can be applied to any contextualized captioning model. We empirically show that when it is used, a gain in performance is achieved on Wikipedia and News captioning domains.
- We exhibit that pre-training models on Wikipedia captioning boosts performance on News captioning, transferring the knowledge from one domain to another.
- We extensively study the limitations of our models and discuss a setting that shows we are far off on mimicking the human ability of context-based story-telling.

Related Work

Classic Image Captioning. Early image captioning methods (Farhadi et al. 2010; Ordonez, Kulkarni, and Berg 2011) focus on hand crafted features with templates. The advent of deep learning with large scale datasets (Lin et al. 2014), in contrast, steers the trend into end-to-end encoder-decoder networks. Centralized by attention mechanisms (Bahdanau, Cho, and Bengio 2014; Devlin et al. 2018), architectures, ranging from LSTM-based (Vinyals et al. 2015; You et al. 2016; Anderson et al. 2018; Lu et al. 2017) to Transformers (Pan et al. 2020; Luo et al. 2021; Ji et al. 2021; Chen et al. 2021) achieve remarkable progress in captioning metrics. Advanced models (Dai et al. 2017; Shetty et al. 2017) employ GANs to favor human-like caption outputs. Works applying Reinforcement Learning (Rennie et al. 2017; Liu et al. 2017) successfully optimize models on non-differential metrics. Nevertheless, such models are restricted by design to the mere description of what is shown, thus unable to attempt any interpretation.

News Articles Captioning. Recent advances in News image captioning (Feng and Lapata 2013) come in both architectures and datasets. Typically, proposed models follow the standard encoder-decoder paradigm while large-scale datasets such as Breaking News (Ramisa et al. 2018), GoodNews (Biten et al. 2019), VisualNews (Liu et al. 2020) etc. have been contributed. To deal with NEs, (Biten et al. 2019; Liu et al. 2020) propose two-stage template-based methods that train models to capture NE-tag distribution followed by post-generation NE insertion. Closely related to our work, (Tran, Mathews, and Xie 2020) employs a transformer (Vaswani et al. 2017) with byte-pair-encoding (BPE) (Sennrich, Haddow, and Birch 2015) to directly decode NEs. Lastly, (Yang et al. 2021) enhances caption quality with template guidance and a pre-trained NE Embedding module based on Wikipedia Knowledge graph (Yamada et al. 2018) to better encode NEs. Given the importance of NEs in this task, we introduce the Named Entity Masking pretext task for accurate NE prediction and coverage.

Self-supervised Pre-training Methods. Self-supervised approaches in NLP (Devlin et al. 2018; Radford et al. 2019; Roberts and Raffel 2020) have greatly inspired many works in vision-and-language (VL) (Li et al. 2020; Kim, Son, and Kim 2021; Zhang et al. 2021) that follow the pre-training and fine-tuning paradigm, elevating the performance in a wide variety of benchmarks. While existing pre-training objectives in VL either focus on representation learning or multi-modal alignments, our proposed method MNEM aims to learn context-entity alignment, which is essential for contextual caption generation. Compared to other entity-centric methods (Sun et al. 2020; Lin et al. 2021), we are the first to explore MNEM in a multi-modal scheme, with careful adaptations to suit the image-text distribution in Wikipedia.

Method

In this section, we first define the Wikipedia captioning task formally and then present our approach which utilizes well-designed transformer-based models to solve the task. We further discuss how we extend these models by employing a novel pre-training objective which we refer to as Masked Named Entity Modeling (MNEM).

Problem Definition

We refer to $I = \{i_1, i_2, ...\}$, $C = \{c_1, c_2, ...\}$ as the image set and its corresponding captions set where each caption contains words defined as c_{k_j} or $c_k = \{c_{k_0}, c_{k_1}, ...\}$. Moreover, $D = \{d_1, d_2, ...\}$ and $S = \{s_1, s_2, ...\}$ refer to description and section sets, respectively. Given a Wikipedia article, a description d_k is tied to each specific image i_k for which it provides Named Entities (when and where the photo was taken, who is depicted in the photo, author, etc.). Each Wikipedia article is divided in sections s_k to enhance readability. To emulate the process that the Wikipedia writers go through during Wikipedia captioning, we define the dataset with quadruples where (i_k, c_k, s_k, d_k) refers to the k^{th} datum. We define the task of generating a caption that is contextualized as $\mathcal{P}(c_k|i_k, s_k, d_k)$.

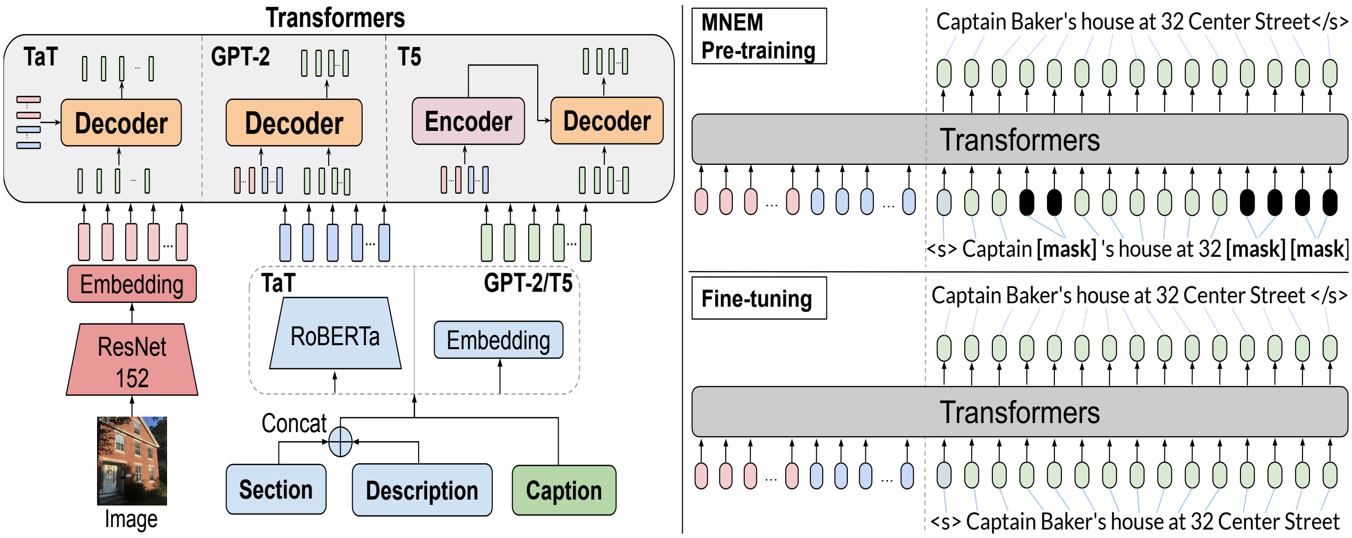

Figure 1: Overview of the Transformer-based models and our MNEM pre-training task. Left: Transformers take multi-modal input features (image, section, description) processed by single-modal encoders. Right: During pre-training (right-top) we replace the NEs in the ground-truth caption by the *MASK* token and train our models to reconstruct the caption. Best view in colors.

Transformer-based Models

The overview of our proposed method is depicted in Figure 1, which follows the commonly used encoder-decoder paradigm. In particular, we investigate the capacity of Transformer-based models, either trained from scratch or initialized with pre-trained weights acquired from self-supervised training. For the latter, we explore Pre-trained language models (PLMs) due to its compelling performances on numerous open-ended text generation tasks. However, as only pre-trained on textual modality, properly adapting PLMs to solve VL tasks remains a challenge. Therefore, we propose a simple strategy to enable PLMs to reason on multi-modal context when generating captions.

Transform and Tell (TaT) (Tran, Mathews, and Xie 2020). Proposed to address the News Image Captioning task in an end-to-end manner, TaT consists of a set of frozen encoding modules followed by a trainable Transformer-decoder. The decoder thus generates captions conditioned on multi-modal input features extracted from the image and the article. Perceiving the similarities between the two tasks, we apply TaT to Wikipedia captioning and modify it by removing the MTCNN (Zhang et al. 2016) for face detection, FaceNet (Schroff, Kalenichenko, and Philbin 2015) for face embedding, YOLOv3 (Redmon and Farhadi 2018) for object representation and the location embedding, thus speed up the training process by decreasing complexity.

The reduced encoder now comprises two pre-trained modules for image and text. More specifically, given an image i_k, we extract image features using a ResNet-152 model pre-trained on ImageNet to obtain (f_{i_k}). We use RoBERTa (Liu et al. 2019) to encode text sequences, yielding two sets of text features for sections and descriptions (f_{s_k}, f_{d_k}) respectively. The auto-regressive decoder consists of the same BPE embedding matrix introduced in RoBERTa, followed by a stack of N identical transformer blocks. For each transformer block, we use a Dynamic Convolution (DC) (Wu et al. 2019) layer for past-token conditioning and a Multi-Head Attention (MHA) (Vaswani et al. 2017) layer for multi-modal conditioning. This architecture design serves as an efficient alternative to the traditional self-attention mechanism, relaxing memory requirements which are quadratic with the input sequence length. At time step t, previously-generated token representations $Z_{\leq t} = \{z_0, z_1, ..., z_t\}$ are efficiently enriched by attending to its leftward context thanks to the DC. After going through the decoder stack, the final representation $z'_t \in Z'_{\leq t}$ is fed to the vocabulary layer to predict the next token at $t+1$.

GPT-2 (Radford et al. 2019) is a transformer decoder variant which performs text generation in auto-regressive manner. We want to leverage the linguistic knowledge stored in the model's pre-trained weights to generate captions while preserving the Wikipedia style. We therefore decide to fine-tune the entire network, including the learned embedding matrix. We start by transforming the description and section sequences into continuous embeddings (f_{s_k}, f_{d_k}) using the internal embedding layer of GPT-2. Concurrently, we represent the image as a set of visual features (f_{i_k}) extracted from ResNet-152, followed by a MLP layer with Layer Normalization to map images into the word embedding space. We form this set of representations (f_{i_k}, f_{s_k}, f_{d_k}) as prefix and pass it to the decoder (referred as GPT-2_{BASE}++) whose output $z'_t \in Z'_{\leq t}$ is used as a logit vector over the vocabulary distribution to predict the caption token at $t+1$.

T5 (Roberts and Raffel 2020). Built on top of the vanilla Transformer encoder-decoder, T5 is another language model pre-trained on C4. Particularly, T5 presents a new self-supervised training objective called *Replace Corrupted Spans* for efficient computation conforming to its encoder-decoder design. To apply the off-the-shelf version of T5 to Wikipedia captioning, we extend its text-only encoder to multi-modal encoder, referred as T5_{BASE}++. We incorporate image features f_{i_k} as additional input to the description and section embeddings (f_{s_k}, f_{d_k}), with each set of embeddings computed the same way as in the previous section. The decoder then attends to previous caption tokens $Z_{\leq t}$ via self-attention and the contextualized joint representations from the encoder via cross-attention to predict the next token.

Masked Named Entity Modeling

Referring to the right NEs is among the most important but challenging tasks of Wikipedia captioning to assure the veracity of information. MLM, as proposed in BERT (Devlin et al. 2018), has shown remarkable success, building richer representations by exploiting the distributional hypothesis present in language to reconstruct masked out words. To extend the applicability of MLM to our captioning task and mitigate the rare NEs insertion issue, we apply the masking strategy explicitly on NEs to model the context-NE relationship. More formally, let $\mathcal{W} = \{w_1, w_2, ..., w_n\}$ be the set of all byte-pair tokens from the caption. Now, let $\mathcal{M}_l = \{j, j+1, ..., j+k\}$ be the l^{th} mask span where j is the starting index to mask such that $\max(\mathcal{M}_l) < \min(\mathcal{M}_{l+1})$. Then, $\{w_j, ..., w_{j+k}\}$ are replaced with the special mask token [MASK] to form the corrupted caption $\tilde{\mathcal{W}} = \{\tilde{w}_1, \tilde{w}_2, ..., \tilde{w}_n\}$. We use cross-entropy to reconstruct the original caption $\mathcal{W}$, with *no restriction to masked tokens* compared to BERT, in order to retain the generation ability:

$$\begin{aligned} \tilde{w_i} &= [MASK], \text{where } i \in \mathcal{M}_l \\ \mathcal{L} &= \sum_{i=0}^{N} -\log p(w_i | \tilde{\mathcal{W}}_{0<i}, f_{i_k}, f_{s_k}, f_{d_k}) \end{aligned} \tag{1}$$

Inspired from T5, we introduce a small modification to MNEM to specifically pre-train encoder-decoder architectures. We first substitute each entity span $\mathcal{M}_l$ with *a unique mask* token to create $\tilde{\mathcal{W}}$ then append it to the encoder input sequence. Consequently, instead of reconstructing the caption, we task the decoder to predict all the masked name entities, which are concatenated in original order as the target sequence. This conforms the training procedure performed on T5 while explicitly tells the model to focus on name entities within context.

Figure 1 depicts our proposed design with masked NEs language modeling during the pre-training stage. We extract NEs contained in captions and randomly mask the NEs with a probability of 0.8 while adopting a whole word masking strategy. We only mask NEs that fall into the categories of: [*Person*, *ORGanization*, *GeoPolitical Entity*] as they appear to be dominant types in Wikipedia captions. Moreover, changing every NEs might be too disrupting of the semantic structure in captions which tend to be short.

By masking the NEs within the captions, the model can learn to precisely capture the contextual information relating to the NEs to produce useful representations. Also, our proposed method aims to mitigate another issue, which is further studied below. We found that when there is a high overlap between the captions and context, the model learns to directly copy NEs to generate captions, making it ineffective to deal with the diversity and semantic richness of Wikipedia Captioning. By limiting the input caption to NE-masked sentence in the pre-training phase, we decrease the influence of the language bias, allowing a better utilization of the entire context given as input.

	Model	B	C	Pr	Re
No Image	Seq2seq	6.04	137.38	26.42	14.05
	Seq2seq+Attn.	15.86	242.87	**44.10**	27.60
	Seq2seq+Ptr.	21.04	255.87	40.81	30.67
	Seq2seq+Ptr.+Cvrg.	21.57	258.84	41.23	31.37
	TaT	18.74	229.77	29.47	27.12
	GPT-2_{BASE}	19.18	254.94	42.28	30.38
	T5_{BASE}	**25.70**	275.31	39.11	**35.09**
	$\text{TaT}^{\dagger}$	19.93	235.9	332.94	28.47
	$\text{GPT-2}^{\dagger}_{BASE}$	19.64	260.4	43.31	30.9
	$\text{T5}^{\dagger}_{BASE}$	24.60	**278.68**	40.51	34.26
	Δ TaT	1.19↑	6.13↑	0.85↑	1.35↑
	Δ GPT-2_{BASE}	0.46↑	5.46↑	1.03↑	0.52↑
	Δ T5_{BASE}	1.1↓	2.08↑	1.4↑	0.83↓
Image	Seq2seq	6.10	136.98	25.97	13.88
	Seq2seq+Attn.	14.71	235.44	**43.49**	26.60
	Seq2seq+Ptr.	20.46	252.79	40.37	29.97
	Seq2seq+Ptr.+Cvrg.	20.22	249.67	40.20	29.68
	TaT	24.47	267.37	32.10	34.64
	GPT-2_{BASE}++	19.4	256.25	41.28	30.06
	T5_{BASE}++	23.83	276.6	40.7	33.72
	$\text{TaT}^{\dagger}$	25.31	272.46	33.34	35.56
	$\text{GPT-2}^{\dagger}_{BASE}$++	20.65	266.29	42.86	31.48
	$\text{T5}^{\dagger}_{BASE}$++	**25.89**	**279.09**	39.65	**35.73**
	Δ TaT	0.84↑	5.09↑	1.24↑	0.92↑
	Δ GPT-2_{BASE}++	1.25↑	10.04↑	1.58↑	1.12↑
	Δ T5_{BASE}++	2.56↑	3.78↑	1.05↓	2.01↑

Table 1: Performance of proposed models on Wikipedia Captioning. Seq2seq performs better than TaT without image while MNEM is effective to align the image and text modality. Metric scores are reported as percentage. † indicates that MNEM pre-training is performed. Δ indicates the improvement with MNEM compared to respective models. Bold numbers are the best performance of each metric.

Experiments

In our experimentation[1] for Wikipedia Captioning, we utilize the WIT (Srinivasan et al. 2021) dataset. All the details regarding the dataset statistics, pre-processing and implementation as well as an in-depth explanation of our baselines can be found in the Appendix[2].

[1]The code, models and data splits are publicly available at https://github.com/khanhnguyen21006/wikipedia_captioning

[2]Full version of this paper can be found on https://arxiv.org/abs/2209.10474

Wikipedia Captioning

Table 1 showcases the results for Wikipedia Captioning in BLEU-4 (Papineni et al. 2002), METEOR (Banerjee and Lavie 2005), ROUGE-L (Lin 2004), CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015), and SPICE (Anderson et al. 2016), as well as precision and recall of the NE insertion as defined by (Biten et al. 2019). The high performing baselines, especially without image, imply strong language priors learned by the models, outperforming transformer language models. The observed drawbacks of such methods indicate two things: (1) the baselines tend to rely on copying NEs excessively coming from the description only and (2) the models prefer to generate few NEs. These hypotheses are confirmed by the high Precision and low Recall, suggesting the models have found a shortcut to generate few NEs by only copying them from the description. Moreover, the addition of images as an input clearly disrupts the performance. According to our examination, we find that the attention weights of seq2seq models are significantly lower for the image than the context. In other words, seq2seq tries to disregard the images while putting more emphasis on the context, failing to correspond these two modalities.

Yet, we see opposite behaviour on the efficacy of images on transformer models, showcasing the superiority of transformers over LSTM encoder-decoder. The addition of images significantly improves the performance on all the captioning metrics. Interestingly, images have a direct impact on the Recall for TaT, improving the results from **27%** to **34%** while not affecting the Precision score. However, this effect is not present in GPT-2 and T5, which is expected, as these PLMs naturally gravitate towards linguistic features, thus require more well-designed adaptation to balance the visual and textual information in VL tasks. MNEM aims to connect two modalities by means of predicting NEs, leading to better NEs selection given multi-modal alignment. It is clearly observed that in all transformer models, MNEM exceeds the performance significantly on captioning metrics (**+10** and **+5** points in CIDEr for GPT-2 and TaT respectively), as well as on Precision and Recall with or without image. Finally, T5 shows a consistent superiority among all models, with MNEM it achieves state-of-the-art in Wikipedia Captioning.

Transfer Learning from Wikipedia to News

We examine the possibility of transfer learning from Wikipedia to News captioning, *i.e.* fine-tuning models on GoodNews with weights **initialized from pre-training on Wikipedia**. To imitate the News captioning setting, we utilize the section as the only context, disregarding the description. We select the T5$_{BASE}$++ model for the experiments given that it is the best model in Wikipedia Captioning. Table 2 shows the results of the transferability from Wikipedia to News with MNEM pre-training. At first, simply fine-tuning T5$_{BASE}$++ on GoodNews can match or surpass previous transformer models (**+6** points in CIDEr compared to TaT), showing PLMs impressive ability to generate captions. Pre-training on Wikipedia further improves the performance over JoGANIC, which is the state-of-the-art method with

	Wiki		News		B	C	Pr	Re
	†	Ft	†	Ft				
TaT				✓	6	53.1	21.3	19.7
TaT*				✓	6.05	53.8	21.85	19.93
VN.*				✓	6.1	55.4	–	–
JoG.				✓	6.34	59.19	–	–
JoG.*				✓	6.83	61.22	–	–
T5$_B$++				✓	6.34	59.68	23.61	20.08
			✓	✓	6.56	60.17	23.28	20.43
		✓		✓	7.17	62.76	23.11	20.92
		✓	✓	✓	**7.19**	**63.89**	**23.37**	**21.48**
	✓			✓	5.89	51.49	20.36	18.85
	✓		✓	✓	6.63	57.03	21.51	20.43
	✓	✓		✓	6.66	58.47	21.84	20.59
	✓	✓	✓	✓	6.75	59.07	21.89	20.55

Table 2: Performance of our proposed method on GoodNews compared to other transformer-based models such as VisualNews (VN.) and JoGANIC (JoG.). $*$ denotes models employing all of its designed modules. † indicates that MNEM pre-training is performed. Ft means fine-tuning the models with the captioning objective in the corresponding dataset. Note that for captioning metrics, we directly use the results that are reported in the original papers.

Wikipedia pre-training, (**+0.36** points in BLEU-4 and **+2.67** in CIDEr) while not including specific template guidance. This illustrates the advantage of proposing this **Wikipedia Captioning** task, in which the acquired knowledge is useful for the News domain. Moreover, incorporating MNEM pre-training on GoodNews constantly results in additive gains, demonstrating its effectiveness in this task. However, the improvement is not transferable as we additionally perform MNEM on Wikipedia (rows 9–12). This extra pre-training stage guides the model to focus on Wikipedia NEs, which come from a different distribution compared to GoodNews, thus degrading the final model performance.

Ablation Study on MNEM

To further probe into the efficacy of MNEM, we compare it against the well established MLM introduced by BERT. The MLM strategy selects 15% of the words in a sentence in which it replaces WordPiece tokens (Wu et al. 2016) 80% of the time with a special [MASK] token, 10% of the time with a random word and 10% of the time without changing the word. We introduce another variant of MLM as a comparison namely Full Masking where given 15% of the words in a sentence, we always mask a word without employing other cases. We perform the pre-training stage with these masking strategies on transformer models and report results on Wikipedia Captioning in Table 3.

We note that our newly defined Full Masking strategy outperforms the well established MLM, challenging the hyper-parameters defined by MLM. Our experiments demonstrate that masking hyper-parameters are rather task dependent which need to be treated selectively. Overall, MNEM yields a significant boost over its counterparts in TaT (**+12** points

Model			B	C	Pr	Re
GPT-2$_B$++	W	Full	20.44	**266.4**	43.24	31.4
		MLM	20.34	266.	**43.4**	31.4
		MNEM	**20.65**	266.29	43.12	**31.55**
T5$_B$++	W	Full	25.02	276.58	38.87	35.08
		MLM	24.57	275.22	38.75	34.94
		MNEM	**25.89**	**279.09**	**39.65**	**35.73**
TaT	W	Full	24.01	260.87	31.83	34.35
		MLM	23.81	258.25	31.47	34.12
		MNEM	**25.31**	**272.46**	**33.34**	**35.56**
	S	Full	8.96	147.43	20.02	15.13
		MLM	8.71	147.40	19.81	14.46
		MNEM	**9.54**	**153.87**	**20.75**	**16.01**
	D	Full	20.95	215.42	28.80	26.55
		MLM	20.92	213.69	28.41	26.16
		MNEM	**22.36**	**224.38**	**29.36**	**27.90**

Table 3: Ablation studies on different masking strategies. W, D, S denote Wiki, Description, Section as context respectively. MNEM outperforms MLM and Full masking strategy on every context and on every metrics.

in CIDEr), while this behaviour is less apparent in T5 and GPT-2. This is justified since TaT is specifically designed for News Image Captioning, highlighting the effectiveness of MNEM in a fine-grained multi-modal scheme. In contrast, T5 and GPT-2 essentially are language models that have been pretrained on text thus require more careful adaptations in architecture for further improvements.

To examine how the context affects the masking strategies, we carry out experiments with TaT, with the context defined in three distinct ways: *description*, *section* and the original *Wiki context*. We observe that MNEM again outperforms other methods on every context type. Notably, the difference between MNEM and the other two masking strategies is much higher with the description compared to section used as a context. This is somehow expected given that a section tends to be longer and it is harder to generate contextualized captions. However, we observe that MNEM is especially effective when section and description are combined (Wiki as context). MNEM performs **2%** better in Recall in Wiki context while improving **1%** more in Recall compared to MLM or Full Masking with other context types, this behaviour is more apparent in CIDEr. This demonstrates that MNEM is not only better on encoding the contextual information but especially better at combining description and section information. We believe the potential of MNEM is unlocked further because it can amalgamate the two contexts better by discovering strong correlations between description and section, resulting in better alignments for NEs.

Analysis of Wikipedia Captioning

In this section, we investigate the performance of our models in Wikipedia Captioning and provide an in-depth analysis in both quantitative and qualitative results. Furthermore, we examine if our models have the ability to generate a caption for "irrelevant" images given a context. This evaluates how close our model performs in terms of generating captions for any given image while staying faithful to the section.

Model	Test Set	GT ol.	Gen. ol.	B	C
TaT	Easy	0.81	0.92	29.93	313.99
	Hard	0.35	0.75	8.48	133.68
	Full	0.71	0.88	24.47	267.37
TaT†	Easy	0.81	0.92	30.76	319.01
	Hard	0.35	0.73	8.91	137.81
	Full	0.71	0.87	**25.31**	**272.46**
GPT-2$_B$++	Easy	0.81	0.93	24.74	304.71
	Hard	0.35	0.81	6.38	128.15
	Full	0.71	0.90	19.4	256.25
GPT-2$_B^\dagger$++	Easy	0.81	0.93	25.63	312.16
	Hard	0.35	0.8	7.01	133.64
	Full	0.71	0.89	**20.65**	**266.29**
T5$_B$++	Easy	0.81	0.93	29.55	325.32
	Hard	0.35	0.8	7.46	130.79
	Full	0.71	0.9	23.83	276.6
T5$_B^\dagger$++	Easy	0.81	0.91	31.96	330.79
	Hard	0.35	0.79	7.95	129.84
	Full	0.71	0.88	**25.89**	**279.09**

Table 4: Performance on Easy and Hard subsets. They are defined based on the Jaccard similarity between the context and caption. Overlap scores measure the similarity compared to the context, reported in average Jaccard score. We refer the (ground truth caption, context) overlap as GT ol. and Gen. ol. is the (generated caption, context) overlap. † indicates that MNEM pre-training is performed.

Easy vs Hard Samples. Previously, we show in Table 1 that seq2seq models exploit a shortcut of copying the NEs from the context, reflected in the high precision at inserting NEs. Hence, to better understand our model's limitations, we devise a procedure to identify the captions that can be produced simply by utilizing the context. To this end, we employ the Jaccard similarity (Jaccard 1912) at the word level where given two sets $\mathcal{C}, \mathcal{S}$ for caption and context respectively composed of word tokens, we calculate $J(\mathcal{C}, \mathcal{S}) = \frac{|\mathcal{C} \cap \mathcal{S}|}{|\mathcal{C} \cup \mathcal{S}|}$. Accordingly, we assume the image has a diminishing impact on the caption when the caption and the context have a Jaccard score of higher than 0.5. Thus, we refer to these samples as *easy* samples (*hard* otherwise) since the caption can be generated simply by copying from the context and ignoring the image.

Table 4 presents the scores in terms of Easy and Hard subset, as well as the average Jaccard score between the caption output and the context. We observe that the average GT overlap between two subsets (GT. ol.) is quite disparate, yet we do not see a similar distribution learned by our models in generated overlap (see Gen. ol. in Table 4). The models instead have higher overlap scores in both Easy and Hard samples (see Gen. ol.) compared to GT overlap. *This highlights the first limitation of our model where the copying shortcut is exploited and learnt through the Easy subset while being carried over to the Hard samples.* This behaviour is also apparent in the captioning metrics as well as Precision and Recall by the excessive gap between Easy and Hard subset regardless of the models. Although the benefits of MNEM are observable, this method still can not effectively deal with the distribution shift that exists within Wikipedia.

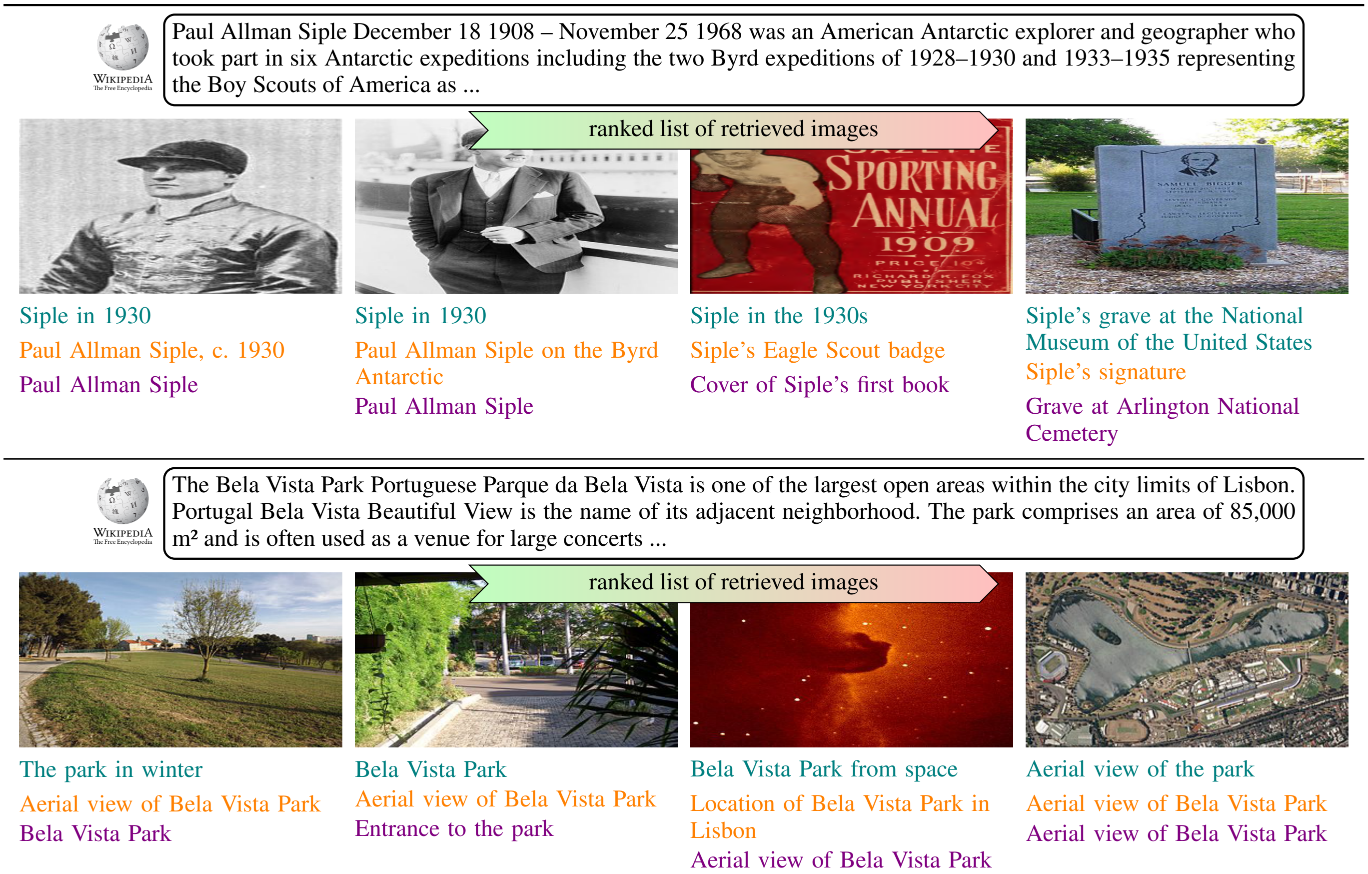

Figure 2: Qualitative samples of generated captions with TAT-MNEM (top, green), GPT2-MNEM (middle, orange) and T5-MNEM (bottom, purple) for unpaired images given a similar Wikipedia article section. The models usually prefer to generate short, repetitive and template-based sentences. However, in some cases, they can integrate relevant visual details into the caption.

Context-based Story Telling. Humans have an innate capability of coming up with a well sounding explanation for an image given any context. To assess this skill in our captioning models, we employed a retrieval system to sort images based on its relevance given a section. Our goal is to evaluate our model's ability to adapt their produced captions based on the context that comes in the form of NEs. In our experiments, we trained two fully-connected layers on top of the visual and textual representations of CLIP (Radford et al. 2021) to obtain ranked images according to the semantic relatedness. In Figure 2, we provide qualitative samples for the studied models with MNEM. We ask the reader to refer to the Appendix for more samples and details. It can be seen that models can generate context-based captions even with irrelevant images to a certain degree. For example, even though the images in 1st and 2nd row, 3rd column are irrelevant to the input section, MNEM models can "hallucinate" better sounding captions that can improve upon the combination of modalities. Also, MNEM models, tend to include more relevant information in the form of NEs, thus integrating relevant incoming information from the section. However, this is not always the case, we can also note that models tend to generate short sentences that are repeated over many images, thus missing some salient details coming from the visual input. All in all, even though improved captions are generated with MNEM, this innate human skill lies far compared to current approaches that favor template-based, short and repetitive sentences, thus opening ways for future work.

Conclusions

In this work, we have proposed the novel task of Wikipedia Captioning. We introduce the importance of the Wikipedia captioning due to the direct impact of it on several applications. We show extensive experimentation to compare our approach with several baselines. By incorporating a novel pretext task MNEM, we improve the caption prediction with better context-entity alignment. We also show that a model pre-trained on Wikipedia generalizes well to the News Captioning task. We find that the models tend to learn a shortcut of directly copying the context to generate a caption. To evaluate this effect, we divide the Wikipedia Captioning test set according to the Jaccard similarity between the context and ground-truth captions. We show that there is still a big gap among these two test splits, which opens the path for future research. Finally, we disclose several limitations of our models and provide insights of the benefits of our approach.

Acknowledgements

This work has received support from Grants PID2020-116298GB-I00 and PLEC2021-007850 funded by MCIN/AEI /10.13039/501100011033 and the project ELSA, 101070617, financed by the European Union's Horizon Europe research and innovation programme.

References

Anderson, P.; Fernando, B.; Johnson, M.; and Gould, S. 2016. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*.

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6077–6086.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Banerjee, S.; and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72.

Biten, A. F.; Gomez, L.; Rusinol, M.; and Karatzas, D. 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*, 12466–12475.

Chen, H.; Wang, Y.; Yang, X.; and Li, J. 2021. Captioning transformer with scene graph guiding. In *2021 IEEE International Conference on Image Processing (ICIP)*, 2538–2542. IEEE.

Dai, B.; Fidler, S.; Urtasun, R.; and Lin, D. 2017. Towards diverse and natural image descriptions via a conditional gan. In *Proceedings of the IEEE international conference on computer vision*, 2970–2979.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Farhadi, A.; Hejrati, M.; Sadeghi, M. A.; Young, P.; Rashtchian, C.; Hockenmaier, J.; and Forsyth, D. 2010. Every picture tells a story: Generating sentences from images. In *European conference on computer vision*, 15–29. Springer.

Feng, Y.; and Lapata, M. 2013. Automatic caption generation for news images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(4): 797–812.

Gardner, H. E. 2011. *Frames of mind: The theory of multiple intelligences*. Hachette Uk.

Geirhos, R.; Jacobsen, J.-H.; Michaelis, C.; Zemel, R.; Brendel, W.; Bethge, M.; and Wichmann, F. A. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11): 665–673.

Jaccard, P. 1912. The distribution of the flora in the alpine zone. 1. *New phytologist*, 11(2): 37–50.

Ji, J.; Luo, Y.; Sun, X.; Chen, F.; Luo, G.; Wu, Y.; Gao, Y.; and Ji, R. 2021. Improving image captioning by leveraging intra-and inter-layer global representation in transformer network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 1655–1663.

Kim, W.; Son, B.; and Kim, I. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, 5583–5594. PMLR.

Lake, B. M.; Ullman, T. D.; Tenenbaum, J. B.; and Gershman, S. J. 2017. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40.

Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, 121–137. Springer.

Lin, C.; Miller, T.; Dligach, D.; Bethard, S.; and Savova, G. 2021. EntityBERT: Entity-centric masking strategy for model pretraining for the clinical domain. Association for Computational Linguistics (ACL).

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *ACL Workshop on Text Summarization*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*.

Liu, F.; Wang, Y.; Wang, T.; and Ordonez, V. 2020. Visual News: Benchmark and Challenges in News Image Captioning. *arXiv preprint arXiv:2010.03743*.

Liu, S.; Zhu, Z.; Ye, N.; Guadarrama, S.; and Murphy, K. 2017. Improved image captioning via policy gradient optimization of spider. In *Proceedings of the IEEE international conference on computer vision*, 873–881.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lu, J.; Xiong, C.; Parikh, D.; and Socher, R. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Conference on Computer Vision and Pattern Recognition*.

Luo, Y.; Ji, J.; Sun, X.; Cao, L.; Wu, Y.; Huang, F.; Lin, C.-W.; and Ji, R. 2021. Dual-level collaborative transformer for image captioning. *arXiv preprint arXiv:2101.06462*.

Martin, L.; Ammanabrolu, P.; Wang, X.; Hancock, W.; Singh, S.; Harrison, B.; and Riedl, M. 2018. Event representations for automated story generation with deep neural nets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Ordonez, V.; Kulkarni, G.; and Berg, T. L. 2011. Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems*, 1143–1151.

Pan, Y.; Yao, T.; Li, Y.; and Mei, T. 2020. X-linear attention networks for image captioning. In *Proceedings of* *the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10971–10980.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W. 2002. BLEU: a method for automatic evaluation of machine translation. *Annual Meeting on Association for Computational Linguistics*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

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

Redmon, J.; and Farhadi, A. 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.

Rennie, S. J.; Marcheret, E.; Mroueh, Y.; Ross, J.; and Goel, V. 2017. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7008–7024.

Roberts, A.; and Raffel, C. 2020. Exploring transfer learning with t5: the text-to-text transfer transformer. *Accessed on*, 23–07.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 815–823.

Sennrich, R.; Haddow, B.; and Birch, A. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Shetty, R.; Rohrbach, M.; Anne Hendricks, L.; Fritz, M.; and Schiele, B. 2017. Speaking the same language: Matching machine to human captions by adversarial training. In *Proceedings of the IEEE International Conference on Computer Vision*, 4135–4144.

Srinivasan, K.; Raman, K.; Chen, J.; Bendersky, M.; and Najork, M. 2021. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2443–2449.

Sun, Y.; Wang, S.; Li, Y.; Feng, S.; Tian, H.; Wu, H.; and Wang, H. 2020. Ernie 2.0: A continual pre-training framework for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 8968–8975.

Terman, L. M. 1916. *The measurement of intelligence: An explanation of and a complete guide for the use of the Stanford revision and extension of the Binet-Simon intelligence scale*. Houghton Mifflin.

Tran, A.; Mathews, A.; and Xie, L. 2020. Transform and tell: Entity-aware news image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13035–13045.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Wu, F.; Fan, A.; Baevski, A.; Dauphin, Y. N.; and Auli, M. 2019. Pay less attention with lightweight and dynamic convolutions. *arXiv preprint arXiv:1901.10430*.

Wu, Y.; Schuster, M.; Chen, Z.; Le, Q. V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Yamada, I.; Asai, A.; Shindo, H.; Takeda, H.; and Takefuji, Y. 2018. Wikipedia2Vec: an optimized tool for learning embeddings of words and entities from Wikipedia. *arXiv preprint arXiv:1812.06280*.

Yang, X.; Karaman, S.; Tetreault, J.; and Jaimes, A. 2021. Journalistic Guidelines Aware News Image Captioning. *arXiv preprint arXiv:2109.02865*.

You, Q.; Jin, H.; Wang, Z.; Fang, C.; and Luo, J. 2016. Image Captioning with Semantic Attention. In *IEEE conference on computer vision and pattern recognition*.

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

Zhang, P.; Li, X.; Hu, X.; Yang, J.; Zhang, L.; Wang, L.; Choi, Y.; and Gao, J. 2021. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5579–5588.