Structural and Functional Decomposition for Personality Image Captioning in a Communication Game

Minh Thu Nguyen[1], **Duy Phung**[1], **Minh Hoai**[2] **and Thien Huu Nguyen**[3]
[1] VinAI Research, Vietnam
[2] Stony Brook University, Stony Brook, New York, USA
[3] Department of Computer and Information Science, University of Oregon, Eugene, OR 97403, USA
{v.thunm15,v.duypv1}@vinai.io,
minhhoai@cs.stonybrook.edu, thien@cs.uoregon.edu

Abstract

Personality image captioning (PIC) aims to describe an image with a natural language caption given a personality trait. In this work, we introduce a novel formulation for PIC based on a communication game between a speaker and a listener. The speaker attempts to generate natural language captions while the listener encourages the generated captions to contain discriminative information about the input images and personality traits. In this way, we expect that the generated captions can be improved to naturally represent the images and express the traits. In addition, we propose to adapt the language model GPT2 to perform caption generation for PIC. This enables the speaker and listener to benefit from the language encoding capacity of GPT2. Our experiments show that the proposed model achieves the state-of-the-art performance for PIC.

1 Introduction

To effectively communicate with human, an important step involves image captioning (IC) that requires systems to describe images using natural language captions. Image captioning (IC) has been studied extensively, featuring deep learning models (i.e., the encoder-decoder architectures) as the dominant approach (Vinyals et al., 2014; Xu et al., 2015; Anderson et al., 2017; Yang et al., 2018). Despite its popularity, the current work on IC has mainly considered the factual setting for IC where the generated captions should faithfully present the visual content of images. A major limitation for this factual IC task concerns its failure to incorporate human factors (i.e., personalities or traits) into the caption generation process. As such, ones might prefer to produce engaging captions where his/her personality traits are explicitly expressed and the visual concepts in the images are not necessarily covered in their full details. Consequently, in this work, we seek to fill in this gap for IC by exploring personality image captioning (PIC) where the models need to further consider a personality/trait in the captioning process. In particular, we leverage PERSONALITY-CAPTIONS (PC) (Shuster et al., 2019), the first dataset for PIC, to evaluate the models in this work.

Which characteristics should a caption have to adequately describe an image in PIC? Motivated by the functional and structural decomposition for language learning (Lazaridou et al., 2016, 2020; Kottur et al., 2017), we argue that an effective caption for PIC should posses two important properties. On the one hand, the captions in PIC should follow the natural language structures to induce effective communication with human (i.e., the structural view or naturalness of the captions). On the other hand, for the functional view, the generated captions from a model should involve sufficient information to enable another system or human to uniquely identify the input images and traits.

In this paper, we propose to achieve these two goals by recasting PIC as a multi-agent communication framework that involves a speaker and a listener (Evtimova et al., 2018; Lazaridou et al., 2018). The speaker attempts to generate a natural language caption for a given image and trait (i.e., for the structural property) while the listener seeks to identify the input images and personality traits based on the generated caption from the speaker (i.e., for the functional property). By training this framework, we expect that the generated captions of the speaker can be regularized to naturally convey the information in the images and express the provided personality traits at the same time. To our knowledge, this is the first work to solve PIC via a multi-agent communication framework.

A bottleneck in the training of the speaker-listener framework concerns the ability to model the language effectively for the captions in PIC. In particular, the speaker would benefit from a high-quality language model that can produce natural captions for PIC while the listener would make

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 4587–4593
November 16 - 20, 2020. ©2020 Association for Computational Linguistics

better predictions for the image and trait identification if it can effectively encode the generated captions. Although ones can attempt to learn those language modeling abilities directly from the provided captions of the PIC datasets, this approach cannot exploit the enormous amount of the external text to boost the performance for PIC.

In this work, we propose to employ the pre-trained language model GPT2 (Radford et al., 2019) as a language prior for both the speaker and listener in the multi-agent communication framework. As GPT2 has been trained on a large amount of unlabeled text, we expect that its incorporation can significantly improve the language modeling/encoding for the speaker and listener. To our knowledge, this is also the first work to consider pre-trained language models for PIC. Finally, we conduct extensive experiments on the PC dataset to demonstrate the benefits of the proposed framework, leading to the state-of-the-art performance for this dataset.

2 Model

Given an image I and a personality trait T (i.e., a word), the goal of PIC is to generate an engaging caption $\hat{C} = \hat{w}_1, \hat{w}_2, \ldots, \hat{w}_{\hat{N}}$ (i.e., of $\hat{N}$ words). In the supervised learning setting, there is a ground-truth caption C for each pair (I, T): $C = w_1, w_2, \ldots, w_N$ (i.e., of N words).

To encode I, we first feed it into the ResNeXt ConvNet model (Mahajan et al., 2018) to obtain a feature map of size 7×7×2048. This can be viewed as a matrix V of size 49×2048, where each row encodes the visual content for a cell of the uniform image grid. V is called the representation of I in the following. Also, we use T to refer to the personality trait or its embedding vector interchangeably in this work (these vectors are randomly initialized and updated during training).

2.1 Adapting the Structure of GPT2 for PIC

Our PIC model involves a multi-agent framework where a speaker and a listener communicate to solve PIC. Our PIC model uses the pre-trained language model GPT2 (Radford et al., 2019) as the starting point for both the speaker and listener to benefit from its language modeling capacity. This GPT2 model is fine-tuned for PIC in the training.

In particular, our goal is to adapt GPT2 so it can accept the representation V of I, the personality trait T, and some sequence of words $\bar{C}_k = \bar{c}_1, \bar{c}_2, \ldots, \bar{c}_k$ as the inputs and produce a representation vector $G(V, T, \bar{C}_k)$ for the input triple as the output. Here, $\bar{C}_k$ can be any sequence of k words in the vocabulary. The representation vector $G(V, T, \bar{C}_k)$ can also be used for different purposes in the speaker and listener (i.e., to predict the next word $\hat{c}_{k+1}$ in the speaker or to estimate a compatible score for the input triple $(V, T, \bar{C}_k)$ in the listener).

In particular, taking as input a sequence of words $\bar{C}_k = \bar{c}_1, \bar{c}_2, \ldots, \bar{c}_k$, the vanilla version of the GPT2 language model would send $\bar{C}_k$ to a stack of transformer layers, producing the hidden vectors $h^l_1, h^l_2, \ldots, h^l_k$ for the words in $\bar{C}_k$ at the l-th transformer layer (modulo the tokenization for the words in $\bar{C}_k$) (i.e., $h^l_i \in \mathbb{R}^{1\times d}$). Afterward, the hidden vector for the last word $\bar{c}_k$ in the last transformer layer L (i.e., h^L_k) is typically used as the representation vector for the input sequence $\bar{C}_k$. In GPT2, the hidden vector h^{l+1}_t ($1 \leq t \leq k$, $0 \leq l < L$) is computed via self-attention:

$$h^{l+1}_t = \text{softmax}((h^l_t W^l_q)(H^l_t W^l_k)^T)(H^l_t W^l_v)$$

where $H^l_t = [h^l_1, h^l_2, \ldots, h^l_t] \in \mathbb{R}^{t\times d}$, and W^l_q, W^l_k, and W^l_v are the query, key, and value weights for the l-the layer. Note that we omit the multiple heads and the biases in this work for simplicity.

Given this version of GPT2, we propose to directly inject the representation vectors for I and T into the self-attention computation for every transformer layer in GPT2 to obtain the representation $G(V, T, \bar{C}_k)$. In particular, the hidden vector h^{l+1}_t in this case would be computed via the *softmax* function sf:

$$h^{l+1}_t = \text{sf}\left((h^l_t W^l_q)\begin{bmatrix} VP^l_k \\ TW^l_k \\ H^l_t W^l_k \end{bmatrix}^T\right)\begin{bmatrix} VP^l_v \\ TW^l_v \\ H^l_t W^l_v \end{bmatrix}$$

where P^l_k and P^l_v are the new key and value weight matrices for GPT2 to transform the image representation vectors in V into the same space as the hidden vectors in H^l_t (i.e., of d dimension). Note that we reuse the key and value weight matrices W^l_k and W^l_v in GPT2 to transform the trait embedding vector T as it comes with the same modality (i.e., text) as $\bar{C}_k$. Finally, the representation vector $G(V, T, \bar{C}_k)$ is set to the hidden vector for the last word $\bar{c}_k$ in the last transformer layer, i.e., $G(V, T, \bar{C}_k) = h^L_k$.

2.2 The Speaker-Listener Framework

PIC is recast as a communication game between a speaker and a listener. We first feed the input image I and personality trait T into the speaker model to generate a caption $\hat{C}$. The listener model then consumes $\hat{C}$ and learns to rank the input image I and trait T higher than another distractor image and trait (i.e., being able to use the information in $\hat{C}$ to identify the input image and trait).

To generate the word $\hat{c}_k$ for the caption $\hat{C}$, the speaker feeds the GPT2 representation vector $G(V, T, \hat{C}_{k-1})$ for the image, the trait, and the previously generated words $\hat{C}_{k-1} = \hat{c}_1, \ldots, \hat{c}_{k-1}$ into a feed-forward network F_{spk}, producing a distribution $\hat{P}(.|V, T, \hat{C}_{k-1})$ over the vocabulary for next word prediction[1]. Note that we differentiate $\hat{P}(.|V, T, \hat{C}_{k-1})$ from the distribution $P(.|V, T, C_{k-1})$ computed via the ground-truth caption $C_{k-1} = c_1, \ldots, c_{k-1}$ with the representation vector $G(V, T, C_{k-1})$ to be used later in this work. As the goal of the listener is to solve a ranking problem, we send the representation vector $G(V, T, \hat{C})$ into another feed-forward net F_{ltn} to produce a compatible score $s(V, T, \hat{C}) = F_{ltn}(G(V, T, \hat{C}))$ for the triple that would be used to perform the ranking later. Note that the speaker and listener share the GPT2 model G in this work.

Pre-training: The training process for our framework involves propagating of the training signals from the listener to the parameters in the speaker. As the speaker and listener are linked via the generated captions $\hat{C}_k$, which is a discrete variable, we use REINFORCE (Williams, 1992) to train the PIC model. As this method requires the reward as the training signals, we first pre-train the feed-forward network F_{ltn} in the listener so it can provide the rewards (i.e., based on the compatible scores) for our later training step. In particular, in the pre-training step, we train the speaker and listener with the following loss:

$$\mathcal{L}_{pretrain} = \alpha \mathcal{L}_{CE} + (1 - \alpha)\mathcal{L}_{comp}$$

where α is a trade-off parameter, $\mathcal{L}_{CE}$ is the cross-entropy loss for the ground-truth caption C: $\mathcal{L}_{CE} = -\sum_{k=1}^{N} \log P(c_k|V, T, C_{k-1})$, and $\mathcal{L}_{comp}$ is the logistic loss: $\mathcal{L}_{comp} = \log(1 + e^{s(V,T,C')-s(V,T,C)})$. Here, C' is a the ground-truth caption for another example/triple in the same batch with the current example (i.e., the distractor caption). Note that $\mathcal{L}_{comp}$ helps to train F_{ltn} using the training signals from the ground truth captions.

Training: In the main training step, our goal is to train the speaker and listener so the generated caption $\hat{C}$ of the speaker can: (i) be similar to the ground truth C, and (ii) provide sufficient information to identify the input image and trait from the distractors. In particular, to achieve the similarity between $\hat{C}$ with C, we employ the CIDEr score (Vedantam et al., 2015) of $\hat{C}$ as one part of the reward for REINFORCE: $R_{lang} = CIDEr(\hat{C})$. In addition, to enforce the sufficient information in $\hat{C}$, we introduce the following rewards R_{img} and R_{trait} for REINFORCE:

$$R_{img} = -\max(0, m + s(V', T, \hat{C}) - s(V, T, \hat{C}))$$
$$R_{trait} = -\max(0, m + s(V, T', \hat{C}) - s(V, T, \hat{C}))$$

where m is a margin parameter for the Hinge losses, and V' and T' are the representation vectors for another image and personality trait that are sampled from the same batch with the current example during training (i.e., the distractors). By maximizing these rewards, we increase the compatible scores of the generated caption with the input image and trait (i.e., $V, T, \hat{C}$) and decrease those with the distractor image and trait (i.e., $V', T, \hat{C}$ and $V, T', \hat{C}$). In this way, we expect that $\hat{C}$ can be enriched to better fit with I and T. Overall, the reward for REINFORCE in this work is:

$$R(\hat{C}) = \beta R_{img} + \gamma R_{trait} + (1 - \beta - \gamma)R_{CIDEr}$$

With REINFORCE, we seek to minimize the negative expected reward R over the possible choices of $\hat{C}$: $\mathcal{L} = -\mathbb{E}_{\hat{C} \sim \hat{P}(\hat{C}|V,T)}[R(\hat{C})]$. The policy gradient is estimated by: $\nabla \mathcal{L} = -\mathbb{E}_{\hat{C} \sim \hat{P}(\hat{C}|V,T)}[(R(\hat{C}) - b) \log \nabla \hat{P}(\hat{C}|V, T)]$ where b is a baseline to reduce the variance. Motivated by (Rennie et al., 2017), we obtain b by evaluating the reward $R(C^*)$ for the greedy decoding caption C^*. Finally, we approximate $\nabla \mathcal{L}$ with one roll-out sample.

3 Experiments

3.1 Dataset and Hyper-parameters

We evaluate our models using the PC dataset (Shuster et al., 2019), which consists of 241,858 triplets of image-personality-caption with 215 personality traits. It is divided into three separate parts for training (186K+ examples), development (5K examples), and testing (10K examples). The hyper-parameters for the models are fine-tuned on the

[1] Note that the generated caption always involves the special symbols SOS as the start word and EOS as the end word.

development set. The selected hyper-parameters include: $1.25e$-4 and $3.25e$-5 respectively for the learning rate of the pre-training step and the main training step (respectively) with the Adam optimizer, 64 and 256 for the batch sizes in mini-batching of the pre-training and main training step, 3 for the beam search size in the inference time, 0.5, 0.3, and 0.2 for the parameters α, β, and γ respectively, and 1 for the margin m in the Hinge losses. We train the proposed model with 20 epochs for the pre-training step and 3 epochs for the main training step using early stopping on the development data. In addition, we use the distilled version of GPT2 in (Sanh et al., 2019) for the GPT2 model in this work. The size of the transformer model in GPT2 follows (Sanh et al., 2019) where the number of layers is $L = 6$, the number of attention heads is 8, the dimensionality of the hidden vectors is $d = 1024$, and the dimension of the input embeddings (i.e., the segmentation embeddings, positional embeddings, and word embeddings) is 768. Finally, we use Byte Pair Encoding (Sennrich et al., 2016) to tokenize the captions in the dataset.

3.2 Comparing to the State of the Art

We compare our proposed model (called GPT-Speaker) with the state-of-the-art models on the PC test data. In particular, we consider the following baselines (reported in Shuster et al. (2019)): (1) **ShowTell**: the encoder-decoder architecture (Vinyals et al., 2014), (2) **ShowAttTell**: a similar model to ShowTell where the visual feature vector is computed via attention (Xu et al., 2015), and (3) **UpDown**: an encoder-decoder model with two LSTM layers for the decoder (Shuster et al., 2019). UpDown, which is adapted from (Anderson et al., 2017), is the current state-of-the-art model on the PC dataset. Following Shuster et al. (2019), standard measures are employed to evaluate the models, including BLEU, ROUGE-L, CIDEr, and SPICE. Table 1 presents the performance of the models on the PC test set. As can be seen, our proposed model significantly outperforms the baseline models across different performance measures, clearly demonstrating the benefits of GPT-Speaker for PIC.

3.3 Ablation Study

The major contribution in this work is the introduction of the speaker-listener communication game for PIC that is trained with REINFORCE using the reward $R(\hat{C})$ and the pre-trained language model GPT2 (i.e., in the main training step). In particular, the overall reward $R(\hat{C})$ involves three components, i.e., R_{img}, R_{trait}, and R_{CIDEr}. This section evaluates the effects of these components for GPT-Speaker by incrementally removing them from the full model. Table 2 reports the performance of the models on the test set.

Models	B@1	B@4	R	C	S
ShowTell	38.4	7.3	24.3	9.6	1.6
ShowAttTell	43.3	7.1	27.0	12.6	3.6
UpDown	44.4	8.0	27.4	16.5	5.2
GPT-Speaker (ours)	**52.1**	**8.4**	**30.2**	**19.9**	**7.3**

Table 1: Comparison with the state-of-the-art models on the PC test set. B@1, B@4, R, C and S represent BLEU@1, BLEU@4, ROUGE-L, CIDEr and SPICE respectively.

Models	B@1	B@4	R	C	S
GPT-Speaker	**52.1**	8.4	**30.2**	**19.9**	**7.3**
- R_{img}	52.1	7.5	29.7	19.2	6.8
- R_{trait}	49.7	8.0	29.3	18.8	6.8
- R_{img} - R_{trait}	51.5	8.8	29.8	19.6	6.1
- R_{img} - R_{trait} - R_{CIDEr}	48.7	**9.3**	29.7	16.9	5.3
-GPT	49.2	9.1	29.0	19.0	6.3
Pretrained with $\mathcal{L}_{CE}$ only	47.1	8.8	29.1	16.3	5.2

Table 2: Ablation study.

From the table, we see that both R_{img} and R_{trait} are important; excluding them would decrease the performance of GPT-Speaker. As these reward components are associated with the listener, it demonstrates the benefits of the listener for PIC. In addition, the exclusion of the main training step, which corresponds to the line "- R_{img} - R_{trait} - R_{CIDEr}" in the table, also leads to a large performance reduction. This clearly testifies to the advantages of the speaker-listener framework and the main training step for PIC. Importantly, in the line "*Pretrained with $\mathcal{L}_{CE}$ only*", we show the performance of the model when it is only trained with the pre-training step using the cross-entropy $\mathcal{L}_{CE}$ (i.e., only training the GPT2-based speaker with $\mathcal{L}_{CE}$). As we can see, this model is worse than "- R_{img} - R_{trait} - R_{CIDEr}", thus proving the advantage of the loss function $\mathcal{L}_{comp}$ for the pre-training step. However, "*Pretrained with $\mathcal{L}_{CE}$ only*" still outperforms the baseline UpDown in Table 1 that is also trained with $\mathcal{L}_{CE}$, clearly showing the effectiveness of GPT2 for language generation in PIC. Finally, some qualitative analysis is presented in Appendix A.

3.4 Human Evaluation

Finally, we perform a human evaluation to further compare the proposed model GPT-Speaker with the

Type of evaluation	WIN PERCENTAGE	
	GPT-Speaker	UpDown
Engagingness	**65.8**	34.2
Image Relevance	**63.8**	36.2
Personality Relevance	**66.9**	33.1

Table 3: Human Evaluation.

UpDown baseline (Shuster et al., 2019). In particular, following (Shuster et al., 2019), we consider two classes of evaluations that examine the Engagingness and Relevance of the generated captions from the models. As such, the engagingness evaluation considers human preference for the naturalness and appropriateness of the generated captions while the relevance evaluation concerns human judgment on the relatedness of the generated captions with the information presented in the input images and personality traits. In particular, we further divide the relevance test into two categories, depending on whether it assesses the relatedness with the input images or personality traits (leading to three actual types of human evaluations in this work). For each of these types, we randomly sample 50 pairs of images and personality traits from the test set (i.e., the samples are different for the three evaluation). Afterward, we apply the trained models (i.e., GPT-Speaker and UpDown) to generate captions for these selected image-personality pairs. We then present the selected image-personality pairs along with their generated captions from GPT-Speaker and UpDown to 12 recruited annotators (i.e., resulting in 600 trials in total for each type of human evaluations). For an image-personality pair, based on its corresponding test, the annotator is asked to determine which generated caption (i.e., from GPT-Speaker or UpDown) is more engaging (i.e., for the engagingness test), more related to the input image (i.e., for the relevance test with the image), and more related to input personality trait (i.e., for the relevance test with the trait). In the next step, for each of the tests, we record the percentage of times the generated captions from GPT-Speaker and UpDown are selected by the annotators (i.e., the win percentages). Table 3 shows the win percentages of GPT-Speaker and UpDown for the three tests. It is clear from the table that GPT-Speaker substantially outperforms UpDown in this human evaluation. This is significant with $p < 0.005$ (using a binomial two-tailed test), thus highlighting the advantage of GPT-Speaker to generate more engaging and relevant captions for PIC.

4 Related Work

The main approach for IC so far involves deep learning models where several datasets have been created (Chen et al., 2015; Young et al., 2014) and different variants of the encoder-decoder architectures have been proposed (Xu et al., 2015; Herdade et al., 2019; Su et al., 2019). PIC is a way to encourage more engaging captions for which several features are considered, i.e., location and age (Denton et al., 2015), reader's active vocabulary (Park et al., 2017), humour (Yoshida et al., 2018), sentiment (Mathews et al., 2016), dialog/conversation (Zhang et al., 2018), and caption styles (Gan et al., 2017; Mathews et al., 2018). The closest work to ours is (Shuster et al., 2019) that examines a different feature of diverse personality traits.

Our work also bears some similarity with the previous IC models that attempts to improve the ability to discriminate images for the generated captions (Liu et al., 2018; Luo et al., 2018; Vered et al., 2019). However, these IC models do not capture personality traits for PIC as we do. We also note the stylized IC model in (Guo et al., 2019) that applies a style classification loss. However, this work does not consider the speaker-listener framework with REINFORCE training as GPT-speaker. Above all, none of these works has exploited pre-trained language models (i.e., GPT2) for PIC.

5 Conclusions

We formulate PIC as a communication framework between a speaker and a listener. A novel training mechanism for this framework is introduced, exploiting the rewards in REINFORCE to encourage the generated captions to be natural and informative about the input images and traits. We also introduce the pre-trained language model GPT2 into the model to benefit from its language modeling/encoding capacity. The experiments demonstrate the effectiveness of the proposed model for PIC.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2017. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. In *CoRR (2015)*.

Emily Denton, Jason Weston, Manohar Paluri, Lubomir D. Bourdev, and Rob Fergus. 2015. User conditional hashtag prediction for images. In *KDD*.

Katrina Evtimova, Andrew Drozdov, Douwe Kiela, and Kyunghyun Cho. 2018. Emergent language in a multi-modal, multi-step referential game. *ICLR*.

Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. 2017. Stylenet: Generating attractive visual captions with styles. In *CVPR*.

Longteng Guo, Jing Liu, Peng Yao, Jiangwei Li, and Hanqing Lu. 2019. Mscap: Multi-style image captioning with unpaired stylized text. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Simao Herdade, Armin Kappeler, Kofi Boakye, and João Paulo Holanda Soares. 2019. Image captioning: Transforming objects into words. In *NeurIPS*.

Satwik Kottur, José Moura, Stefan Lee, and Dhruv Batra. 2017. Natural language does not emerge 'naturally' in multi-agent dialog. In *EMNLP*.

Angeliki Lazaridou, Karl Moritz Hermann, Karl Tuyls, and Stephen Clark. 2018. Emergence of linguistic communication from referential games with symbolic and pixel input. *ICLR*.

Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. 2016. Multi-agent cooperation and the emergence of (natural) language. *CoRR*.

Angeliki Lazaridou, Anna Potapenko, and Olivier Tieleman. 2020. Multi-agent communication meets natural language: Synergies between functional and structural language learning. In *arxiv 2005.07064v1*.

Xihui Liu, Hongsheng Li, Jing Shao, Dapeng Chen, and Xiaogang Wang. 2018. "show, tell and discriminate: Image captioning by self-retrieval with partially labeled data". In *ECCV*.

Ruotian Luo, Brian L. Price, Scott Cohen, and Gregory Shakhnarovich. 2018. Discriminability objective for training descriptive captions. In *CVPR*.

Dhruv Kumar Mahajan, Ross B. Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. 2018. Exploring the limits of weakly supervised pretraining. In *ECCV*.

Alexander Patrick Mathews, Lexing Xie, and Xuming He. 2016. Senticap: Generating image descriptions with sentiments. In *AAAI*.

Alexander Patrick Mathews, Lexing Xie, and Xuming He. 2018. Semstyle: Learning to generate stylised image captions using unaligned text. In *CVPR*.

Cesc Chunseong Park, Byeongchang Kim, and Gunhee Kim. 2017. Attend to you: Personalized image captioning with context sequence memory networks. In *CVPR*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. In *TechReport OpenAI*.

S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel. 2017. Self-critical sequence training for image captioning. In *CVPR*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *NeurIPS EMC^2 Workshop*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *ACL*.

Kurt Shuster, Samuel Humeau, Hexiang Hu, Antoine Bordes, and Jason Weston. 2019. Engaging image captioning via personality. In *CVPR*.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. *ArXiv*, abs/1908.08530.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *CVPR*.

Gilad Vered, Gal Oren, Yuval Atzmon, and Gal Chechik. 2019. Joint optimization for cooperative image captioning. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2014. Show and tell: A neural image caption generator. In *CVPR*.

Ronald J. Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. In *Kluwer Academic*.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*.

Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. 2018. Auto-encoding scene graphs for image captioning. In *CVPR*.

Kota Yoshida, Munetaka Minoguchi, Kenichiro Wani, Akio Nakamura, and Hirokatsu Kataoka. 2018. Neural joking machine : Humorous image captioning. In *arXiv preprint arXiv:1805.11850*.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *ACL*.