Zero-Shot Visual Storytelling

Simo Ryu [1] **Seungyun Baek** [2]

1. Introduction

In computer vision, large research efforts have been devoted to generating natural language description for visual contents (Gao et al., 2017), (Liang et al., 2017), (Rohrbach et al., 2014), (Yu et al., 2016), (Krishna et al., 2017), (Xu et al., 2018), (Yu et al., 2016), (Yu et al., 2019).

Task of generating natural language description of images can be classified into two lines of research, Image Captioning (Gao et al., 2017), (Liang et al., 2017) and Visual Storytelling (Huang et al., 2016), (Huang et al., 2019). While image captioning is the task of directly generating single caption per image, visual storytelling extends this task and targets to generating multiple sequence of coherent story that matches visual semantics of given sequences of images.

Existing literature on generating natural language description of images are largely driven by framework of language model conditioned on visually encoded features (Hossain et al., 2019). Other than some researches that explores capability of non-autoregressive language models (Guo et al., 2020), most models can be thought of special case of conditional language models. It is also the case that all of these methods are specifically trained on some large image-caption pairs, in some way or another, trying to optimize the conditional language model. To our best knowledge, there is no visual-description model that directly generates natural sentences without any particular training process, i.e., zero-shot way. Consequently, there seems to be lack of research in zero-shot task of Visual Storytelling.

We propose novel zero-shot visual storytelling method that leverages pre-trained language model and multi-modal similarity model. Although both models were not particularly created to generate descriptions of image sequences, our Visually Guided Beam Search method (ViGS) can achieve the goal in zero-shot way. We further evaluate our method both quantitatively and qualitatively to verify its significance. If one regards Visual Storytelling's stream of images with as stream of single image, our method is indeed generalization of Image Captioning. Naturally our method can perform image captioning as well. We also share the codebase for running similar experiments in the future.[1]

2. Backgrounds

Beam Search Beam search is an algorithm used in NLP models as a final decision making layer to choose the best output. Beam search is an improved version of greedy search. It keeps track of k states rather than one. At each step, k states generate their successors and selects the k best successors from the outputs and repeats. Many studies used beam search for NLP models (Sutskever et al., 2014), (Wu et al., 2016).

Language Model Language modelling is the development of probabilistic models that are able to predict the next word in the sequence. Given sequence $W = \{w_1, \dots, w_n\}$, a standard language modelling objective is maximizing the following likelihood:

$$L(W) = \sum_i \log P_l(w_i | w_{i-1}, \dots, w_i; \Theta)$$

where the conditional Probability P_l is modeled with parameters Θ. (Radford et al., 2018), (Radford et al., 2019) are well-known Language Models.

Text-Image Joint representations Learning Contrastive Language-Image Pre-training (CLIP) learns a multi-modal embedding space, which can be used to estimate the semantic similarity between a text and an image (Radford et al., 2021). We use text-image similarity model D based on CLIP. D is the cosine distance between the embedding of text $W = (w_1, \dots, w_n)$, and image I.

3. Related Works

Visual Storytelling Though a lot of works have been achieved impressive results in visual captioning, the task of generating story from sequences of image, which requires a deeper understanding, has been rarely studied. (Park &

[*] Equal contribution [1] Designed experiments, written paper and code, ran experiments. [2] Curated datasets, written paper and code, ran experiments..

This is a report submitted to COSE474-2019F Deep Learning @ Korea University

[1] Code and training configurations are available at `github.com/cloneofsimo/zeroshot-storytelling`

Kim, 2015) firstly proposed an approach for retrieving sentence sequences for an image stream. (Huang et al., 2016) introduced the first dataset (VIST) for sequential vision-to-language for evolving AI towards more human-like understanding. We used VIST for training our model. (Wang et al., 2018) introduced an adversarial reward learning algorithm to generate human-like stories given image sequences. However, many studies have not yet shown impressive results in Visual Storytelling.

Zero-Shot Transfer In computer vision, zero-shot learning models learned parameters for seen classes along with their class representations, and generalizing to unseen classes in image classification. To our best knowledge, (Chang et al., 2008) firstly proposed zero-shot learning in natural language processing (NLP). In computer vision, (Li et al., 2017) first studied zero-shot transfer to existing image classification datasets. In the case of GPT-2(Radford et al., 2019), they use zero-shot learning and outperform on a lot of task in natural language processing. CLIP(Radford et al., 2021) is trained on a large corpus of image-text pairs and demonstrates impressive zero-shot transfer capabilities.

4. Method

In this section, we describe our ViGS (Visually Guided Beam Search) framework. We begin by formulating our problem, and how we approached the problem in zero-shot way. Rest of the section will describe in detail how our method was implemented and executed.

4.1. Problem Definition

We consider the task of visual storytelling. Objective of Visual storytelling is generating coherent sentences (sequence of words) $(w_1, w_2, ...w_r)$, according to input sequence of images, $(I_1, I_2, I_3, ...I_n)$. Naively, given that we have language modelling conditioned on sequence of images,

$$P(w_1, w_2, \ldots, w_r | I_1, I_2, \ldots, I_n) = \prod_{i=1}^{r} P(w_i | w_1, \ldots, w_{i-1}, I_1, ...I_n)$$

we wish to sample $w_1, w_2, ...w_r$ that maximizes the likelihood given set of images $I_1, I_2, ...I_n$. However, in visual storytelling scenario, we wish to generate story that has local description of the image, but globally coherent. Therefore, we modify our objective probability as the following,

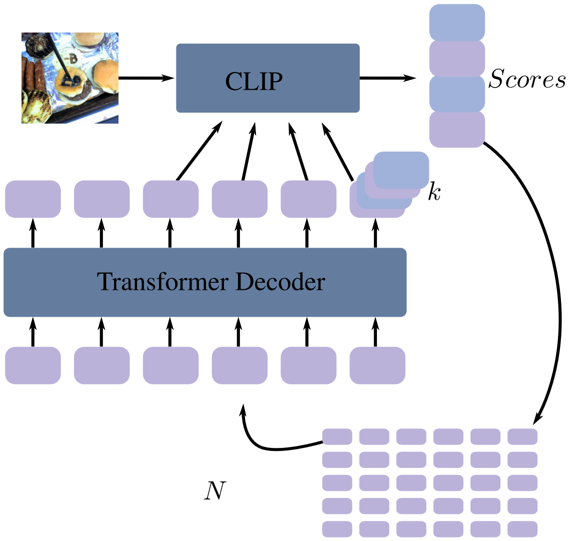

Figure 1. Overview of our method. We first sample k child candidates from N candidates via pre-trained language model. From these kN candidates, we select N elements with greatest visual similarity, which is evaluated with CLIP.

$$P_m(w_1, w_2, \ldots, w_r | I_1, I_2, \ldots, I_n) := \prod_{i=1}^{r} P_m(w_i | w_1, \ldots, w_{i-1}, I_{g_i})$$

where g_i is the index of image that ith word is set to describe. P_m is incorporated with an appropriate model with image-text modality, to model the conditional probability regarding both image-text information.

Accordingly, our objective becomes

$$\arg\max_{w_1, w_2, \ldots, w_r} P_m(w_1, \ldots, w_r | I_1, \ldots, I_n)$$

4.2. Visually Guided Beam Search

We wish to generate sequence of words that is globally coherent and natural, that is also describing sequences of given images. As formulated above, our goal is ultimately modelling joint conditional language model, $P_m(w_i | w_1, \ldots, w_{i-1}, I_{g_i})$. One core contribution of our work is that we approach this modelling without any actual dataset.

To achieve this task zero-shot, assume that we have pre-trained auto-regressive language model P_l, and text-image similarity model D.

We propose our model P_m as conditional probability distribution using P_l and D as the following.

Algorithm 1 ViGS

```
Input: Initial string Images $I_1, \dots, I_n$, Language Model $P_l$, Image-Text Model $D$
Paths ⟵ Initialize $N$ initial string.
for $i = 1$ to $r$ do
    Candidates ⟵ Generate $k$ random candidates from each Paths string with $P_l$.
    Paths ⟵ Score Candidates with $D$, sample Top $N$ elements.
end for
```

$$P_m(w_i|w_1,\dots,w_{i-1},I_{g_i}) \propto P_l(w_i|w_1,\dots,w_{i-1})D((w_1,w_2,\dots,w_i),I_{g_i})$$

In practice, sampling from this distribution is computationally difficult, as one evaluation of the value $P_l D$ takes $O(n^2)$ amounts of computation. Thus, we decompose our sampling method into two heuristic stage while naturally integrating beam search.

Refer to Figure 1. and Algorithm 1. for overview of our method. Initially, starting with N candidates, first explores k child candidates from each candidates according to given language model. This can be done with various methods, including wildly used top-k sampling (Fan et al., 2018), Nucleus sampling (Holtzman et al., 2019), or stochastic beam search (Kool et al., 2019). We end up with total of Nk sentences.

Next, we keep N candidates that has greatest similarity score with the given image. This can be done with model D, which we have utilized with CLIP (Radford et al., 2021).

5. Experiments

In this section, we will discuss evaluation metrics and experiments for our proposed method.

5.1. Evaluation Metric

Unlike Neural Machine Translation, generative results of visual narrative has vastly large space. Thus, metrics such as BLUE or METEOR would not be a good fit. For these reasons, we have used average accuracy for our metric.

5.2. VIST Dataset Experiment

Overall, our goal is to generate (possibly multiple) sentences that is both visually and linguistically coherent.

For this purpose, we've used VIST (Huang et al., 2016) dataset, which is most widely used dataset with image-sequences, sentence-sequences annotation. To evaluate how our method performs, we will use 5 stories (40470, 40471, 40472, 40473, 40474) and evaluate the accuracy score. To perform our method, we require both auto-regressive language model and Image-text similarity model. For the language model, we have used DistilGPT2 and GPT2 (Radford et al., 2019). Sampling from auto-regressive language model, we have used sampling with temperature 0.6. For other hyperparameters, we have used $N = 800$, $k = 400$.

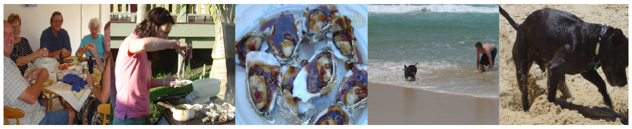

Figure 2. ViGS Generated Caption : *The visitors and family members of the tributary roof were invited to join in a family gathering on the state's Gran Lucha Park in Wellington.. The cuisine and food, featuring a combination of strong sauces and clams, was presented by a three year old boy who came down to the Tackles Centre on the Sunshine Coast and died when he fell in the water.Members of the Tackles Centre have been growing to enjoy the comfort of the three-year old boy in the water.*

Table 1. VIST Dataset evaluation with different language model decoder. Compared to naively generating with only decoder, having visual guidance by CLIP improves accuracy.

Decoder	Guided	Unguided
DistilGPT2	12.53	11.94
GPT2-Small	13.12	12.57
GPT2-Medium	13.07	10.23

Our method is fully implemented with PyTorch (Paszke et al., 2017), supplemented with pre-trained Transformers implementation from Huggingface (Wolf et al., 2019). Our implemented code was executed from Linux OS with CUDA RTX 3090 device.

5.3. Effect of Visual Guidance

Our primary results are shown in Table 1. First, we conduct a controlled experiment to evaluate the generation accuracy with and without visual guidance. Across 3 different language model decoders, we found that having visual guidance consistently improved accuracy on VIST dataset.

In Figure 2, some zero-shot generated examples from VIST input image sequences is presented. Note that our generated image is vastly different from ground truth, yet, is still meaningfully coherent and natural.

5.4. User Study

We qualitatively evaluate our methods regarding its expressiveness and coherency by performing the user study.

Table 2. User evaluation on preference between Human generated captions and ViGS generated captions.

DECODER	VIGS	HUMAN
GPT2-MEDIUM	33%	67%

Table 3. User evaluation on preference between captions generated by two distinct decoders.

GPT2-MEDIUM	DISTILGPT2
67 %	33 %

Summary of our survey could be found in Table 2. Regarding our resources, our user study was performed in relatively small scale. We asked the users to perform pairwise selection between two visual description that was created by the authors and ViGS (GPT2-Medium). Users were asked to select visual description that was more natural.

We next asked the users to perform pairwise selection between ViGS generated captions that were generated by GPT2-Medium and DistilGPT2, to determine which method was more natural.

User study suggests that even though users were mostly able to distinguish between human generated caption and ViGS generated caption, they mostly preferred captions that were generated with larger model, which aligns with the result from VIST Dataset experiment.

Overall, we argue that experiments suggests our method is both qualitatively and quantitatively meaningful and effective.

6. Future Direction

Our method sufferes from limits of computational resources that are yet to be solved. ViGS is currently dependent on many iterative forward passes of relatively large language models, which is both very power-consuming and time consuming. Since our work showed that it is possible to use both generative capability of pre-trained language model and multi-modality of pre-trained Transformer model to create coherent visual descriptions, next natural step would be to make it more efficient.

We have also not fully explored which sampling method works the best. Although we noted ”beam search”, the overall formulation and methodology does not require the sampling step of the language model to be restricted to beam search.

References

Chang, M.-W., Ratinov, L.-A., Roth, D., and Srikumar, V. Importance of semantic representation: Dataless classification. In *Aaai*, volume 2, pp. 830–835, 2008.

Fan, A., Lewis, M., and Dauphin, Y. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*, 2018.

Gao, L., Guo, Z., Zhang, H., Xu, X., and Shen, H. T. Video captioning with attention-based lstm and semantic consistency. *IEEE Transactions on Multimedia*, 19(9):2045–2055, 2017.

Guo, L., Liu, J., Zhu, X., He, X., Jiang, J., and Lu, H. Non-autoregressive image captioning with counterfactuals-critical multi-agent learning. *arXiv preprint arXiv:2005.04690*, 2020.

Holtzman, A., Buys, J., Du, L., Forbes, M., and Choi, Y. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019.

Hossain, M. Z., Sohel, F., Shiratuddin, M. F., and Laga, H. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CsUR)*, 51(6):1–36, 2019.

Huang, Q., Gan, Z., Celikyilmaz, A., Wu, D., Wang, J., and He, X. Hierarchically structured reinforcement learning for topically coherent visual story generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 8465–8472, 2019.

Huang, T.-H., Ferraro, F., Mostafazadeh, N., Misra, I., Agrawal, A., Devlin, J., Girshick, R., He, X., Kohli, P., Batra, D., et al. Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1233–1239, 2016.

Kool, W., Van Hoof, H., and Welling, M. Stochastic beams and where to find them: The gumbel-top-k trick for sampling sequences without replacement. In *International Conference on Machine Learning*, pp. 3499–3508. PMLR, 2019.

Krishna, R., Hata, K., Ren, F., Fei-Fei, L., and Carlos Niebles, J. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pp. 706–715, 2017.

Li, A., Jabri, A., Joulin, A., and van der Maaten, L. Learning visual n-grams from web data. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4183–4192, 2017.

Liang, X., Hu, Z., Zhang, H., Gan, C., and Xing, E. P. Recurrent topic-transition gan for visual paragraph generation. In *Proceedings of the IEEE international conference on computer vision*, pp. 3362–3371, 2017.

Park, C. C. and Kim, G. Expressing an image stream with a sequence of natural sentences. *Advances in neural information processing systems*, 28:73–81, 2015.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. Automatic differentiation in pytorch. 2017.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. Improving language understanding by generative pre-training. 2018.

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

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

Rohrbach, A., Rohrbach, M., Qiu, W., Friedrich, A., Pinkal, M., and Schiele, B. Coherent multi-sentence video description with variable level of detail. In *German conference on pattern recognition*, pp. 184–195. Springer, 2014.

Sutskever, I., Vinyals, O., and Le, Q. V. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pp. 3104–3112, 2014.

Wang, X., Chen, W., Wang, Y.-F., and Wang, W. Y. No metrics are perfect: Adversarial reward learning for visual storytelling. *arXiv preprint arXiv:1804.09160*, 2018.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

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

Xu, N., Liu, A.-A., Wong, Y., Zhang, Y., Nie, W., Su, Y., and Kankanhalli, M. Dual-stream recurrent neural network for video captioning. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(8):2482–2493, 2018.

Yu, H., Wang, J., Huang, Z., Yang, Y., and Xu, W. Video paragraph captioning using hierarchical recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4584–4593, 2016.

Yu, J., Li, J., Yu, Z., and Huang, Q. Multimodal transformer with multi-view visual representation for image captioning. *IEEE transactions on circuits and systems for video technology*, 30(12):4467–4480, 2019.