W266 Section 3 Final Project - DRAFT
Neural Narratives for Multiple Images

Sundeep Kumar Mahensaria , Eric Lee,
{smahensaria, ericlee.2020}@berkeley.edu

Abstract

We extend the visual storytelling (VIST) task of [Huang et. al., 16] that uses a 'story-language' paradigm instead of mere 'image captions' to develop a narrative across multiple 'discrete' images. Most existing research uses image captioning to generate narratives across multiple images. In this paper we exploit neural text generation techniques to develop a story across multiple images by GPT-2 models on VIST dataset [Huang et. al,, 16]. We find that the METEOR scores of story-language generated using vanilla GPT-2 on individual image captions in [Huang et. al, 16] are much higher than results reported in [16]. However, METEOR scores from a GPT-2 model fine-tuned with the VIST training dataset is lower. Additionally, METEOR scores for the combined narratives are much lower for the machine generated stories.

Introduction

Narratives and stories, as opposed to just viewing discrete images, help humans comprehend an idea much better and retain it for a long time. For example, narratives can be more effective than reviewing crime scene photos taken by law-enforcement agencies [15]. The text narrative can also be enhanced with audio/speech for the benefit of the visually impaired. Numerous models exist in captioning single images [1, 3] and videos [2]. However, the task of automated generation of narratives across multiple discrete, yet related, images is challenging because it is difficult to blend discrete images to tell a coherent story without proper context [18].

Most existing research uses image captioning to generate narratives across multiple images. In this paper we used the GPT-2 model on VIST dataset [16] to generate story language from the image captions. We adopt a novel approach in which we do not generate our own captions from the images. Instead we rely on image captions generated from the state-of-the-art MS COCO based model and then apply SOTA text generation techniques for story generation. Effectively, we leverage models that are SOTA in their respective individual tasks to address a challenging composite task.

We find that the story languages generated from vanilla GPT-2 [17] have much higher METEOR scores for individual image captions when compared to our baseline and existing work [16]. These results open the door to two key areas - a) evaluation of machine-generated stories need

not rely on human-generated stories as gold standard; b) visual storytellers can use state-of-the-art transformer based models to generate stories.

Background

Visual Storytelling (VIST)

VIST [16, Huang et. al. 2016] was the first dataset for sequential vision-to-language and has since been the reference dataset for visual storytelling. This research developed a narrative across 5 discrete images which were not necessarily visually temporal events. Prior to this, most research had focused on image captioning ((Lin et al., 2014; Karpathy and Fei-Fei, 2015; Vinyals et al., 2015; Xu et al., 2015; Chen et al., 2015; Young et al., 2014). In [16] the image sequence is encoded using an RNN model and the story language was generated using the Gated Recurrent Units (GRU) for both image encoders and story decoder. All results are compared against human generated story languages.

Visual Story Post-Editing

In VIST-Edit [23] Hsu et. al. introduce the first dataset for human edits of machine-generated visual stories and explore how these collected edits may be used for the visual story post-editing task. The stories were generated by two state-of-the-art visual storytelling models, each aligned to 5 human-edited versions. The research shows that a relatively small set of human edits can be leveraged to boost the performance of large visual storytelling models.

A Hierarchical Approach for Visual Storytelling Using Image Description

Nahian et. al. [26] state the challenges of visual storytelling is in developing techniques that can maintain the context of the story over long event sequences to generate human-like stories. They propose a hierarchical deep learning architecture based on encoder-decoder networks to address the problem. To better help their network maintain this context while also generating long and diverse sentences, they incorporate natural language image descriptions along with the images themselves to generate each story sentence. They evaluate the system on the Visual Storytelling (VIST) dataset and show that their method outperforms state-of-the-art techniques on a suite of different automatic evaluation metrics BLEU, CIDEr, METEOR, and ROUGE-L.

Methods

Dataset

We use the VIST dataset [Huang et. al. 16] which includes 2 sets of text annotations:

1. **Descriptions of Images-in-Isolation (DII)**: These are captions of the images generated using MS COCO dataset.

2. **Stories of Images-in-Sequence (SIS)**: These are human-generated stories on sets of 5 images appearing in a sequence in the image dataset.

SIS comprises 40,155 5-image stories in the training set across 167528 images, 4990 stories in the validation set (21048 images) and 5055 stories (21075 images) in the test set. We used the SIS dataset to fine-tune our model.

Model & Evaluation Metric

We generated texts using multiple GPT-2 models, using METEOR [24] as the evaluation metric because it was the primary metric used and proposed by Huang et al. [16], our baseline research.

Vanilla GPT-2: We used the vanilla GPT-2 model from huggingface.co [17]. We experimented with generated text length sizes of 25 to 30 depending on the prefix text. We used VIST DII text from validation dataset, i.e., regular image captions as the prefix text for our model. For each prefix text we generated 5 sequences and picked the one with the highest METEOR score. In case multiple texts returned the same score we picked the most coherent one using human judgement.

Fine-tuned GPT-2 using VIST training dataset: We next fine-tuned the GPT-2 model using the VIST SIS training dataset (40,155 5-image stories) and generated 5 sequences of texts for each VIST DII text. We generated texts having lengths of 15 and 250. We evaluated the 5 generated text sequences (hypothesis) against the corresponding VIST validation set SIS stories (reference) using METEOR as our evaluation metric. Here again, for each prefix text we picked the one with the highest METEOR score. In case multiple texts returned the same score we picked the most coherent one using human judgement.

A key learning was that we had to limit the length of generated text to the SIS text length, to get higher METEOR scores, our primary evaluation metric. We believe this is expected as METEOR relies on the lengths of the reference and hypothesis texts and similar text lengths should result in a higher score.

Results and Discussion

Below are our key findings and are summarized in Table.1 (we did a spot check using a few examples from VIST validation set):

- Compared to baseline [16] METEOR scores are higher for text generated from vanilla GPT-2 for individual images

- Shorter generated text results in higher METEOR scores (score from longer texts not reported for brevity)
- METEOR score of combined narratives was much lower than baseline [16] (Table.2)

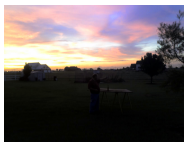
Image Seq ->	**1**	**2**	**3**	**4**	**5**
VIST - Val set					
a) VIST - DII	**Six or seven bottle rockets next to a specialty firework.**	**the burgers are being grilled on the grill.**	**A woman has big forks with meat on them.**	**An older man stands in the entry of a garage, manning his station at the grill.**	**A man is standing in front of some saw horses at dusk.**
METEOR	23.55	23.55	**23.55**	23.55	23.55
b) VIST - SIS (Ground Truth story)	We gathered up some fire works to set off at dark.	We grilled some burgers that were oddly shaped.	Our friends trickled in one by one!	We got all the food on the grill to cook and ready to eat.	The sunset was amazing that night!
c) Fine Tuned GPT-2	A cool ceramics set for the firework show. Our first fire	The food is being served and will be eaten. The food is served and	Her husband is taking a picture. The woman is eating the meat. They	Next to the grill, a man stands in the entryway, making sure	The man has seen a lot of horses .the man looks around .the'
METEOR	25.00	21.74	0.00	21.74	22.73
d) Vanilla GPT-2	And there really never was a rocket launcher to say it	It\'s kind of like a burger to my mind.\n\n"Are they	If someone wants to cut their fork open to taste their milk, their fork	\n\nHe is carrying a heavy bag of chips.	The horse, called a "mewing-rope
METEOR	**25.00**	**25.00**	0.00	**29.41**	**35.71**

Table.1: Generated texts with METEOR Scores. a) MS COCO image captions, b) Human generated story language; c) Machine generated text with GPT-2 fine-tuned on VIST train set, d) Machine generated story language using vanilla GPT-2.

Baseline	Fine-tuned GPT-2	Vanilla GPT-2
23.13	6.76	8.62

Table.2: METEOR scores of combined narratives (story across all 5 images)

Our primary evaluation metric is METEOR and much has been written about the need for a better evaluation metric than BLEU [13], ROUGE [14], METEOR [24] that rely on the occurrences of words in references rather than on coherence or similarities of context [19]. We would also like to propose using a different evaluation metric than METEOR as we believe we were able to get higher METEOR scores by optimizing the length of generated text and not necessarily improving the coherence of the story. This is similar to findings of [McCoy et. al. 27].

Our assertion is evident from the image sequence 3 in Table 1. We would argue for a prefix text of "**A woman has big forks with meat on them",** the generated text **"Her husband is taking a picture. The woman is eating the meat."** seems more coherent than human-generated text, **"Our friends trickled in one by one!".** Thus, a metric such as Bert Score may have been more appropriate [25] for this task.

Next Steps

So far we have tested the performance of the 2 models (vanilla GPT-2 and fine-tuned GPT-2) on few examples from VIST validation dataset [16]. Over the remaining weeks, we plan to run the models on the entire validation dataset and test our most optimal model on the VIST test dataset. Additionally, we will continue to experiment for improving our GPT-2 model performance for the combined narrative of the story languages generated for individual images.

Some modeling techniques we plan to explore include:

- Using coherence techniques [28]
- Fine-tuning using relevant genre based story corpus [29]

Code Repository

Evaluation: w266-final-project/W266-Neural-Narratives-Using-GPT-2-DRAFT.ipynb at main · smahensaria/w266-final-project (github.com)

Fine-Tuning: w266-final-project/W266_Neural_Narratives_Fine_Tune_GPT_2_DRAFT.ipynb at main · smahensaria/w266-final-project (github.com)

References

1. MD. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, Hamid Laga. *A Comprehensive Survey of Deep Learning for Image Captioning. ACM Computing Surveys.* October, 2018 - https://arxiv.org/pdf/1810.04020.pdf

2. Nayyer Aafaq, Ajmal Mian, Wei Liu, Syed Zulqarnain Gilani, and Mubarak Shah. *Video Description: A Survey of Methods, Datasets and Evaluation Metrics. ACM Computing Surveys. Vol. 52 Issue 6*. 2020 - https://arxiv.org/pdf/1806.00186.pdf

3. Pix2Story: Neural storyteller which creates machine-generated story in several literature genre

4. How do you watch a movie if you can't see? Blind people answer (cnet.com)

5. Image captioning with visual attention | TensorFlow Core

6. Deep Learning, NLP, and Representations, Chris Olah's blog, 2014

7. Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, Piotr Dollár. "*Microsoft COCO: Common Objects in Context."* 2015 - [1405.0312] Microsoft COCO: Common Objects in Context

8. Kenneth Tran, Xiaodong He, Lei Zhang, Jian Sun, Cornelia Carapcea, Chris Thrasher, Chris Buehler, Chris Sienkiewicz. "*Rich Image Captioning in the Wild."* 2016 - ImageCaptionInWild-1.pdf

9. Takako Aikawa, Lee Schwartz, Michel Pahud. "*NLP Story Maker."* 2005- Microsoft Word - NLP_Story_Maker_revisedFinal.doc

10. Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean. "*Efficient Estimation of Word Representations in Vector Space."* 2013 - 1301.3781.pdf: Efficient Estimation of Word Representations in Vector Space

11. Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, Sanja Fidler. "*Skip-Thought Vectors."* 2015 - https://arxiv.org/pdf/1506.06726.pdf

12. Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, Dario Amodei. "*Language Models are Few-Shot Learners."* 2020 - 2005.14165.pdf: Language Models are Few-Shot Learners

13. BLEU: a Method for Automatic Evaluation of Machine Translation. *Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu.* bleu.dvi (aclweb.org)

14. ROUGE: A Package for Automatic Evaluation of Summaries. *Chin-Yew Lin* - Lin.PDF (aclweb.org)

15. Crime scene investigation, a guide for law enforcement - https://www.nist.gov/system/files/documents/forensics/Crime-Scene-Investigation.pdf
16. Visual Storytelling. *Ting-Hao (Kenneth) Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, Margaret Mitchell*.

17. OpenAI GPT2 — transformers 4.7.0 documentation (huggingface.co)

18. Automatic Story Generation: Challenges and Attempts. *Amal Alabdulkarim, Siyan Li, Xiangyu Peng*. 2021 - Proceedings of the Third Workshop on Narrative Understanding

19. Fabula Entropy Indexing: Objective Measures of Story Coherence. *Louis Castricato, Spencer Frazier, Jonathan Balloch, and Mark O. Riedl.* 2021 - Proceedings of the Third Workshop on Narrative Understanding

20. Proceedings of the First Joint Workshop on Narrative Understanding, Storylines, and Events

21. Proceedings of the Second Workshop on Storytelling

22. Proceedings of the First Workshop on Storytelling

23. Visual Story Post-Editing. *Ting-Yao Hsu, Chieh-Yang Huang, Yen-Chia Hsu, Ting-Hao (Kenneth) Huang*. 2019 - [1906.01764] Visual Story Post-Editing (arxiv.org)

24. METEOR - Wikipedia

25. A Survey of Evaluation Metrics Used for NLG Systems. *Ananya B. SAi, Akash Kumar Mohankumar, Mitesh M. Khapra.* 202. - A Survey of Evaluation Metrics Used for NLG Systems (arxiv.org)

26. A Hierarchical Approach for Visual Storytelling Using Image Description. *Md Sultan Al Nahian, Tasmia Tasrin, Sagar Gandhi, Ryan Gaines, and Brent Harrison*. 2019 - 1909.12401v1.pdf (arxiv.org)

27. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. *R. Thomas McCoy, Ellie Pavlick, & Tal Linzen*. 2019 - 1902.01007.pdf (arxiv.org)

28. Centering-based Neural Coherence Modeling with Hierarchical Discourse Segments. *Sungho Jeon and Michael Strube*. 2020 - https://aclanthology.org/2020.emnlp-main.604.pdf

29. Project Gutenberg - https://www.gutenberg.org/