

# W266 Section 3 Final Project

## Neural Narratives for Multiple Images

[Eric Lee](#), [Sundeep Kumar Mahensaria](#),  
{ericlee.2020, smahensaria}@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’ and develops a narrative across multiple ‘discrete yet related’ images. Most existing research [[16](#), [26](#), [35](#)] starts by generating image captions and then using the captions to generate narratives across images. Additionally, we are not aware of any research that uses pre-trained language models (LM) to generate abstract story narratives. In this paper we explore neural text generation techniques to develop a story across multiple images on VIST dataset [Huang et. al., [16](#)] applying transfer learning using a powerful language model. Instead of first captioning the images, we use the image captions already available in the VIST data set as proxies for the images and generate our stories using a fine-tuned GPT-2 Medium model. We find that though the generated stories are coherent and contextual, our model under-performs the state-of-the-art (SOTA) results on VIST. Additionally, we observe that the scores of story-languages generated using our fine-tuned GPT-2 Medium model out-perform the original (vanilla) GPT2 Medium model by a factor of up to 2.7.<sup>1</sup>

## 1. 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 for 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 capturing appropriate emotions without proper context [[18](#)].

GPT-2 is a powerful transformer-based language model trained on a massive data set, which can be used for text generation. In this paper we use the GPT-2 Medium (345M parameters) model on the VIST dataset [[16](#)] to generate story language from the image captions. For the scope of this NLP related project, we do not generate our own captions from the images. Instead, for model benchmarking we leverage descriptions of images in isolation (DII) and stories of images in sequence (SIS) from the VIST dataset (Section 3.1), as image captions and reference texts respectively. Effectively, we research whether existing models that are SOTA in their respective individual tasks can be combined to address a challenging composite task.

We find that the story languages generated from our fine-tuned model GPT-2 [[17](#)] outperforms the vanilla (pre-trained) GPT-2 model by a factor of up to 2.7. Although our fine-tuned model underperforms the models in existing work [[16](#), [26](#), [35](#)], these results could inform areas of research such as - a) evaluation of machine-generated stories need not rely on human-generated stories as gold standard; b) visual storytellers can use transformer-based LM models to generate stories.

## 2. 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 [[39](#)]; Karpathy and Fei-Fei, 2015 [[38](#)]; Vinyals et al., 2015 [[41](#)];

---

<sup>1</sup> Code available at <https://github.com/smahensaria/w266-final-project>

Xu et al., 2015 [42]; Chen et al., 2015 [40]; Young et al., 2014 [43]). 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 and reported as METEOR scores.

**Adversarial Reward Learning for Visual Storytelling (AREL):** Wang et al. [35, 2018] apply a reinforcement learning framework for visual storytelling. The framework consists of two modules, a policy model, and a reward model. The policy model, a CNN-RNN architecture, takes an image sequence as input and chooses words from a vocabulary (generated from VIST training set) to form the story narrative. The reward model, a CNN-based architecture, is optimized for adversarial objectives. The AREL model is trained and evaluated on the VIST data set [16] and the authors report BLEU, METEOR, ROUGE-L-Recall, and CIDEr scores. Additionally, they employ Amazon Mechanical Turk for human evaluation.

**GLocal Attention Cascading Networks for Multi-image Cued Story Generation (GLAC Net):** Kim et al. [36, 2019] propose a deep learning network model that generates visual stories by combining global-local (glocal) attention and context cascading mechanisms. The model uses a two-level attention architecture and maintains sentence coherence by cascading the information of previous sentence to the next sentence serially. The model is also evaluated on the VIST data set and authors report METEOR score.

**A Hierarchical Approach for Visual Storytelling Using Image Description (HCBNet):** Nahian et al. [26, 2019] develop 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, HCBNet, 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.

AREL is the most performant amongst these. The results of these research are compared in Table 2. In this paper we explore text generation using GPT-2 using image captions already available in VIST data set.

### 3. Methods

#### 3.1 Dataset

We use the VIST dataset [Huang et. al. 16] which includes 2 sets of text annotations (Figure 1):

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

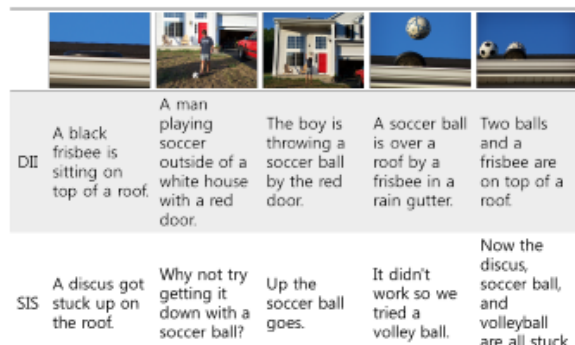


Figure 1. Image annotations in VIST data set

SIS comprises 40,155 5-image stories in the training set across 167,528 images, 4,990 stories in the validation set (210,48 images) and 5,055 stories (21,075 images) in the test set. An image can have multiple DII captions as well as multiple SIS story languages. We use both DII and SIS to fine-tune our model.

#### 3.2 Evaluation Metric

We use METEOR [24] as the core evaluation metric because it was the primary metric used and proposed by Huang et al. [16], and is the common metric in all our other reference research (Section 2). Some authors also report BLEU, ROUGE-L-Recall, and CIDEr scores. We also report ROUGE-L-Recall and BERTScores - primarily to measure the coherence of our stories to human generated texts. All reported scores are average across all stories in the test data set.

### 3.3 Models

**Baseline Model:** We use the vanilla GPT-2 model from huggingface.co [17] as our baseline model for text generation. We concatenate the longest DII texts (i.e., image captions) and pass them as prefixes to GPT-2’s text generation pipeline. We evaluate the generated text with corresponding SIS for each story (i.e., narrative across 5-images) and report the METEOR, ROUGE-L-Recall and BERTScore scores.

We explore several hyperparameters with 2 decoders – 1) Beam Search decoder with beam sizes of 1, 3, 5, and 10; 2) KP Sampling decoder with K=50, P= 0.8, 0.9, and 0.95. Considering the coherency of the generated text, we observed the most optimal average METEOR scores for Beam Size=5 and P=0.95 and thus, used Beam Size 5 and KP Sampling with P=0.95 for our fine-tuned models. Table 1 shows the METEOR scores resulting from different hyperparameters on partial validation set.

| Decoder | Beam Search |      |      |       | KP Sampling |       |       |        |
|---------|-------------|------|------|-------|-------------|-------|-------|--------|
|         | BS-1        | BS-3 | BS-5 | BS-10 | K50         | P=0.8 | P=0.9 | P=0.95 |
| METEOR  | 9.8         | 14.5 | 15.1 | 15.1  | 7.1         | 5.4   | 7.3   | 7.2    |

Table 1: METEOR Scores with different hyperparameters on partial validation set using Baseline model

#### Fine-tuned GPT-2 using VIST training dataset:

We next fine-tune the GPT-2 Medium model using the VIST DII and SIS training data set (40,155 5-image stories). We use the nouns in DII text as hints and concatenate them with SIS text as the ‘label’ using the list of tokens in (1)

<BOS> <HINT> unique nouns extracted from the longest DII <SENT> SIS text <EOS> (1)

Our intuition is that the Nouns serve as proxies for objects in the images. We experimented with different <HINT> types – full DII text, longest DII text, unique nouns in longest DII, all nouns in DII (repetitions allowed), nouns and verbs in DII, nouns in SIS, – and ultimately observed unique nouns from the longest DII to report most optimal results. We noticed a decrease in perplexity - 34.73 for all nouns in DII (repetitions allowed) to 28.33 for unique nouns only in longest DII.

We generate texts having lengths of 500 and keep the first 5 sentences or first 75 words, as applicable. We evaluate this 5-sentence/75 word generated text

sequence (hypothesis) against the corresponding SIS stories (reference) using METEOR as our evaluation metric.

Initially we started the model fine-tuning with 3 epochs and noticed overfitting behavior after 1 epoch. We thus trained our model for 1 epoch only [44].

For both models (baseline and fine-tuned) we experiment with the same two decoding methods (i.e., Beam Search and KP Sampling) and hyperparameters as the baseline model. We also experimented with different temperature settings of 0.5, 0.7, and 1.2. We observed the most optimal average METEOR scores for Beam Size=5 and KP Sampling with P=0.95 and temperature = 0.7. Table 2 shows the METEOR scores resulting from different hyperparameters on partial validation set.

| Decoder | Beam Search |      |      |       | KP Sampling |       |       |        |
|---------|-------------|------|------|-------|-------------|-------|-------|--------|
|         | BS-1        | BS-3 | BS-5 | BS-10 | K50         | P=0.8 | P=0.9 | P=0.95 |
| METEOR  | 16.1        | 17.9 | 17.3 | 16.7  | 12.5        | 13.2  | 15.1  | 14.8   |

Table 2: METEOR scores with different hyperparameters on partial validation set using Fine-tuned model

Additionally, we experimented with hyperparameters such as generated text lengths, repetition penalty (Beam Search), and repeat n-grams. We also experimented with GPT-2 Large (762M parameters) and observed the performance of the baseline model to be better by a factor of 2 when compared to GPT-2 Medium baseline. However, it’s performance was marginally lower than the Medium fine-tuned model. We thus decided to use the GPT-2 Medium model to optimize for computation time. We observed that GPT-2 Large (Figure 3 in Appendix) was underfitting the validation data set while GPT-2 Medium converged (Figure 2 in Appendix).

## 4. Results and Discussion

Table 3 illustrates our results using an example. Row a) is the 5 images sequence, b) Human-generated DII (image captions), c) Human-generated SIS (stories) and the ground truth/gold standard, d) Fine-tuned GPT-2 Medium model generated story using Beam Search decoding, e) Fine-tuned GPT-2 Medium model generated story using KP Sampling decoding, f) Vanilla GPT-2 Medium’s generated story. Table 4 shows the scores of our Vanilla and Fine-tuned GPT-2 models

along with other research [16, 26, 35, 36] with their respective SOTA results.

Our key findings are in Table 4:

- All our GPT-2 based models underperform the prior research [16, 26, 35, 36]. Our fine-tuned KP Sampling model achieves a maximum METEOR score of 19.4 while all prior research exceeds 30. We were unable to find the source IDs for examples quoted in these research so we were unable to compare our scores with them for the individual examples.
- GPT-2 models: KP Sampling decoded model scores higher on METEOR compared to Beam Search – 19.4 vs 17.9. However, the Beam Search decoded model results in a much higher BERTScore – 26.0 vs 15.5.
- Since our models rely on DII text (image captions) as prefix hints, we also calculated the scores of DII vis a vis SIS as reference. We notice that the METEOR and ROUGE scores are consistent with our models’ generated text.
- AREL [35] was the only research that reported ROUGE-L-Recall. Our model underperformed AREL w.r.t. ROUGE-L-Recall, 19.8 vs 29.6.






## Analysis

Though GPT-2 is a very powerful language model, we notice that it under-performs non-transformer based SOTA models. On analyzing the data, we believe the following reasons likely contribute to these results:

- The authors of our related prior research all trained their models on VIST training data set only, while a pretrained model like GPT-2 has a very generic corpus. Although we notice the METEOR scores are up to 2.7 times better for the fine-tuned model, we believe the VIST training data set, used for fine-tuning is a very small proportion of the GPT-2 original corpus.
- The prior work generates text on an image-to-image basis, while GPT-2 generates contiguous

text using a prefix of <HINT> tokens (Section 3.3). Thus, the coherence of generated text relies heavily on pre-trained corpus – this could be a limitation of using pre-trained language models.

- Though DII and SIS are both human generated, very few tokens are common between the two, because the SIS is an abstractive version of DII. This is highlighted by their low METEOR and ROUGE-L (Recall) scores in Row A) in Table 4.
- GPT-2 generated texts are an extractive version of DII text and hence are expected to be more ‘aligned’ to DII than SIS.
- We did a spot check of GPT-2 generated texts and they all seemed quite coherent and with minimal repetitions (Table 3, Rows d & e). Since we could not find any related prior work that calculates BERTScore, we were unable to compare our results with any SOTA results.
- Our primary evaluation metric is METEOR and much has been written about the need for a better automated 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 5 (Appendix). 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 BERTScore may have been

| Image Seq ->  | 1  | 2   | 3  | 4   | 5   |
|---|--|---|--|---|---|
| a) VIST – Test Data set<br>(Story ID: 47653)                                |   |  |  |  |  |
| b) VIST – DII (Description-In-Isolation)<br>Human-generated image captions  | A woman is at a craft fair and appears to have just blown up a beautiful marbled balloon. She is reaching into a yellow pail. She is smiling and happy.  | An Asian woman is playing the cello.  | Two Asian girls in white blouses playing wind instruments.                         | A young woman carrying a large instrument in front of a crowd.                      | Young boy consuming cotton candy at outdoor event.                                  |
| c) VIST - SIS (Story-In-Sequence)<br>Human-generated stories (Ground Truth) | I went to the festival today.  | There were a lot of people performing.  | The orchestra was very good  | Some of the musicians were very talented.   | I had a great time there.   |
| d) Fine-tuned GPT-2 generated story (Beam Search – Beam = 5)                | We went to the fair today. There were a lot of people there. Some of them were playing music. We had a great time there. We got to play some music.  |   |  |   |   |
| e) Fine Tuned GPT-2 generated story (KP Sampling)                           | I went to the arts and crafts festival today. There were a lot of interesting things there. I had a great time playing with all kinds! We played some folk songs. Afterward, my family came over for dinner at my house. |   |  |   |   |
| f) Vanilla GPT-2 generated story  | He licks his lips. The words 'hillbilly' were played during the first half of The Big Bang Theory, but was cut after episode 10 ended when Sam became gay.   |   |  |   |   |

**Table 3:** Generated texts. a) VIST image sequence; b) Human-generated VIST image captions; c) Human generated story language; d) Machine generated text with GPT-2 fine-tuned on VIST train set, e) Machine generated story language using vanilla GPT-2.

| Method   | METEOR                                      | ROUGE-L-R                                   | BERTScore-P                                 |
|--|---|---|---|
| A) Human generated; DII vs SIS   | 16.5  | 14.3  | -   |
| B) Our Models (Vanilla GPT-2 Medium)<br>Beam Search (beam size=5) - Baseline<br>KP Sampling (P=0.95) - Baseline                                      | 13.8<br>7.3                                 | 15.0<br>8.1                                 | -3.0<br>-14.3                               |
| C) Our Models (Fine-tuned GPT-2 Medium)<br>Beam Search (beam size=5)<br>(Validation data set*)<br><br>KP Sampling (P=0.95)<br>(Validation data set*) | 17.9<br>(22.6)<br><br><b>19.4</b><br>(22.5) | 15.0<br>(20.9)<br><br><b>15.8</b><br>(18.4) | <b>26.0</b><br>(15.6)<br><br>15.5<br>(10.9) |
| D) Huang et al [16] (Beam Size = 10)<br>Huang et al [16] (Grounded)  | 23.1<br>31.4                                | -<br>-                                      | -<br>-                                      |
| E) AREL-t-50 [35]<br>F) GLAC Net [36]<br>G) HCB Net [26]   | <b>35.2</b><br>30.6<br>34.0                 | <b>29.6</b><br>-<br>-                       | -<br>-<br>-                                 |

**Table 4:** Evaluation metrics of combined narratives (story across all 5 images). A) Human generated VIST DII, SIS texts; B) Our baseline model generated texts; C) Our Fine-tuned models generated texts (\* Results for the validation data set); D) VIST data set; E) AREL [35], GLAC Net [36], HCB Net SOTA scores [26]



more appropriate [25] for this task. We were not able to find any research on this task that reported BERTScore.

## 5. Conclusion

Our model applies existing SOTA task models to the composite task of visual storytelling. We use human generated image captions to generate stories using a fine-tuned GPT-2 Medium language model, instead of generating our own. When evaluated against the VIST data set, our fine-tuned GPT-2 Medium model underperforms the SOTA models but reports up to 2.7x improvement over vanilla GPT-2 model. We observe that when we evaluate human-generated image captions (DII) against human-generated stories (SIS) for the same images, the METEOR scores are aligned with our model generated scores. A key promising outcome is on spot checking, we find the generated stories to be coherent and contextual. However, we were not able to compare coherency of our results with any SOTA work, since we were not able to find any research that reports coherency-based scores such as BERTScore.

We believe our research has a potential to reduce dependance on human text generation and evaluation.

In the future, we plan to continue exploring fine-tuning techniques and optimize for automated scores such as BERTScore to get closer to human-generated text coherence. We expect to find prior research to validate our results against.

## Acknowledgements

We thank Daniel Cer for his feedback and numerous suggestions to keep our research and model development on track. We are also thankful to Mark Butler, Joachim Rahmfeld and Zack Alexander for their guidance in evaluating our model.

## References

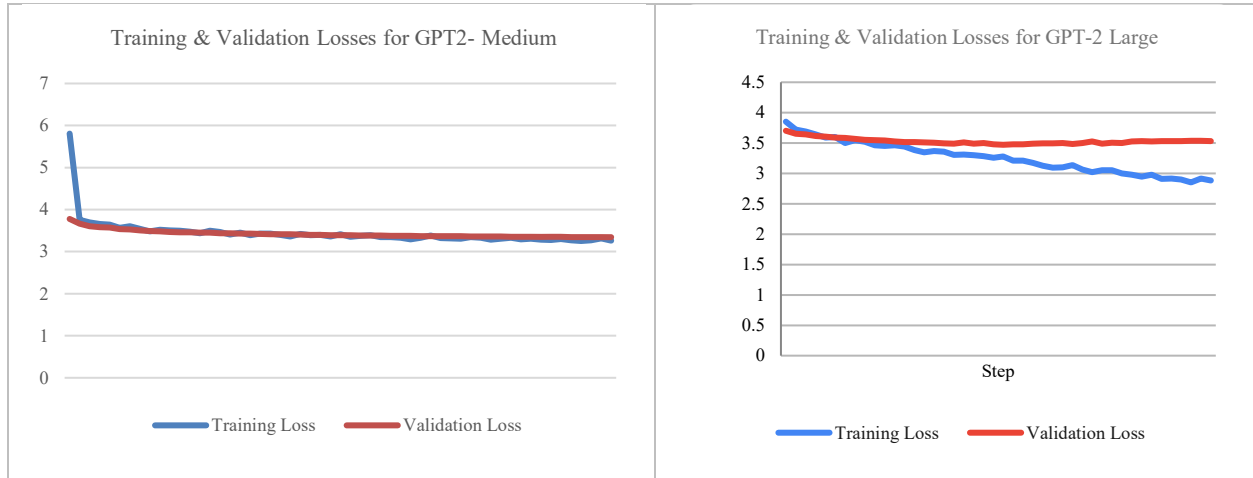
1. MD. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, Hamid Laga. *A Comprehensive Survey of Deep Learning for Image Captioning*. *ACM Computing Surveys*. October, 2018 - <https://arxiv.org/pdf/1810.04020.pdf>
2. Nayyer Aafaq, Ajmal Mian, Wei Liu, Syed Zulqarnain Gilani, and Mubarak Shah. *Video 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/>
30. BERTScore: Evaluating Text Generation with BERT. Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, Yoav Artzi. 2019 - [\[1904.09675\] arxiv.org](#)
31. Contextualize, Show and Tell: A Neural Visual Storyteller. Diana Gonzalez-Rico and Gibran Fuentes-Pineda. 2018 - [1806.00738.pdf \(arxiv.org\)](#)
32. Hierarchically-Attentive RNN for Album Summarization and Storytelling. Licheng Yu and Mohit Bansal and Tamara L. Berg. 2017 - [licheng\\_emnlp2017.pdf \(tamaraberg.com\)](#)
33. [Plan-and-Write: Towards Better Automatic Storytelling | Proceedings of the AAAI Conference on Artificial Intelligence](#). Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, Rui Yan. 2019

34. How to generate text: using different decoding methods for language generation with Transformers - <https://huggingface.co/blog/how-to-generate>
35. No Metrics Are Perfect: Adversarial Reward Learning for Visual Storytelling. *Xin Wang, Wenhui Chen, Yuan-Fang Wang, William Yang Wang*. 2018 - <https://aclanthology.org/P18-1083.pdf>
36. GLAC Net: GLocal Attention Cascading Networks for Multi-image Cued Story Generation. *Taehyeong Kim, Min-Oh Heo, Seonil Son, Kyoung-Wha Park, Byoung-Tak Zhang*. 2019 - [1805.10973.pdf \(arxiv.org\)](https://arxiv.org/abs/1805.10973)
37. [How to generate text: using different decoding methods for language generation with Transformers \(huggingface.co\)](https://huggingface.co/blog/how-to-generate)
38. Andrej Karpathy and Li Fei-Fei. 2015. Deep visual semantic alignments for generating image descriptions. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June.
39. Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In Computer Vision–ECCV 2014, pages 740–755. Springer.
40. Jianfu Chen, Polina Kuznetsova, David Warren, and Yejin Choi. 2015. D’ej’a image-captions: A corpus of expressive descriptions in repetition. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 504–514, Denver, Colorado, May–June. Association for Computational Linguistics.
41. Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In Computer Vision and Pattern Recognition.
42. Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. arXiv preprint arXiv:1502.03044.
43. 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.
44. One Epoch Is All You Need. *Aran Komatsuzaki*. 2019 - <https://arxiv.org/abs/1906.06669>



## APPENDIX



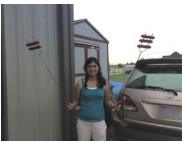
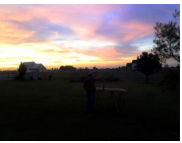
| Image Seq ->                          | 1  | 2  | 3   | 4  | 5  |
|---------------------------------------|--|--|---|--|--|
| VIST - Val set<br>Story ID: 40471     |  |  |  |  |  |
| a) VIST - DII                         | Six or seven bottle rockets next to a specialty firework.                          | 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.                               |
| b) VIST - SIS<br>(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.                      | Next to the grill, a man stands in the entryway, making sure <truncated>             | The man has seen a lot of horses. The man looks around.                              |

Table 5: Text generated for each image caption. a) Human generated image captions, b) Human generated story language; c) Machine generated text with Fine-tuned GPT-2 model.