Predicting Visual Futures with Image Captioning and Pre-Trained Language Models

Anonymous ACL submission

Abstract

The task of visual forecasting deals with predicting future events from a sequence of input images. Purely pixel-based approaches find this challenging due to the presence of abstract concepts and temporal events at different timescales. In this paper, we present an approach that combines image captioning with pre-trained language models to predict visual futures. By leveraging language as an intermediate medium, our model is able to perform more effective temporal reasoning on two different tasks – visual story cloze and action forecasting. Despite making the final predictions using only the generated captions, our approach outperforms state-of-the-art systems by 4% and 6% respectively on the two tasks. We find that our model consistently picks images/actions that are semantically relevant to the given image sequence instead of simply relying on visual similarity.[1]

1 Introduction

Predicting future events based on past observations is useful for autonomous agents to navigate the world. Several recent works in computer vision and reinforcement learning have developed models that learn to predict or generate future observations (Xu et al., 2018; Isola et al., 2017; Ebert et al., 2018), with one goal being to use such predictions to inform control policies (Ha and Schmidhuber, 2018; Hafner et al., 2019a; Schrittwieser et al., 2020; Hafner et al., 2019b).

However, such approaches usually work directly on pixel-based inputs (or build on top of visual features from pre-trained models), which makes it challenging to accurately capture and reason over varying levels of temporal abstraction. In this paper, we explore the use of natural language as a medium for predicting visual futures, building on recent insights that pre-trained language models can perform temporal reasoning (Vashishtha et al., 2020; Han et al., 2020). Specifically, we first use image captioning to describe frames in a sequence of images, and then train a model that can reason temporally over the generated captions to predict future events. For the latter, we make use of pre-trained language models such as RoBertA (Liu et al., 2019) and fine-tune them to predict the required quantity in the future (e.g. picture that completes a story or an anticipated action). As our experiments show, our use of captions allows for temporal reasoning over a diverse set of abstract concepts and timescales.

We compare our method with existing models on two tasks – (1) visual story cloze, where the goal is to pick an image that completes a sequence of images to form a coherent story, and (2) action forecasting, where a model has to predict a future action. Surprisingly, despite not using image features to make the final predictions and relying only on captions, our approach outperforms the baselines on both tasks, by 4% and 6%, respectively. Our analysis reveals that most of this gain comes from the language model leveraging the high level concepts in the generated captions to predict semantically coherent future events.

2 Related Work

Future forecasting in vision and NLP Recent work has explored ideas around generating future images (Villegas et al., 2019; Ha and Schmidhuber, 2018; Hafner et al., 2019a; Schrittwieser et al., 2020; Hafner et al., 2019b), inferring trajectories and future actions based on past observations (Zeng et al., 2017), or predicting temporal orderings (Sigurdsson et al., 2016). These approaches require learning good visual feature representations that can capture temporal structure, which inherently makes it challenging to model long-range temporal events since capabilities like object tracking (Yilmaz et al., 2006) and optical flow (Fortun et al.,

[1] Code provided in supplementary material.

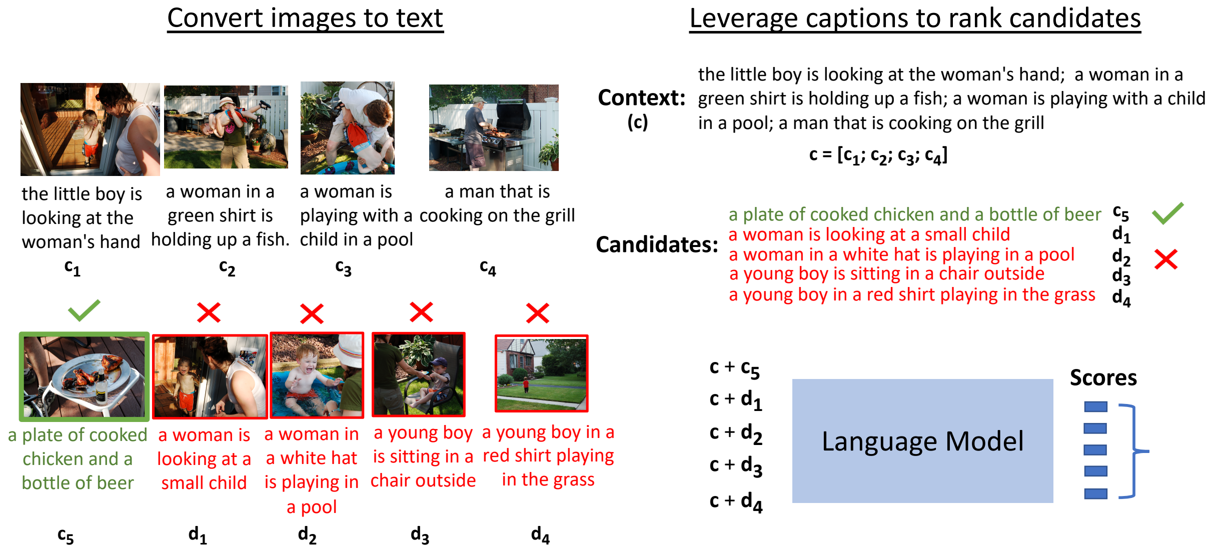

Figure 1: (Left) Visual forecasting for *Story Cloze*: given a set of 4 context images, a model is tasked to predict the most likely future image among 5 candidate images. Pixel-based approaches such as (Zeng et al., 2017) make an incorrect prediction (d_4) since they rely heavily on visual similarities rather than semantic consistency or temporal reasoning (e.g. "cooking on the grill" results in a "plate of cooked chicken"). Our approach generates captions for all the images and uses the generated text to rank all the candidate completions with a language model (right).

2015) are more suited for prediction over shorter timescales ($\sim$ 10-20 seconds). In our work, we leverage the textual modality to better reason over various timescales (e.g. minutes, hours, days).

Future forecasting in NLP includes story ending prediction (Mostafazadeh et al., 2016; Cui et al., 2020; Cai et al., 2017; Chaturvedi et al., 2017; Li et al., 2019; Chen et al., 2019), temporal ordering anticipation (Ning et al., 2020, 2018; Zhou et al., 2019), future information retrieval (Baeza-Yates, 2005), and language models for storytelling (Ammanabrolu et al., 2019; Li et al., 2019; Yang and Tiddi, 2020). These works demonstrate the use of modern language models for temporal modeling of events, which forms a core part of our hypothesis.

Image captioning in downstream tasks Recent work has explored the use of image captioning (Lin et al., 2014; Li et al., 2020; You et al., 2016) in downstream tasks like visual question answering (Wu et al., 2019; Fisch et al., 2020) and image retrieval (Luo et al., 2018). While their primary goal is to improve captioning and its applicability to downstream tasks, our focus is on using the generated captions as a medium to perform temporal reasoning for predicting visual futures.

3 Our Approach

Task Setup Given a sequence of k temporally ordered images $I_1, ..., I_k$, our goal is to predict a quantity $y(I_{k+1})$ where I_{k+1} represents a future image continuing the temporal sequence, and y represents a property based on that image (e.g. an action or an image that completes a story). In this work, we consider only discriminative predictions and do not generate I_{k+1}.

Prior approaches train a model to directly predict $y(I_{k+1})$ using the input image frames. We wish to leverage image captioning to assist with this prediction. Therefore, we first caption the set of input images to produce a set of captions (captioning systems are described later in Section 3):

$$c_j = Caption(I_1, ..., I_j), \text{for j } \in [1 \cdots k] \quad (1)$$

Note that the generated caption might be conditioned on the entire history of past images.

Once we have captions, we simply concatenate them together with the relevant separator tokens and feed them into a pre-trained language model (LM) such as RoBERTa (Liu et al., 2019) to predict the required property $\hat{y}$:

$$\hat{y} = LM([c_1, ..., c_k]) \quad (2)$$

This LM is then fine-tuned using standard loss functions such as cross-entropy loss. The parameters of the captioning model are held fixed during this training. Given this general framework, we provide specific details for tasks below.

Visual Story Cloze In visual story cloze (Mostafazadeh et al., 2016), the goal is to predict the image that best completes (or closes) a story from a set of candidate choices. Formally, the goal is to predict the right I_{k+1} from a set of images that also contain m distractors $D_1, ..., D_m$. We generate captions for all the candidates to obtain c_{k+1} and $d_1, ..., d_m$, respectively. Each of these captions is concatenated with the context captions $c_1, ..., c_k$ and input into the language model to produce a score, $s = LM([c_1, ..., c_k, C])$ where $C \in \{c_{k+1}, d_1, ..., d_m\}$. These scores are then optimized with binary cross entropy loss.

Action Forecasting For this task (Patron et al., 2010), $y(I_{k+1})$ is an action in the future to be predicted. We pass the context captions $c_1, ..., c_k$ into the language model to predict y and fine-tune the language model with standard cross-entropy loss.

Converting Images to Captions We consider two options for generating captions:
1. Independent image captioning: Here, we generate captions for each image independently, i.e. $c_j = Caption(I_j)$. We use Oscar (Li et al., 2020), a state-of-the-art image captioning approach pretrained on millions of aligned image text corpora (Sharma et al., 2018; Plummer et al., 2015; Hudson and Manning, 2019) and finetuned on COCO captions (Lin et al., 2014), and label the model as "Oscar(pretrained)". For story cloze, we also finetuned an Oscar variant on captions from the training data and label this variant as "Oscar(finetuned)".
2. Story captioning: We experiment with with Reco-RL (Hu et al., 2020) and AREL (Wang et al., 2018) storytelling models that jointly produce captions for an entire sequence of images. Given the $Story$ operator, which extracts the last sentence from the generated story of the input image sequence, we generate text for the context and distractor images as follows:

$$c_j = Story([I_1, ..., I_j])\text{for j } \in [1 \cdots 5]$$

$$d_k = Story([I_1; I_2; I_3; I_4; D_k])\text{for k } \in [1 \cdots 4]$$

4 Experiments

Datasets: For visual story cloze, we follow (Zeng et al., 2017) and construct the future prediction task through storylines from the Visual Storytelling Dataset (Huang et al., 2016). The dataset consists of temporally-ordered sequence of 5 photos from a large subset of Flickr albums and provides GT stories and captions. Following Zeng et al. (2017), we randomly select 1 storyline from each album and sample 4 distractor images from the same Flickr album. Using the original split, we get 8024 training, 1011 testing, and 998 validation storylines.

Model	Validation R@1 ↑	Validation R@3 ↑	Test R@1 ↑	Test R@3 ↑
GAIL (Zeng et al., 2017)	24.77	65.80	22.48	64.95
Nearest Neighbor	22.67	63.09	24.26	62.27
LSTM	19.96	58.58	21.68	59.11
Oscar(finetuned) + RoBERTa	**29.66**	**68.54**	**28.39**	**69.14**
Oscar(pretrained) + RoBERTa	29.15	68.54	26.80	67.26
AREL + RoBERTa	27.38	64.79	22.97	62.08
ReCo-RL + RoBERTa	25.67	64.79	23.96	63.66
Human Baseline	-	-	31.00	-
Random	20.00	60.00	20.00	60.00

Table 1: Summary of results on the future image prediction task on both the validation and test splits.↑ indicates higher is better. ↓ indicates lower is better.

For action forecasting, we use the TV Human Interactions dataset (Patron et al., 2010), with 300 videos of 4 interactive actions ("Hug", "Kiss", "HighFive", "HandShake"), with a 50-50 split between train/test. We follow the same setup in Zeng et al. (2017) and use context images upto 1 second before the start of the action. We sample 3 images from the context images to make the prediction.
Baselines: We compare with several baselines, following Zeng et al. (2017):
1. *LSTM* (Hochreiter and Schmidhuber, 1997): This uses ResNet-101 (He et al., 2016) features for the context images to predict $y(I_{k+1})$.
2. *Nearest Neighbor(NN):* We extract ResNet-101 features for all candidates and pick candidate with the lowest L2 difference with the context feature.
3. *GAIL* (Zeng et al., 2017): This leverages General Adversarial Imitation Learning (GAIL) (Ho and Ermon, 2016) to model sequences of images (details in appendix A.5).
We also collected human baseline performances for the tasks (details in Appendix A.2).
Evaluation metrics: We rank scores of all the candidates for y(I_{k+1}), calculate the rank of the GT candidate and report Recall@k. We set k to 1, 3 for visual story cloze and 1 for action forecasting.
Pre-trained LMs: We experiment with the pretrained and randomly initialized variants of the RoBERTa (Liu et al., 2019), GPT-2 (Radford et al., 2018) and BERT (Devlin et al., 2019) LMs.

5 Results

Visual story cloze. From Table 1, we see that our best model, Oscar(finetuned) + RoBERTa, outper-

Figure 2: Comparing predictions on samples from the test split across different variants (GAIL in dashed purple, NN in dashed red, our Oscar(finetuned) + RoBERTa model in green) with captions generated from Oscar(finetuned). Our model predicts candidates which are most likely to occur in the future by leveraging the concepts in the captions, as opposed to the vision baselines which predict candidates which are visually similar to the context images (Best viewed in color).

Model	R@1
GAIL (Zeng et al., 2017)	45.8
Deep Regression ($K = 3$) (Vondrick et al., 2016)	43.6± 4.8
Oscar(pretrained) + RoBERTa	**52.0** $\pm$ 13.1
Oscar(pretrained) + GPT-2	51.0 $\pm$ 17.0
Oscar(pretrained) + BERT	49.0 $\pm$ 14.5
Human (Vondrick et al., 2016)	71.7
Random	25

Table 2: Performance on the TV Human Interaction dataset (Baselines from Zeng et al. (2017)).[2]

forms the closest vision-only baselines, GAIL and NN, by more than 4% on both R@1 and R@3 respectively. This is significant given that R@1 performance for humans is only $\sim 31\%$. The distractor images tend to be visually similar to the context images as they belong to the same Flickr album and might explain why vision baselines, which rely mostly on pixel similarity, do worse than our models, which are able to leverage language pre-training to predict the most likely concepts to occur in the future. We note that our approach is competitive even without access to GT captions(Oscar(pretrained) and Oscar(finetuning) differ only by $\sim$ 1.5% on R@1). An extensive comparison between different pre-trained LMs is in Appendix A.1.

Captions vs Stories (Table 1): The storytelling variants (ReCo-RL, AREL) perform much worse than the captioning variants. This is likely due to the storytelling models generating generic stories("They had a great time"), which are accurate but not descriptive, as opposed to captioning models which tend to generate more descriptive captions("Picture of man eating cake in the garden").

Qualitative samples (Figure 2): Both rows demonstrate examples where the vision baselines such as NN (**marked in red**) and GAIL **(marked in purple)** incorrectly predict candidates that are visually similar to the context images. In contrast, our model **(marked in green)** encodes all the important concepts in the sequence of images ("eating dinner", "man holding a baby on a couch") through captions and leverages language pretraining to correctly predict the future concept (e.g. "sitting on a couch") that is most likely to occur.

Action Forecasting Table 2 shows that our model Oscar(pretrained) + RoBERTa, outperforms the best vision baselines, GAIL, by more than 6% and thus show that language pretraining might be capturing meaningful information about action dynamics (e.g: "high five" is the likely action following "two men standing at a table").

6 Conclusion

We propose a novel approach that combines image captioning with pre-trained language models to predict visual futures. By leveraging language as an intermediate medium, our model is able to perform more effective temporal reasoning on two different tasks – visual story cloze and action forecasting. Surprisingly our system, which makes final predictions using only the generated captions, outperforms state-of-the-art systems by 4% and 6% respectively on the two tasks. Our model successfully encodes all the important concepts in the sequence of images through captions and leverages language pre-training to correctly predict the concepts likely to occur in the future.

[2] Standard deviation not available for Zeng et al. (2017)

References

Prithviraj Ammanabrolu, Ethan Tien, W. Cheung, Z. Luo, William Ma, Lara J. Martin, and Mark O. Riedl. 2019. Guided neural language generation for automated storytelling.

Ricardo Baeza-Yates. 2005. Searching the future. In *SIGIR Workshop MF/IR*, volume 5. Citeseer.

Zheng Cai, Lifu Tu, and Kevin Gimpel. 2017. Pay attention to the ending: Strong neural baselines for the roc story cloze task. In *ACL*.

Snigdha Chaturvedi, H. Peng, and D. Roth. 2017. Story comprehension for predicting what happens next. In *EMNLP*.

Gang Chen, Y. Liu, Huanbo Luan, Meng Zhang, Qun Liu, and Maosong Sun. 2019. Learning to predict explainable plots for neural story generation. *ArXiv*, abs/1912.02395.

Yiming Cui, Wanxiang Che, Wei nan Zhang, T. Liu, Shijin Wang, and Guoping Hu. 2020. Discriminative sentence modeling for story ending prediction. *ArXiv*, abs/1912.09008.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Frederik Ebert, Chelsea Finn, Sudeep Dasari, Annie Xie, Alex Lee, and Sergey Levine. 2018. Visual foresight: Model-based deep reinforcement learning for vision-based robotic control. *arXiv preprint arXiv:1812.00568*.

Adam Fisch, Kenton Lee, Ming-Wei Chang, J. Clark, and R. Barzilay. 2020. Capwap: Captioning with a purpose. *EMNLP*.

Denis Fortun, Patrick Bouthemy, and Charles Kervrann. 2015. Optical flow modeling and computation: A survey. *Computer Vision and Image Understanding*, 134:1–21. Image Understanding for Real-world Distributed Video Networks.

David Ha and Jürgen Schmidhuber. 2018. Recurrent world models facilitate policy evolution. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 2455–2467.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. 2019a. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. 2019b. Learning latent dynamics for planning from pixels. In *International Conference on Machine Learning*, pages 2555–2565. PMLR.

Rujun Han, Xiang Ren, and Nanyun Peng. 2020. Deer: A data efficient language model for event temporal reasoning. *arXiv preprint arXiv:2012.15283*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *CVPR*.

Jonathan Ho and Stefano Ermon. 2016. Generative adversarial imitation learning. In *NIPS*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*.

Junjie Hu, Yu Cheng, Zhe Gan, Jingjing Liu, Jianfeng Gao, and Graham Neubig. 2020. What makes a good story? designing composite rewards for visual storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7969–7976.

Ting-Hao Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. 2016. Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1233–1239.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6700–6709.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134.

Diederik Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer.

Zhongyang Li, X. Ding, and T. Liu. 2019. Story ending prediction by transferable bert. In *IJCAI*.

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 *ECCV*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

R. Luo, Brian L. Price, S. Cohen, and Gregory Shakhnarovich. 2018. Discriminability objective for training descriptive captions. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6964–6974.

N. Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, P. Kohli, and James F. Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *NAACL*.

Qiang Ning, Z. Feng, H. Wu, and D. Roth. 2018. Joint reasoning for temporal and causal relations. In *ACL*.

Qiang Ning, Hao Wu, Rujun Han, Nanyun Peng, Matt Gardner, and Dan Roth. 2020. Torque: A reading comprehension dataset of temporal ordering questions. *arXiv preprint arXiv:2005.00242*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Alonso Patron, Marcin Marszalek, Andrew Zisserman, and Ian Reid. 2010. High five: Recognising human interactions in tv shows. In *Proceedings of the British Machine Vision Conference*, pages 50.1–50.11. BMVA Press. Doi:10.5244/C.24.50.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, et al. 2020. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*.

Gunnar A Sigurdsson, Xinlei Chen, and Abhinav Gupta. 2016. Learning visual storylines with skipping recurrent neural networks. In *European Conference on Computer Vision*, pages 71–88. Springer.

Siddharth Vashishtha, Adam Poliak, Yash Kumar Lal, Benjamin Van Durme, and Aaron Steven White. 2020. Temporal reasoning in natural language inference. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4070–4078, Online. Association for Computational Linguistics.

Ruben Villegas, Arkanath Pathak, Harini Kannan, Dumitru Erhan, Quoc V Le, and Honglak Lee. 2019. High fidelity video prediction with large stochastic recurrent neural networks.

Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. 2016. Anticipating visual representations from unlabeled video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 98–106.

Xin Wang, Wenhu Chen, Yuan-Fang Wang, and William Yang Wang. 2018. No metrics are perfect: Adversarial reward learning for visual storytelling. In *ACL (1)*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jialin Wu, Zeyuan Hu, and R. Mooney. 2019. Generating question relevant captions to aid visual question answering. *ACL*.

Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 2018. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324.

Xinran Yang and Ilaria Tiddi. 2020. Creative storytelling with language models and knowledge graphs. In *CIKM*.

Alper Yilmaz, Omar Javed, and Mubarak Shah. 2006. Object tracking: A survey. *ACM Comput. Surv.*, 38(4):13–es.

Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659.

Kuo-Hao Zeng, William B Shen, De-An Huang, Min Sun, and Juan Carlos Niebles. 2017. Visual forecasting by imitating dynamics in natural sequences. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2999–3008.

Ben Zhou, Daniel Khashabi, Qiang Ning, and D. Roth. 2019. "going on a vacation" takes longer than "going for a walk": A study of temporal commonsense understanding. *ArXiv*, abs/1909.03065.

A Appendix

A.1 Varying LMs (Table 3)

Pre-training significantly improves R@1(3.5% for GPT-2, 5.5% for BERT) over randomly initialized models, thus validating the need for pre-training. All the pre-training approaches tend to perform similarly with RoBERTa performing the best.

A.2 Human Baselines

We ask annotators on Mechanical Turk platform to do the visual story cloze task (Figure 3), i.e pick one of the 5 candidate images given 4 context images, on 200 samples from the test split. All annotators are highly rated and belong to United States. We get 3 annotations for each sample and measure annotator agreement by calculating the number of times 2 or more of the 3 annotators made the same prediction. We find that the annotators agree 77% of the time. For action forecasting, we cite the human study from Vondrick et al. (2016).

A.3 Qualitative samples (Figure 4)

Row 1 shows a family having an outdoor barbeque party. The first three images show the family playing with a child and the last image shows an old man barbequing. While the vision models predict visually similar but semantically unrelated candidates, our model correctly captures the correlation between "a plate of cooked chicken" and "a man on the grill". Now, consider row 2 which depicts a father-son duo watching a baseball game. While our model predicts the wrong candidate, the corresponding caption "two young boys playing baseball" is a likely event in post-game celebrations.

A.4 Implementation details

We use a batch size of 16. We use a maximum learning rate of 2e-5 and decay to 1e-5 over the length of training and optimize with Adam (Kingma and Ba, 2015). We set the maximum length of the generated caption to 20. We train the visual story cloze experiments for 10-20 epochs and the action forecasting experiment for 60 epochs. All are models are implemented in PyTorch (Paszke et al., 2019) and we use the Hugging Face transformers library (Wolf et al., 2020) for all pre-trained LMs.

Model	Validation R@1 ↑	Validation R@3 ↑	Test R@1 ↑	Test R@3 ↑
RoBERTa	29.66	68.54	**28.39**	**69.14**
BERT	**29.96**	**70.04**	27.30	68.74
GPT-2	29.86	69.04	28.29	68.25
Random init. BERT	26.95	66.03	21.66	64.00
Random init. GPT-2	27.56	67.64	24.83	64.69

Table 3: Performance of different pretrained models on the future image prediction task, with captions generated from Oscar(finetuned).

A.5 Reproducing Image GAIL Model

We recreate the model in Zeng et al. (2017)(Figure 5) for benchmarking and fine-tune components that were not concretely described in the original paper. The overall model architecture uses ResNet-101 as the network ϕ, an autoencoder as the policy network π, and a discriminator, the latter two of which are described in the supplemental material for (Zeng et al., 2017). During training, we use the Adam optimizer and 10^{-4} as the initial learning rate for all three models, and decay the learning rate by a factor of 0.1 after every 20 epochs. We also set the batch size to be 16 and use a dropout rate of 0.5 across all dropout layers. Additionally, we freeze the weights of ResNet-101 for the first 5 epochs, and unfreeze them afterwards until the end of training. To calculate rewards for the policy network, we set a discount rate 0.99.

During training, a batch of sequences of 5 temporally-ordered images are fed into ϕ to produce a batch of sequences of 5 temporally ordered 2048-dimensional vectors. We then take the first vector h_1 of each sequence in the batch and pass them through the policy network π to produce a corresponding prediction, a_2, and feed these into a normal distribution with fixed variance σ^2 to produce the predicted state h_2'. We then repeat this process to produce h_3' from h_2', h_4' from h_3', and h_5' from h_4'. We treat $[h_t, h_{t+1}]$ as the ground-truth state-action pair, and $[h_t, h_{t+1}']$ as the policy prediction state-action pair.
During discriminator updates, we compute the discriminator loss with binary cross-entropy on the expert trajectory state-action pairs and policy trajectory state action pairs, then taking the mean loss across the batch for gradient computation. During policy updates, we employ the Monte Carlo search described in the supplemental material for (Zeng et al., 2017), where we compute the expected return $Q(h, h_{t+1})$ as the sum of all discriminator outputs on the trajectory of states from the policy output. Finally, we compute the policy gradient loss as the

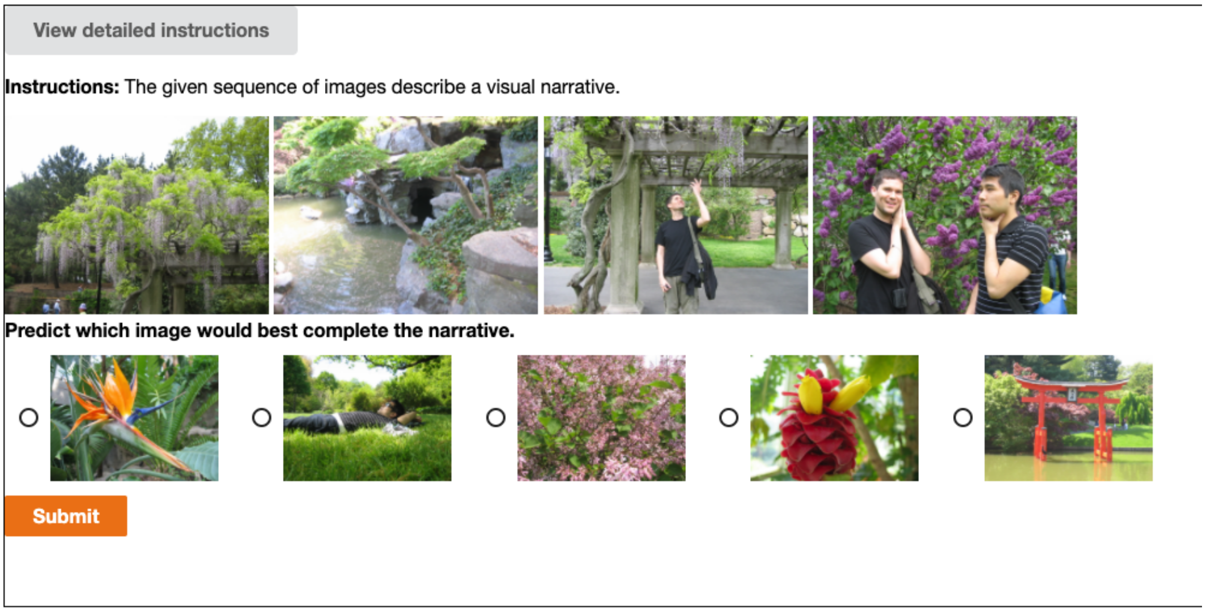

Figure 3: Task interface used to get annotations for the human baseline for the visual cloze task on the Visual Storytelling dataset (Mostafazadeh et al., 2016). Workers are shown the 4 context images and are asked to determine which image, among 5 candidates, best completes the narrative defined by the context images.

Figure 4: Comparing predictions on samples from the test split across different variants (GAIL in dashed purple, NN in dashed red, our Oscar(finetuned) + RoBERTa model in green) with captions generated from Oscar(finetuned). Best viewed in color.

sum of the negative product of the log probability and the expected reward $Q(h, h_{t+1})$ across the state trajectory.

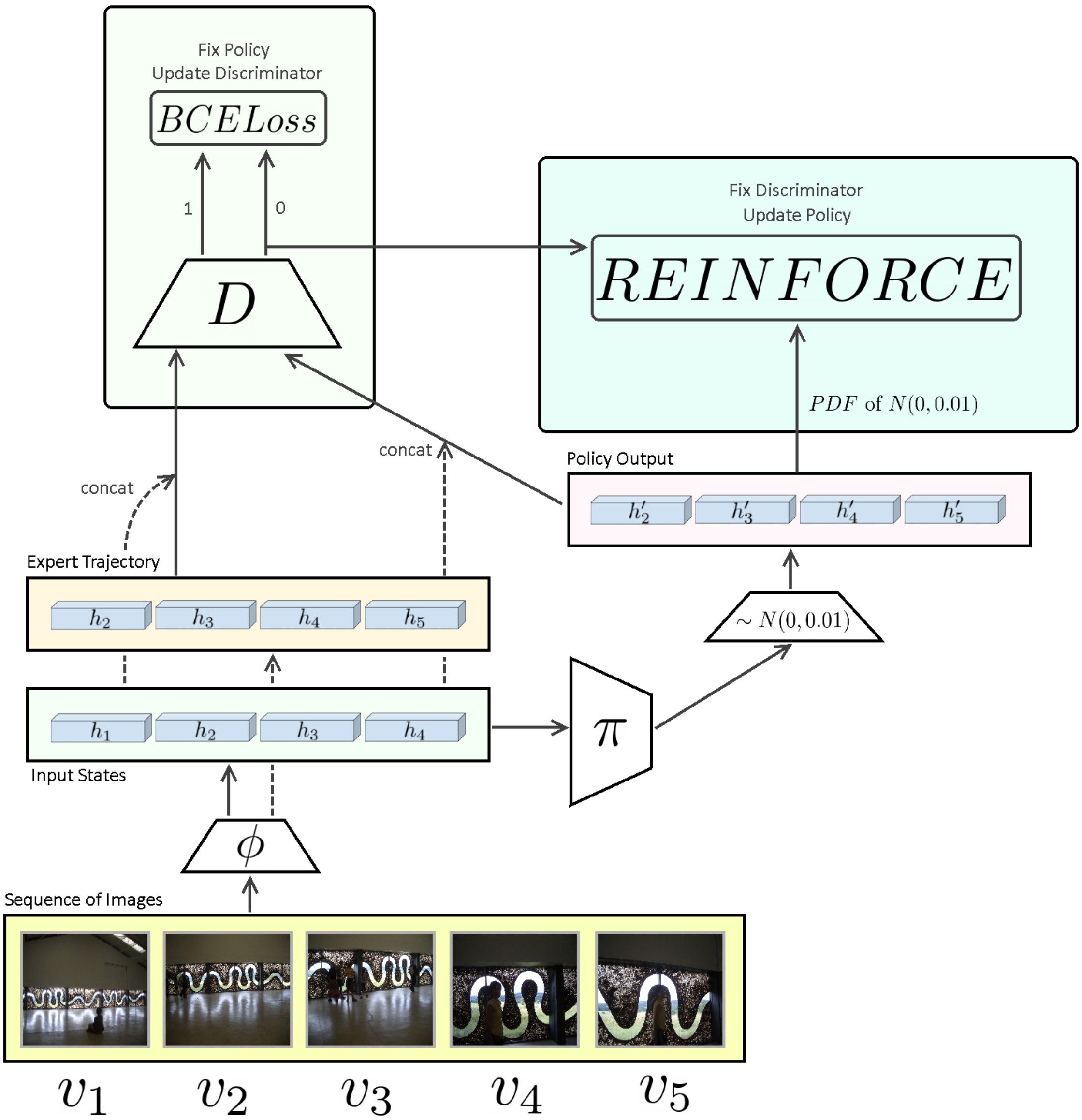

Figure 5: Shows training loop for image-based GAIL model. Given a sequence of 5 images, the model transforms them into 2048-dimensional vectors and splits them such that the vectors representing the first 4 images represent the input states, while the last 4 images represent the expert trajectory. These two sequences are then used to compute both the discriminator and policy loss.