Do Transformers Understand Time?

Shawn Jain
Microsoft Research
{jains,

Hamid Palangi
Microsoft Research
hpalangi,

Yonatan Bisk
Microsoft Research
yobisk,

Jianfeng Gao
Microsoft Research
jfgao}
@microsoft.com

Abstract

Transformers-based models have repeatedly created new state-of-the-art results in NLP and Vision-Language tasks like language modeling, general question answering, visual question answering, information retrieval, and image captioning (Devlin et al., 2018; Lu et al., 2019; Chen et al., 2019). However, such models have not been evaluated on their temporal reasoning abilities. Temporal reasoning is essential; it is easy for humans, yet hard for AI systems.

We attempt to define a metric for temporal reasoning, and evaluate the ability of Transformers models to do temporal reasoning. We propose the Strided Sentence Order Prediction (SOP) as an evaluation metric for temporal reasoning. We evaluate on modified versions of the ActivityNet-Captions (Krishna et al., 2017) and HowTo100M-Captions (Miech et al., 2019) datasets. We find that this metric is flawed, as the task of ordering two grammatical sentences provides insufficient temporal context for humans to solve the problem. This finding breaks our intuitive assumptions.

We release our our codebase: models, experiments, and data preparation pipelines. [1]

1 Introduction

Transformers-based models have repeatedly created new state-of-the-art results in NLP and Vision-Language tasks like language modeling, general question answering, visual question answering, information retrieval, and image captioning (Devlin et al., 2018; Lu et al., 2019; Chen et al., 2019). However, these models have shortcomings. For example, in text generation, they drift off topic (Liu et al., 2019). In vision-language problems, they don't effectively ground language to image regions (Yang et al., 2019). Such models have not been evaluated on their temporal reasoning abilities. Temporal reasoning is essential; it is easy for humans, yet hard for AI systems (Figure 1).

Temporal reasoning is a subset of commonsense reasoning. Commonsense reasoning is the broad field of knowledge, skills, and reasoning abilities that all humans use in everyday situations. Commonsense reasoning also includes spatial, spatial-temporal, and intent reasoning, among other forms of reasoning and knowledge (Davis and Marcus, 2015). Commonsense reasoning is a pre-requisite for AI systems to perform everyday human tasks. While these state-of-the-art Transformers models have been indirectly evaluated for certain forms of commonsense reasoning, like Visual Commonsense Reasoning (Zellers et al., 2018a) and SWAG (Zellers et al., 2018b), they have never been directly evaluated for temporal reasoning.

In this work, we attempt to define a metric for temporal reasoning, and evaluate the ability of Transformers models to do temporal reasoning. We propose the Strided Sentence Order Prediction (SOP) as an evaluation metric for temporal reasoning. This is adapted from the pretraining literature (Lan et al., 2019). We evaluate on modified versions of the ActivityNet-Captions (Krishna et al., 2017) and HowTo100M-Captions (Miech et al., 2019) datasets. We find that this metric is flawed, as the task of ordering two grammatical sentences provides insufficient temporal context for humans to solve the problem. Furthermore, the models actually outperform the human baselines, suggesting the models are learning some other pattern in the data and providing a degenerate solution.

2 Methods and Experimental Setup

We propose the Strided Sentence Order Prediction (SOP) as a metric for temporal reasoning. This is

[1] https://github.com/darkmatter08/VLP_preview

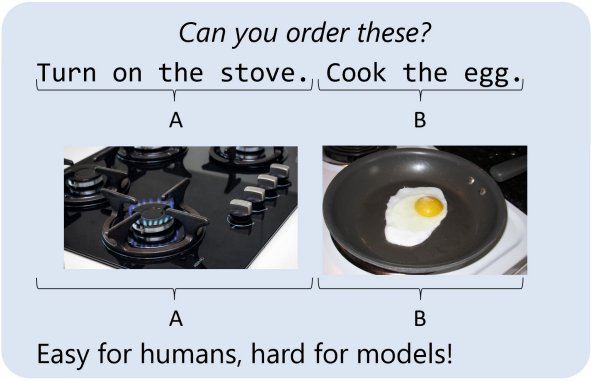

Figure 1: Ordering sentences and images is easy for humans but difficult for AI systems.

adapted from the pretraining literature (Lan et al., 2019). In Strided SOP, sentence pairs are formed from the document. When stride is set to n, n+2 consecutive sentences are selected to form a span. From the span, the first and last sentences are picked to form the sentence pair. If n = 0, the sentences are immediately consecutive sentences. Figure 2 illustrates this process. Note that these are grammatical sentences from the source document. These "strided datasets" are generated offline from the source dataset, before training or testing the model. We generated all strides in the range 0 to 3 (inclusive).

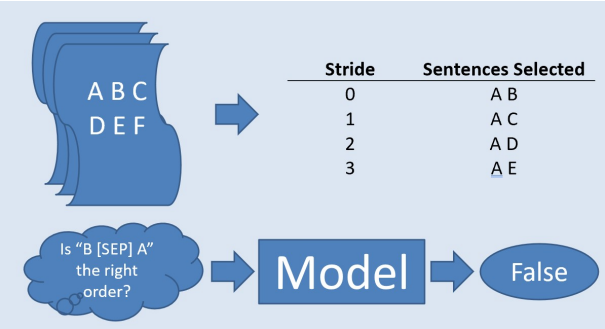

Figure 2: A visualization of the Strided Sentence Order Prediction (SOP) task.

For each strided dataset, instantiate and load the pretrained checkpoint for the BERT-base model. Train and evaluate each model using Strided SOP as the training objective. Note the Masked Language Model (MLM) objective is not used in this experiment. Train for 10 epochs, monitoring the train and eval loss curves, using early stopping. We used a learning rate of 2e-6, and a batch size of 64. We report the peak evaluation accuracy. This procedure was run on two datasets: ActivityNet-Captions (Krishna et al., 2017) and HowTo100M-Captions (Miech et al., 2019). These datasets were selected because both daily activity videos and HowTo/Tutorial-style videos have a clear temporal relationship; ordering of the steps is inherent in this style of content. ActivityNet-Captions consists of 20,000 videos of common daily human activities with 100,000 human generated caption text. HowTo100M is a dataset of 1.2 million YouTube How-To/Tutorial-style videos with ASR-generated caption text. From both datasets, we use only the caption data. We pre-process both datasets given their noisy sources; We discard any videos with fewer than five sentences and discard the first and last sentence per video. Besides BERT-base, we also experimented with BERT-Large, RoBERTa, and ALBERT. However, we do not report these results. We also manually evaluated 20 random samples of data from each strided dataset; this was our human baseline score.

3 Results and Analysis

On ActivityNet-Captions, model performance increases with stride, contradicting our hypothesis (Figure 3 left). On HowTo100M-Captions, model performance decreases with stride, matching our hypothesis (Figure 3 right). Therefore, the SOP – Stride relationship is dataset dependent. On both datasets, model performance beats human baseline performance. To understand these results, an error analysis was conducted. We provide a visualization of our error analysis in Table 1. The leftmost "Error Type" column indicates our error categorization based on our analysis of the example. Notice that many of the examples are difficult to label correctly even for humans. HowTo100M-Captions is particularly difficult for humans, because of the noisy ASR-generated captions.

4 Conclusion

Because the model is consistently outperforming humans, and because the SOP – Stride relationship is dataset dependent, we conclude that the model is not learning the true underlying task. Instead it learns some other underlying patterns in the data and provides a degenerate solution. We conclude that defining the appropriate temporal context for sentence ordering is difficult. This breaks our previous intuitive assumption – that the two sentences provide enough "temporal context" to solve the strided SOP task. This is true for both datasets used. Therefore, we are unable to con-

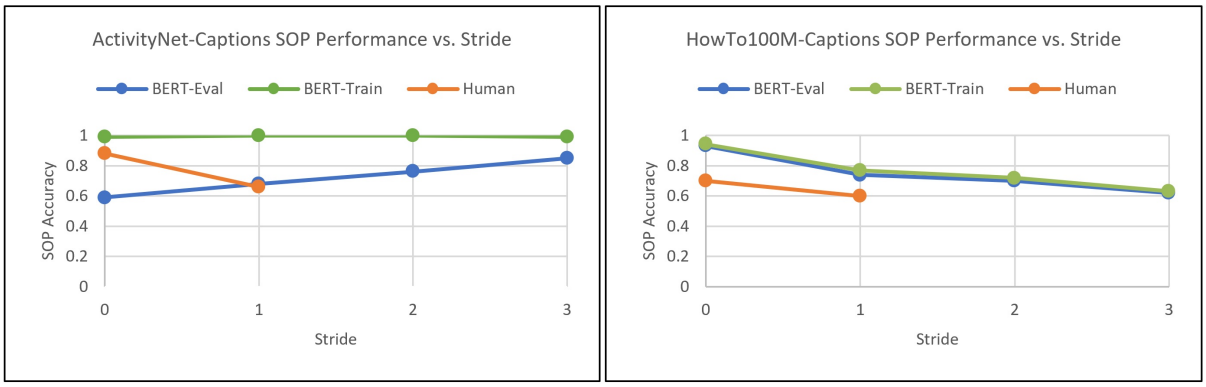

Figure 3: Strided SOP Performance vs. Stride for the ActivityNet-Captions (left) and HowTo100M-Captions (right) datasets.

clude anything about the original hypothesis. This also suggests that these vision-language datasets do not provide enough text-only temporal context. Our error analysis (Table 1) suggests that visual grounding would disambiguate many examples.

4.1 Future Work

Future investigations should focus on three directions. First, tackling the problem of defining temporal context. Second, developing stride-invariant models. Third, generalizability and transferability of learned temporal reasoning to new datasets and domains. The definition of temporal context needs to make the task easy for humans but difficult for existing models, in addition to being semantically meaningful. It needs to serve as a strong proxy for the motivating task of temporal reasoning, as in Figure 1. In this direction, we propose several ideas. First, redefine sentences to include n-seconds of captions instead of grammatical sentences. Longer sentences provide additional context to order; we noticed this when manually labelling examples for the human baseline. Second, order three (or more) sentences instead of two sentences. This is redefining the strided SOP training objective. Ordering three sentences both provides additional context and requires stronger temporal reasoning abilities. More bits of supervision provides the model a stronger signal to learn from. Additionally, this more complex task would reduce the likelihood of overfitting on small datasets like ActivityNet-Captions. Third, investigate alternate types of datasets. The text-only portion of vision-language datasets often need visual grounding to have enough context to solve the strided SOP task. Existing and widely used datasets like HellaSwag could fill this role. Any new definition would have to show that the model is not exceeding human performance. With the appropriate definition of temporal context in place and existing Transformers-type models benchmarked against this metric, future investigations could tackle the idea of developing a stride invariant model. The ideal stride-invariant model would maintain a consistently high level of strided SOP performance across all strides. Such a model would have 'solved' temporal reasoning by this metric. Towards this direction, models could infer the stride of the output data, in addition to inferring the ordering relationship. This would offer more supervision to the model, and the model would have to maintain some notion of 'temporal distance.' In this framework, a model could be trained with multiple strided datasets. The model would then be exposed to reasoning about multiple strides of data, enhancing its temporal reasoning ability. This appeals to the fact that humans do temporal reasoning over multiple time scales naturally. Finally, Transformers-style models have gained broad popularity because after pretraining, they can be easily adapted to a variety of downstream tasks (Chen, et al. 2019). Towards making the learned temporal reasoning transferable, future investigations could investigate evaluating on additional out-of-domain tasks. In this framework, this means treating the ActivityNet-Captions and HowTo100M-Captions as a 'second stage pre-training,' and finetuning the models on the target downstream tasks. HellaSwag could be one such downstream task. The Masked Language Model (MLM) objective could be used as a second finetuning objective to allow for domain adaptation. More broadly, the techniques shown here could be used in the vision-language domain, per-

ActivityNet-Captions		
Error Type	**Example**	**Stride**
Success	The man drops the weight to the ground. The man lifts the weight over his head.	0
Need Visual Grounding	The director of the race gives an interview as people pass behind him. The people are seen running the marathon.	0
Unrelated Concepts	A man walks holding a paper and pen. A gym ##nast stands on his arms.	3
GT doesn't make sense	He is scrap ##ping off the excess wax from the ski. He uses the buffer on the ski again.	0
HowTo100M-Captions		
Error Type	**Example**	**Stride**
Success	mom was told me you should cut the. like a couple of inches I don't know why	1
Success - Repeated words	straight in in kinda even with you know enjoys the j ##ois ##t sometimes people get. enjoys the j ##ois ##t sometimes people get	0
Need Visual Grounding	brown that's what we want them they're. not gonna get as dark as your regular	1
Unrelated Concepts	We're gonna mix that together make sure that it's not c ##lump ##y you've actually. some ha ##m and cheese	0
GT doesn't make sense	mystical powers so as you can see. the power of magic	1

Table 1: Visualized successful and erroneous examples. The leftmost "Error Type" column indicates our error categorization based on our analysis of the example. All sentence pairs are shown in the correct order (ground truth).

haps by using the datasets' visual features and a Vision-Language model.

Acknowledgments

Yonatan Bisk and Hamid Palangi have been my steadfast mentors in this project. Microsoft Research, Susan Dumais, and Jianfeng Gao also enabled me to pursue this project. This work was performed as part of the 2019-2020 Microsoft Research AI Residency Program.

References

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. UNITER: Learning UNiversal Image-TExt Representations. *arXiv e-prints*, page arXiv:1909.11740.

Ernest Davis and Gary Marcus. 2015. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Commun. ACM*, 58(9):92–103.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv e-prints*, page arXiv:1810.04805.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *International Conference on Computer Vision (ICCV)*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *arXiv e-prints*, page arXiv:1909.11942.

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 e-prints*, page arXiv:1907.11692.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. *arXiv e-prints*, page arXiv:1908.02265.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *ICCV*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le.

2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *arXiv e-prints*, page arXiv:1906.08237.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2018a. From Recognition to Cognition: Visual Commonsense Reasoning. *arXiv e-prints*, page arXiv:1811.10830.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018b. SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference. *arXiv e-prints*, page arXiv:1808.05326.