An Improved Baseline for TRIP with Attention and New Backbones

Yicheng Tao* **Leheng Lu***
Computer Science and Engineering Division, University of Michigan
{yctao, leheng}@umich.edu

Abstract

Language understanding models have significant progress recently. They achieve human-level performance on state-of-the-art benchmarks for commonsense inference. However, recent studies show that these models have the sign of overfitting because they are evaluated based on end task performance. The experiment results indicate that they perform poorly on learning the underlying reasoning process and use it to support their predictions. In order to address this problem, Storks et al. (2021) proposed a baseline tiered system to perform a physical commonsense reasoning task on their TRIP (Storks et al., 2021) dataset. In this work, we aim at improving the performance of the baseline system by adjusting the conflicting detector and the story choice prediction module. Our code is available at https://github.com/yctao7/EECS-595-Final-Project.

1 Introduction

Modern state-of-the-art large-scale language models (LMs), such as BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), and DeBERTa (He et al., 2020), have reached human-level performance on language understanding tasks. The recent advances in NLP models are definitely exciting, however, there exists bias in language benchmarks that leads to high performance in these language models (Bender and Koller, 2020). Although these language models have impressing results on end task performance predicting class labels on different context, studies show that they are in fact not good at understanding the reasoning and meaning in the context (Niven and Kao, 2019).

In order to address this problem, Storks et al. (2021) introduced TRIP, a benchmark targeting physical commonsense reasoning along with a baseline tiered system. We want to evaluate language models not only on classification of class labels, but also on performance of commonsense reasoning in order to make sure that the models are learning based on context reasoning, not the correlation between context and class labels. In this work, we aim at improving the baseline tiered reasoning system to solve commonsense reasoning tasks, and it will be evaluated on the TRIP benchmark. We propose two approaches–attention-based story classification and gated state representation–that can be easily integrated with the baseline system and result in improvements in several metrics.

1.1 Problem Statement

TRIP is a physical commonsense reasoning dataset coupled with reasoning evidence. It is composed of human-written stories describing sequences of physical actions with plausible sentences, where each story can be paired with another story so that two stories only differ by one sentence and one story is much more plausible than the other. Given such a pair, the end task of TRIP is to find out the plausible story. To reduce subjectivity of the stories, TRIP only includes concrete actions and sentences in a simple declarative form. To further explore plausibility in longer context than previous benchmarks (Roemmele et al., 2011; Zellers et al., 2018; Bisk et al., 2020), TRIP also requires that each story must have at least five sentences. Three levels of labels are provided with the stories. The first level label indicates the plausible stories, the second level label identifies a conflicting sentence pair for each implausible story, and the third level label gives physical states of entities before and after actions (i.e., precondition states and effect states) in each sentence of all the stories. An example sample in TRIP is shown in 1. The method intended to solve TRIP should predict these three levels of labels simultaneously to show its reasoning process, i.e., not only make story plausibility classification but also find the conflicting sentence pair that leads to the implausibility and the entity's

*Equal contribution.

physical states that lead to the conflict. Overall, the dataset contains 675 plausible stories and 1472 implausible stories. Despite its small size, it is perfect for testing models' reasoning accountability toward the end task. Following Storks et al. (2021), we use accuracy, consistency, and verifiability as metrics for evaluation, which are described below.

Accuracy. The proportion of testing examples where the plausible story is correctly identified.

Consistency. The proportion of testing examples where the plausible story and the conflicting sentence pair for the implausible story are correctly identified.

Verifiability. The proportion of testing examples where the plausible story, the conflicting sentence pair for the implausible story, and the entity's physical states that contribute to the conflict are correctly identified.

2 Related Work

Storks et al. (2021) proposed a baseline tiered system for solving TRIP. Given a pair of stories differing by one sentence which makes one of them implausible, the system first takes as input an entity-sentence pair, uses a pre-trained Transformer to get the contextual embedding of the entity, and feeds the embedding into two network-based classifiers to predict the precondition and effect states of the entity. Next, for each entity, the system concatenates its contextual embedding, precondition classification logits, and effect classification logits for each sentence, and uses a pre-trained Transformer-based conflict detector followed by linear projection and sigmoid to predict the probability of each sentence conflicting with another in each story. Lastly, given the predicted conflict probability for each entity-sentence pair, the system makes story choice classification by summing up the negative probabilities within each story and applying softmax to the sums to predict the plausibility probability of each story. In summary, the overall loss can be written as

$$\mathcal{L} = \lambda_p \mathcal{L}_p + \lambda_f \mathcal{L}_f + \lambda_c \mathcal{L}_c + \lambda_s \mathcal{L}_s,$$

where λ_p is the precondition classification loss, λ_f is the effect classification loss, λ_c is the conflict detection loss, λ_s is the story choice loss, and λ_p, λ_f, λ_c, λ_s are balancing weights summing to 1. The details of the system are provided in Figure 2 (Storks et al., 2021).

3 Approaches

In this section, we present our approaches that improve the baseline system for TRIP in the end task accuracy as well as the reasoning interpretability.

3.1 Attention-based Story Classification

The way that the baseline system makes story plausibility prediction forces the non-conflicting sentences in the implausible story to have a larger belief probability of conflicting with another sentence. This could offset the supervision given by the ground-truth conflicting sentences and thus harm system's performance on consistency. Plus, since the story choice classification of the baseline system directly depends on the result of the conflict detection, the error from the latter could be inherited by the former, which could lead to a drop in accuracy. To address these problems, we let each story attend to its sentences most likely to contribute to a conflict and use a learning-based story choice classifier. Specifically, we first apply softmax to conflict probability logits of sentences in each story and use the results as attention weights to sum up the sentence representations to form the story representation. An MLP followed by softmax is then employed to predict the story plausibility probability. Formally, it can be written as

$$T_j = \sum_{i=1}^{n} \mathrm{softmax}_i(g_{ij}) S_{ij},$$

$$p_j = \mathrm{softmax}_j(\mathrm{MLP}(T_j)), \quad j = 1, 2,$$

where T_j is the representation of story j, $n = \#\text{entities} \times \#\text{sentences}$ in story j, g_{ij} is the conflict probability logit of sentence i in story j, S_{ij} is the representation of sentence i in story j given by the conflict detector, and p_j is the plausibility probability of story j.

3.2 Gated State Representation

We notice that for each entity TRIP uses 20 categorical attributes to depict its precondition and effect states. These attributes could be classified into three distinct categories in terms of their relevance with the entity: (1) relevant in context; (2) irrelevant in context but relevant in semantics; (3) irrelevant in semantics. For attributes in the third category, TRIP has provided a special "irrelevant" value for each attribute to handle them. So, the left problem is how can we tell the model to focus on attributes in the first category instead of those

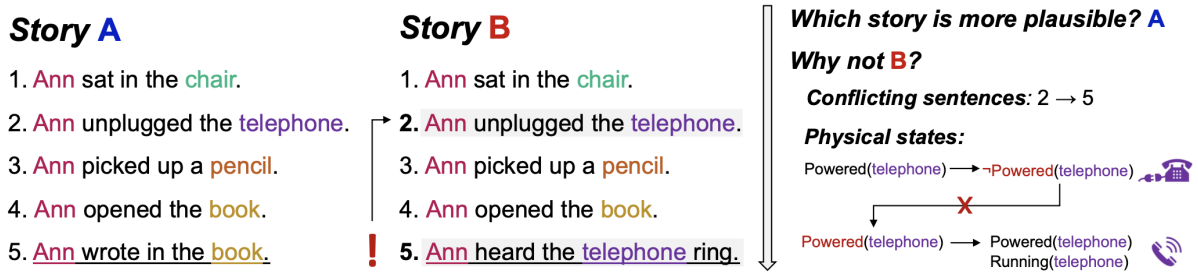

Figure 1: A sample in TRIP.

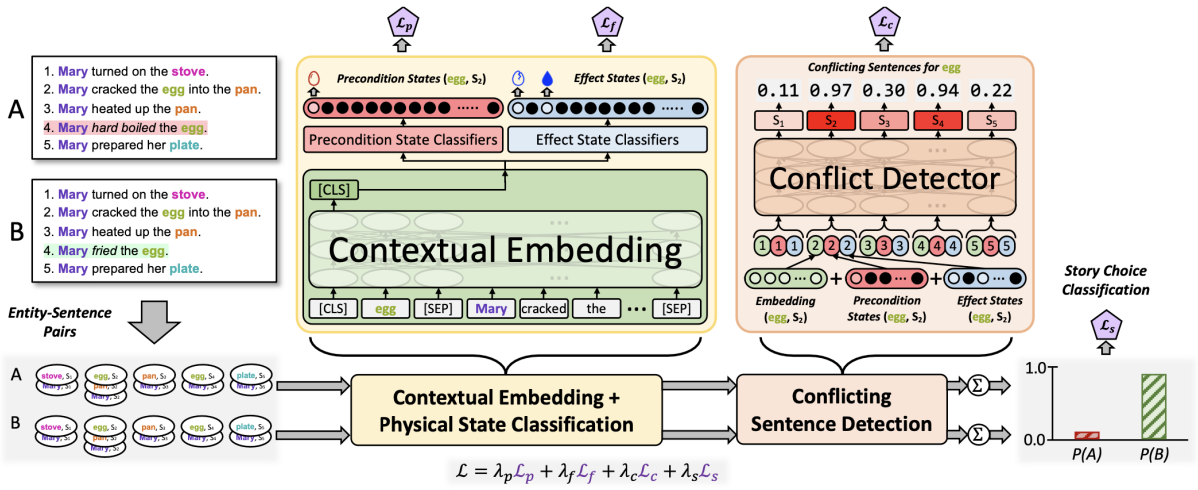

Figure 2: Model architecture of Storks et al. (2021).

in the second category. We address it by employing learned gates to select attributes. Specifically, we first average the entity's contextual embeddings in all sentences of the story and feed the result to an MLP followed by sigmoid to build a "binary" gate for each attribute. Then we apply the gates to the precondition and effect classification logits of the entity to obtain the gated state representations, which are the inputs to the conflict detector. Formally, it can be written as

$$T = \text{sigmoid}(\text{MLP}(\frac{1}{n}\sum_{i=1}^{n} e_i)) \odot S,$$

where T is the gated state representation, $n = \#\text{sentences}$ in the story, e_i is the contextual embedding of the entity in sentence i, $\odot$ is the Hadamard product, and S is the precondition or effect classification logits of the entity.

3.3 Positional Encoding in Transformers

Recurrent neural network (RNN) reads input data sequentially to include inductive bias, while Transformer-based models process data in parallel so they are less sensitive to position. The order of physical states in the tiered reasoning system is important, so, the position encoding methods used in Transformers might be relevant. That being said, we want to inject the position information into attention calculation. Therefore, in this work, we investigate the performance of Transformers with different positional encoding methods by replacing them with the conflict detector module in the baseline system.

There are different ways to encode positions. The original Transformer uses Absolute Positional Encoding (Vaswani et al., 2017). There are also other approaches such as Relative Positional Encoding and its variants. For example, those approaches focus on absolute positional encoding: (Gehring et al., 2017), (Lan et al., 2020). There are also other work focus on relative positional encoding (Parikh et al., 2016), (Raffel et al., 2020), (Dai et al., 2019). In this work, we are using RoFormer (Su et al., 2021), an enhanced Transformer with rotary position encoding (the structure is shown in Figure 4), and Xlnet (Yang et al., 2020), a transformer equipped with relative positional encoding

approach. We use these transformers to replace the original transformer used in the conflict detector module. Then, we evaluate these models using the metrics describe in Section 1.1.

Roformer is an enhanced transformer with rotary positional encoding, which incorporates relative position information using rotation matrix product. Su et al. (2021). Xlnet is also an enhanced transformer that integrates the relative encoding scheme and the segment recurrence mechanism to improve performance (Yang et al., 2020).

In order to adapt those transformers we propose to use, we make use of the transformers library on Hugging Face, which contains pre-trained models for many different kinds of transformers. The input of those pre-trained models is the same as the input used in the conflict detector module in the baseline system. The output is going to be the output of the last hidden layer in the pre-trained transformer. Then, we apply our attention-based story classification and gated state representation methods to produce the final result. In the experiment, we will present the results after making those changes.

4 Evaluation

4.1 Attention and Gate

Following the same evaluation scheme as Storks et al. (2021), we first evaluate our approaches in Section 3.1 and Section 3.2 on TRIP by including all losses or omitting the story choice loss with BERT, RoBERTa, or DeBERTa as backbones. We adopt the original implementation of the baseline system and add our approaches to it. We also use the same batch size, learning rates, loss balancing weights, and model selection methods as Storks et al. (2021). The only difference in experimental settings is the number of training epochs, which we use 15 instead of 10 to allow more training time for the newly added networks. The MLPs used in our approaches are all 2-layer linear layers with ReLU activation.

The results on the validation set are shown in Table 1. Under the "all losses" scenario, using attention-based story classification individually gives a slight increase in highest accuracy from 78.3% to 81.7% while great improvements in highest consistency and verifiability by 19.3% and 6.9% respectively. Together with gated state representation, the approaches again improve the highest consistency a little bit by 1.9%. Under the "omitting the story choice loss" scenario, using gated state representation individually does not seem to work, leading to a huge decrease in accuracy and consistency. Comparing the best systems of Storks et al. (2021) (omitting the story choice loss) and the best systems with our approaches (with attention-based story classification and gated state representation and including all losses), our systems have an about 3% lead in accuracy, are competitive in consistency, and lag behind in verifiability by about 4%.

Model	Accuracy (%)	Consistency (%)	Verifiability (%)
random	47.8	11.3	0.0
Storks et al. (2021): *All Losses*			
BERT	78.3	2.8	0.0
RoBERTa	75.2	6.8	0.9
DeBERTa	74.8	2.2	0.0
Ours (w/ Attention): *All Losses*			
BERT	76.1	17.4	4.0
RoBERTa	**81.7**	26.1	7.8
DeBERTa	78.6	23.3	5.0
Ours (w/ Attention and Gate): *All Losses*			
BERT	77.6	**28.0**	5.6
RoBERTa	78.9	27.3	6.8
DeBERTa	78.3	23.0	4.0
Storks et al. (2021): *Omit Story Choice Loss*			
BERT	73.9	**28.0**	9.0
RoBERTa	73.6	22.4	**10.6**
DeBERTa	75.8	24.8	7.5
Ours (w/ Gate): *Omit Story Choice Loss*			
BERT	57.5	23.3	9.3
RoBERTa	57.8	18.0	5.3
DeBERTa	57.1	16.8	5.9

Table 1: Validation metrics of Storks et al. (2021)'s and our tiered systems trained on varied combinations of loss functions and model architectures. "Attention" and "Gate" are the approaches introduced in Section 3.1 and 3.2 respectively. Random baseline (averaged over 10 runs) makes tiered predictions at random.

The results on the test set are shown in Table 2. Compared to the validation results, we also experience slight drops in consistency and verifiability as Storks et al. (2021) but no loss (actually increase) in accuracy. The best results in accuracy and consistency are given by the RoBERTa system with attention-based story classification and gated state representation and including all losses (the best model architecture and loss configuration found by the validation), which are 7.7% and 2.3% higher than the best results from Storks et al. (2021). Besides the obvious gains given by attention-based story classification, we also see systemic small improvements in all metrics when it is coupled with gated state representation. These facts prove the

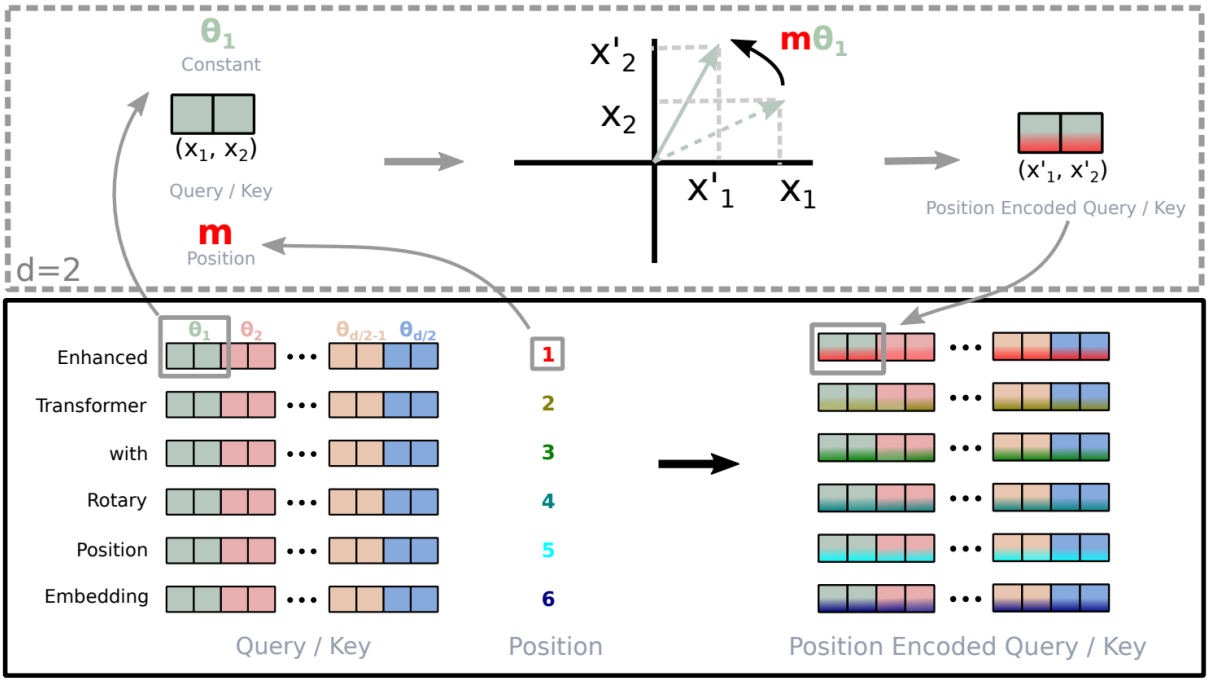

Figure 3: Rotary Position Embedding(RoPE) Su et al. (2021). It uses $\theta_i = 10000^{-2i/d}$. With this setting, there will be a long-term decay property which makes a pair of tokens with long relative distance to have less connection. This will replace the absolute positional encoding in the original Transformer.

supposed effectiveness of our approaches. On the other hand, our best system also narrows the gap in verifiability between two loss scenarios. It is competitive with the RoBERTa and DeBERTa system from Storks et al. (2021) omitting the story choice loss, though still 2.3% behind compared to their best result.

Model	Accuracy (%)	Consistency (%)	Verifiability (%)
random	49.5	10.7	0.0
Ours (w/ Attention): *All Losses*			
BERT	81.2	13.1	1.7
RoBERTa	81.2	23.6	4.6
DeBERTa	82.3	19.7	2.3
Ours (w/ Attention and Gate): *All Losses*			
BERT	82.1	17.9	2.3
RoBERTa	**82.9**	**24.5**	6.0
DeBERTa	81.5	20.8	2.3
Storks et al. (2021): *Omit Story Choice Loss*			
BERT	70.9	21.9	**8.3**
RoBERTa	75.2	18.8	5.7
DeBERTa	72.9	22.2	6.6
Ours (w/ Gate): *Omit Story Choice Loss*			
BERT	58.7	14.5	3.7
RoBERTa	54.1	13.7	4.0
DeBERTa	56.1	14.5	6.3

Table 2: Test metrics of Storks et al. (2021)'s best and our tiered systems.

Table 3 shows the validation macro-F1 scores of the baseline systems and our best systems on the tasks of precondition and effect classification, as well as conflicting sentence detection. We believe the results are reasonable. The losses in precondition and effect classification F1 scores may be due to the overfitting caused by our longer training time. The slightly lower conflicting sentence detection F1 scores are in the reasonable fluctuating range of large-scale network-based systems.

4.2 Positional Encoding

In our experiment on positional encoding, we build upon the approach proposed in Section 3.1, 3.2. Then, we replace the conflict detector module with a new pre-trained transformer model. Our first attempt is to use Roformer (Su et al., 2021), a rotary positional encoding transformer. We evaluate the new system on TRIP by including all losses powered by BERT, RoBERTa, and DeBERTa. We use the same batch size, learning rate, loss balancing weights, and number of training epochs. In the Roformer model, the size of the hidden layer we used is 768, and we use ReLU as the activation function.

The result on the test sets are shown in Table 4. The tests are run under the "All Losses" scenario. The accuracy scores are decreased to about 60 % using the Roformer or Xlnet in our conflict

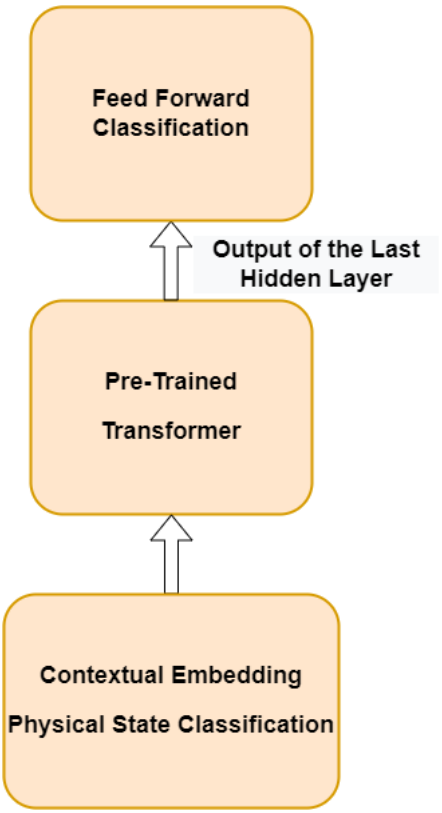

Figure 4: The new conflict detector. The pre-trained transformer will be replaced by Roformer or Xlnet.

detector module. However, the consistency and verifiability are both zero. It indicates that the result of replacing pre-trained models in the conflict detector module is not a good choice. More details will be dicussed below.

5 Discussion

The current evaluation results demonstrate that our approaches–attention-based story classification and gated state representation–are good complements to the baseline systems with all losses included. By adding these two approaches, the improved baseline systems largely increase their performances in accuracy, consistency, and verifiability, and become overall competitive with those omitting the story choice loss. In addition to the methodological designs described in Section 3, we believe these increases are credited to two reasons. First, keeping the story choice loss in the system gives necessary supervision that allows the inclusion of additional trainable networks. Second, removing irrelevant attributes for the entity according to the context alleviates the overfitting problem brought by the more complex model architectures and longer training time.

While the evaluation results on attention-based story classification and gated state representation approach look promising, the evaluation results on positional encoding method are not looking good. It looks like that the conflict detector module should be as simple as possible. With the original transformer, the input only needs to be passed into the encoder layer, followed by a feed-forward classification layer. There is nothing fancy here, but it can achieve a relatively more accurate result. Moreover, the positional encoding might not be relevant to the tiered reasoning system, but the structure of the transformer could be important since other transformers we have used have different structures from the original one. However, it is also possible that the implementation of those pre-trained transformers or our tiered reasoning system is not working as expected since both the consistency and verifiability are zero, and comparing to the random baseline, this is not normal. It is worthy to read the documentation and check the code more carefully, but due to the time limit, we are not able to go further on this. Lastly, the pre-trained transformers model might not be suitable in our task. We could re-train their models and even make some changes to fit in our tiered reasoning system. In which case, due to the limit in computational re-

Model	Prec. F1 (%)	Eff. F1 (%)	Confl. F1 (%)
Ours (w/ Attention): *All Losses*			
BERT	38.5	39.4	69.0
RoBERTa	40.4	41.5	66.5
DeBERTa	38.3	39.2	66.9
Ours (w/ Attention and Gate): *All Losses*			
BERT	38.3	39.0	68.8
RoBERTa	41.5	41.7	67.8
DeBERTa	38.2	38.7	65.9
Storks et al. (2021): *Omit Story Choice Loss*			
BERT	**54.9**	57.2	66.3
RoBERTa	51.2	51.2	**69.6**
DeBERTa	52.8	**57.3**	63.6

Table 3: Validation macro-F1 scores of Storks et al. (2021)'s and our best tiered systems on aggregate precondition, effect, and conflicting sentence classification. Scores are averaged over all attributes for state classification.

source, we won't be able to go further on this, but it could be an approach to be considered in future research.

Despite of the results we already get, we believe our work is not totally complete due to the limitation of computation resources and time. Future work includes: (1) tuning hyperparameters such as learning rates and MLP architectures in our approaches; (2) trying data augmentation techniques to provide more training data and alleviate overfitting; (3) performing downstream analysis for our results as Storks et al. (2021) did to better understand them; (4) designing approaches that can increase verifiability while maintaining our gains in accuracy and consistency; (5) trying Perceiver IO (Jaegle et al., 2021) as the backbone, which beats BERT on the GLUE benchmark (Wang et al., 2018).

6 Conclusion

We propose two approaches that improve the end task accuracy and reasoning consistency of the baseline tiered system proposed by Storks et al. (2021) for the TRIP dataset. For the positional encoding approach, there will be more work to be done in order to prove the feasibility of this approach.

Model	Accuracy (%)	Consistency (%)	Verifiability (%)
random	49.5	10.7	0.0
Ours (w/ Attention): *All Losses*			
BERT	81.2	13.1	1.7
RoBERTa	81.2	23.6	4.6
DeBERTa	82.3	19.7	2.3
Ours (w/ Attention and Gate): *All Losses*			
BERT	82.1	17.9	2.3
RoBERTa	**82.9**	**24.5**	6.0
DeBERTa	81.5	20.8	2.3
Ours (Roformer): *All Losses*			
BERT	61.0	0.0	0.0
RoBERTa	63.0	0.0	0.0
DeBERTa	61.8	0.0	0.0
Ours (Xlnet): *All Losses*			
BERT	64.4	0.0	0.0

Table 4: Test metrics of Storks et al. (2021)'s best and our tiered systems with positional encoding methods.

7 Division of Work

Yicheng Tao writes the Section 1.1, 2, 3.1, 3.2, 4.1, 5. Leheng Lu writes the Section 1, 3.3, 4.2, part of the discussion and Abstract.

References

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, et al. 2020. Experience grounds language. *arXiv preprint arXiv:2004.10151*.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. 2021. Perceiver io: A general architecture for structured inputs & outputs. *arXiv preprint arXiv:2107.14795*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations.

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*.

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments.

Ankur P. Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

Shane Storks, Qiaozi Gao, Yichi Zhang, and Joyce Chai. 2021. Tiered reasoning for intuitive physics: Toward verifiable commonsense language understanding. *arXiv preprint arXiv:2109.04947*.

Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. 2021. Roformer: Enhanced transformer with rotary position embedding.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2020. Xlnet: Generalized autoregressive pretraining for language understanding.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. *arXiv preprint arXiv:1808.05326*.