Advancing Transformers' Capabilities in Commonsense Reasoning

Yu Zhou, Yunqiu Han, Hanyu Zhou, Yulun Wu
University of California, Los Angeles
{yu.zhou, yunqiu21, alexiazhou726, gloriawuyl}@ucla.edu

Abstract

Recent advances in general purpose pre-trained language models have shown great potential in commonsense reasoning. However, current works still perform poorly on standard commonsense reasoning benchmarks including the Com2Sense Dataset (Singh et al., 2021). We argue that this is due to a disconnect with current cutting-edge machine learning methods. In this work, we aim to bridge the gap by introducing current ML-based methods to improve general purpose pre-trained language models in the task of commonsense reasoning. Specifically, we experiment with and systematically evaluate methods including knowledge transfer, model ensemble, and introducing an additional pairwise contrastive objective. Our best model outperforms the strongest previous works by $\sim$ 15% absolute gains in Pairwise Accuracy and $\sim$ 8.7% absolute gains in Standard Accuracy. [1]

Figure 1: Example data pairs from the Com2Sense test set along with model predictions. Here SOTA refers to (Jung et al., 2022) and OURS refers to Model 9 in Table 1.

1 Introduction

Endowing NLP models with human-like commonsense knowledge has remained a challenge for decades (Sap et al., 2020). In 2021, researchers proposed Com2Sense (Singh et al., 2021), a reliable and comprehensive commonsense reasoning benchmark with strict pairwise accuracy metrics. It consists of natural language sentence pairs labeled True/False based on whether they adhere to intuitive commonsense knowledge (Fig.1). The central evaluation criteria: Pairwise Accuracy, required that the model predict correctly for both sequences to count as successful.

Initial works on the dataset revealed that neither general purpose language models (Devlin et al., 2019), (Liu et al., 2019), (Raffel et al., 2019), etc. nor dedicated commonsense understanding models (Khashabi et al., 2020b), (Khashabi et al., 2020a) performed well on the dataset. All current models suffer from significant performance drops from Standard Accuracy to Pairwise Accuracy, displaying a huge discrepancy from human-like behaviour.

In this work, we examine possible methods of improving the performance of general purpose language models on the task of commonsense reasoning. Specifically, we study the effects of:

- Knowledge Transfer from relevant datasets containing commonsense knowledge, including the SemEval-2020 Dataset[2] and the SQuAD2 QA Dataset (Rajpurkar et al., 2018)
- Introducing a pairwise contrastive loss objective that forces models to distinguish commonsensical and non-commensensical statements
- Ensemble general purpose language models of different backbone architectures to study and compare their effects on overall performance.

[1]Our code and data are publicly available for research purposes at https://github.com/bryanzhou008/Improving-Commonsense-Reasoning/

[2]https://alt.qcri.org/semeval2020/

2 Methods

2.1 Knowledge Transfer

SemEval-2020 Dataset Similar to Com2Sense, Wang et al. (2020) provided a commonsense-related dataset, SemEval-2020 Task 4, Commonsense Validation and Explanation(ComVE). Each instance included a pair of sentences, one of which makes sense while the other does not.

We hypothesize that using a language model pretrained on the SemEval dataset, we are able to achieve better performance than finetuning that model directly on Com2Sense. We trained the DeBERTaV3$_{\text{large}}$ model on SemEval dataset to get a checkpoint model. The parameters used for pretraining are: *batch size = 48*, *lr = 4e-5*, *weight decay = 0.01*, *adam eps = 1e*-6, trained for *100 steps*. We then finetuned the obtained model on the Com2Sense dataset with the same parameters as our best-performing model (Table 2, Line 10) and compare the two models.

SQuAD2 QA Dataset We hypothesize that the language model will achieve higher performance after pretraining on question-answering datasets. We compared the results from RoBERTa$_{\text{base}}$ and DeBERTaV3$_{\text{large}}$ with those from RoBERTa$_{\text{base}}$-SQuAD2 and DeBERTaV3$_{\text{large}}$-SQuAD2, respectively, by finetuning them on Com2Sense with the same parameters. We used the parameters in line 3 of Table 2 for RoBERTa$_{\text{base}}$ and parameters of our best model for DeBERTaV3$_{\text{large}}$.

2.2 Pairwise Contrastive Loss

The Com2Sense dataset is complementary in nature. That is, for each statement, there is a complementary statement constructed with small perturbation on certain words making it concerning similar common sense concepts but with different (opposite) labels. This unique setting makes Com2Sense an ideal case for the use of contrastive learning.

We hypothesize that for the model to capture the semantic difference between commonsensical inputs vs their syntactically similar counterparts, it would be beneficial if we can push apart the hidden representation of each complementary input pair in the embedding space.

In practice, inspired by the InfoNCE Contrastive Loss by van den Oord et al. (2018), we propose a Pairwise Contrastive Loss (PCL) function:

$$\mathscr{P}^{\mathrm{x_i,x_j}}(\mathrm{W}) = \frac{\mathrm{e}^{\mathrm{sim}\left(\mathrm{g}(\mathrm{x_i}),\mathrm{g}(\mathrm{x_j})\right)/\tau}}{\sum_{\mathrm{k=1,k\neq i}}^{2\,\mathrm{N}} \mathrm{e}^{\mathrm{sim}(\mathrm{g}(\mathrm{x_i}),\mathrm{g}(\mathrm{x_k}))/\tau}} \quad (1)$$

Here for each complementary input sample pair $(\mathrm{x_i}, \mathrm{x_j})$ with embedding vectors $\mathrm{g}(\mathrm{x_i}), \mathrm{g}(\mathrm{x_j})$, where $\mathrm{sim}\left(\mathrm{g}\left(\mathrm{x_i}\right), \mathrm{g}\left(\mathrm{x_j}\right)\right)$ is the dot product of the L2 normalised inputs and τ is the constant temperature parameter which we set to 0.5.

The total contrastive loss, $\mathcal{L}$, is defined as the arithmetic mean over all pairs in the batch of the cross entropy of their normalised similarities, i.e.

$$\mathcal{L}_{\text{total}} = -\frac{1}{\mathrm{N}} \sum_{\mathrm{j=1}}^{\mathrm{N}} \log \mathscr{P}^{\mathrm{x_j,x_j}}(\mathrm{W}) \quad (2)$$

2.3 Model Ensemble and Rule-Based Perturbation

Due to the complementary nature of the Com2Sense dataset, each input data pair should have one positive sample and one negative sample. With this fact in mind, we propose a posterior model ensemble pipeline that aims to reduce the number of Same-Output Pairs where two prediction labels are the same (either both positive or both negative). This method further helps our ensembled model distinguish between syntactically similar sentence pairs that represent different ideas.

In practice, we take n finetuned models and rank them by their pairwise accuracy score on the dev set to represent our confidence in each model. Then we use the highest performing model as a base predictor to generate predictions on the test set, which would contain a number of Same-Output Pairs. For each Same-Output Pair, we move down the list of ranked models by confidence and generate their predictions. If the new model can differentiate the two samples (generating one positive and one negative), we then adopt the new model's prediction. In the end, we have an (ideally very small) number of test pairs that none of the models is able to differentiate between, in which case we randomly assign different prediction values to the pair.

3 Results

3.1 Different Model Backbones

To find the best model backbone architecture, we compare the results of BERT$_{\text{base}}$, RoBERTa$_{\text{base}}$, DeBERTa$_{\text{base}}$, and DeBERTaV3$_{\text{base}}$ with the best

Line No.	Model	Pairwise Acc %	Standard Acc %
1	UnifiedQA-3B (Khashabi et al., 2020b)	51.26	71.31
2	Maieutic Prompting (Jung et al., 2022)	68.70	75.00
3	DeBERTaV3$_{large}$ (w/ our best parameters)	63.40	77.87
4	DeBERTaV3$_{large}$ + KT	64.19	78.10
5	DeBERTaV3$_{large}$ + KT + CV	66.74	79.39
6	DeBERTaV3$_{large}$ + KT + CV + Contrastive + RP	82.07	82.07
7	DeBERTaV3$_{large}$ + KT + CV + Contrastive + Ensemble(5) + RP	82.62	82.62
8	DeBERTaV3$_{large}$ + KT + CV + Contrastive + Ensemble(8)	78.96	80.53
9	**DeBERTaV3$_{large}$ + KT + CV + Contrastive + Ensemble(8) + RP**	**83.69**	**83.69**
10	Human	95.00	96.50

Table 1: Summary of our results on the Com2Sense test set: KT stands for Knowledge Transfer, CV stands for Cross Validation, RP stands for Rule-based Perturbation, Ensemble(5) stands for a 5-model ensemble between DeBERTaV3$_{large}$ and DeBERTaV3$_{base}$, Ensemble(8) stands for an 8-model ensemble between DeBERTaV3$_{large}$, DeBERTaV3$_{base}$, and RoBERTa$_{base}$. The best model and method is highlighted with bold texts. For fairness of comparison, all the above model performances are measured on the official Com2Sense test set.

finetuning parameters used by their respective authors. Our results show DeBERTaV3$_{base}$ to be the best structure with 48.74% pairwise acc., while DeBERTa$_{base}$ and RoBERTa$_{base}$ have similar performance at ~18%. BERT$_{base}$ is the lowest performing model at ~3%

To find the best model size, we conduct multiple experiments with DeBERTaV3$_{base}$ and DeBERTaV3$_{large}$ under the best finetuning parameters used by He et al. (2021). The results show that DeBERTaV3$_{large}$ reaches 68.34% pairwise acc. while DeBERTaV3$_{base}$ reaches 52.76%. This supports the hypothesis that larger models have stronger common sense reasoning ability.

After choosing the best performing model DeBERTaV3$_{large}$ as our base model, we perform hyperparameter tuning on model parameters including: batch size (equivalent batch size after gradient accumulation), learning rate, and warmup steps. In each case, we fix all other parameters and test the effect of different values for the parameter under investigation. The testing results are documented in Table 2 lines 9-14, and the best set of parameters is highlighted in line 10.

3.2 Knowledge Transfer

3.2.1 SemEval-2020 Dataset

As shown in the following table, the model with transferred knowledge from SemEval-2020 performs better than the best model directly applied to Com2Sense, with a 0.364% improvement on pairwise accuracy.

Model	Pairwise Acc %	F1-Score
best DeBERTaV3$_{large}$	68.34	0.8103
SemEval-pretrained	68.84	0.8139

3.2.2 SQuAD2 QA Dataset

From the following table, we can see that models pretrained on question-answering data did not perform as well as those that have not. This can be because question-answering is a vastly different task than binary classification.

Model	Pairwise Acc %	F1-Score
RoBERTa$_{base}$	18.84	0.546
RoBERTa$_{base}$-SQuAD2	13.56	0.554
DeBERTaV3$_{large}$	68.34	0.810
DeBERTaV3$_{large}$-SQuAD2	65.83	0.789

3.3 Contrastive Learning and Random Perturbation

From Table 1 lines 5-8, we observe that in practice contrastive learning together with the Random Perturbation helped to improve test performance by 16%. In this case, we count that random perturbation changed 371 out of 2790 pairs in the test set. While this can have a maximum influence of 13.29% if all changed pairs turn out to be correct, since it is a purely random perturbation, on average it should have improved pairwise accuracy by 6.65%.

After removing the benefits of Random Perturbation, we conclude that Contrastive Learning yields an improvement of 8.77% on average, 2.04% in the worst case. The improvement can be attributed to the fact that the Com2Sense dataset comes in a natural contrastive fashion, with uniform true/false pairs that need to be differentiated from each other.

3.4 Model Ensemble

From Table 1 lines 6-8, we observe that in practice model ensemble as a post-processing technique helps the model perform better compared to straight-through Random Perturbation, likely because Random Perturbation only has a 50% chance of correctly predicting a pair while models used in the ensemble have a much higher accuracy.

In addition, the ensemble among DeBERTaV3$_{\text{large}}$, DeBERTaV3$_{\text{base}}$, and RoBERTa$_{\text{base}}$ models outperforms the ensemble between DeBERTaV3$_{\text{large}}$ and DeBERTaV3$_{\text{base}}$ models by 1.07% pairwise acc. This results supports the common understanding that diversity in model structures is beneficial for the ensemble.

4 Discussion

4.1 Results Analysis

We make use of the domain, scenario, and numeracy dimensions of Com2Sense, take the best-performing model of BERT$_{\text{base}}$, DeBERTaV3$_{\text{base}}$, DeBERTaV3$_{\text{large}}$, and SemEval-pretrained DeBERTV3$_{\text{large}}$, and then calculate each model's pairwise accuracy on Com2Sense dev set in every possible combination of the three dimensions.

In general, BERT$_{\text{base}}$ gives the lowest pairwise accuracy, as shown in Table 2. The top graph in Figure 2 further reveals that the model correctly predicts none of the numeracy data. Comparatively, it gives better predictions on sentences with comparison than with causal relationship, and yields higher pairwise acc. on temporal sentences than physical and, lastly, social.

DeBERTaV3$_{\text{base}}$ gives boosted pairwise accuracy. From the bottom graph in Figure 2, it performs better on comparative data than causal for all domains. We get slightly better results with numeracy than without numeracy for the physical domain, but in reverse for the social domain. The pattern for the temporal domain is more mixed: data with numeracy information has higher pairwise accuracy than comparisons but decreases for causal scenarios.

DeBERTaV3$_{\text{large}}$ improves the pairwise accuracy for all categories, but in particular more salient for social domain and numeracy data, as shown in Figure 3 top graph. It mostly preserves the pattern of DeBERTaV3$_{\text{base}}$, while performing better on data with numeracy information for temporal, causal sentences.

DeBERTV3$_{\text{large}}$ pretrained on SemEval dataset (Figure 3 bottom graph) generally performs better in social domain than physical, worst in temporal; it also improves on data without numeracy. We do observe, though, the exceptionally higher performance on temporal, comparative, and numeric sentences, probably affected by knowledge transferred from SemEval. The pretraining might have also improved DeBERTV3$_{\text{large}}$'s ability to learn causal reasoning as its pairwise accuracy increases for causal scenarios.

4.2 Limitations

The following are areas that we could improve during the stages of training and finetuning. Firstly, due to different hardware limitations on each of our virtual machines, we were not able to maintain consistent per-GPU batch size when training models throughout the project. While some of the trials used 6-instance batches over 8 accumulation steps, others were only able to use 4-instance batches over 12 accumulation steps. Such inconsistency in parameters could have impacted our final results. Secondly, due to time constraints, we only tested a limited range of hyperparameters which was not guaranteed to be the global optimum.

5 Conclusion

In this on-going research project, we have experimented with various methods to improve general purpose language models on the commonsense learning and reasoning task benchmarked by Com2Sense. We showed that: knowledge transfer from existing commonsense datasets; pairwise contrastive learning from commonsensical statements and their perturbed counterparts; and ensembling models with heterogeneous backbones yielded the greatest overall performance gains among other methods. Experiments applying the methods demonstrate substantial improvements in all metrics over current SOTA works in the field.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *CoRR*, abs/2111.09543.

Jaehun Jung, Lianhui Qin, Sean Welleck, Faeze Brahman, Chandra Bhagavatula, Ronan Le Bras, and Yejin Choi. 2022. Maieutic prompting: Logically consistent reasoning with recursive explanations. In *Conference on Empirical Methods in Natural Language Processing*.

D. Khashabi, S. Min, T. Khot, A. Sabhwaral, O. Tafjord, P. Clark, and H. Hajishirzi. 2020a. Unifiedqa: Crossing format boundaries with a single qa system.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020b. Unifiedqa: Crossing format boundaries with a single qa system. In *Findings*.

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. *CoRR*, abs/1907.11692.

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *ArXiv*, abs/1910.10683.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *Annual Meeting of the Association for Computational Linguistics*.

Maarten Sap, Vered Shwartz, Antoine Bosselut, Yejin Choi, and Dan Roth. 2020. Commonsense reasoning for natural language processing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 27–33, Online. Association for Computational Linguistics.

Shikhar Singh, Nuan Wen, Yu Hou, Pegah Alipoormolabashi, Te-lin Wu, Xuezhe Ma, and Nanyun Peng. 2021. COM2SENSE: A commonsense reasoning benchmark with complementary sentences. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 883–898, Online. Association for Computational Linguistics.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748.

Cunxiang Wang, Shuailong Liang, Yili Jin, Yilong Wang, Xiaodan Zhu, and Yue Zhang. 2020. SemEval-2020 task 4: Commonsense validation and explanation. In *Proceedings of The 14th International Workshop on Semantic Evaluation*. Association for Computational Linguistics.

6 Appendix

6.1 Performance Breakdown

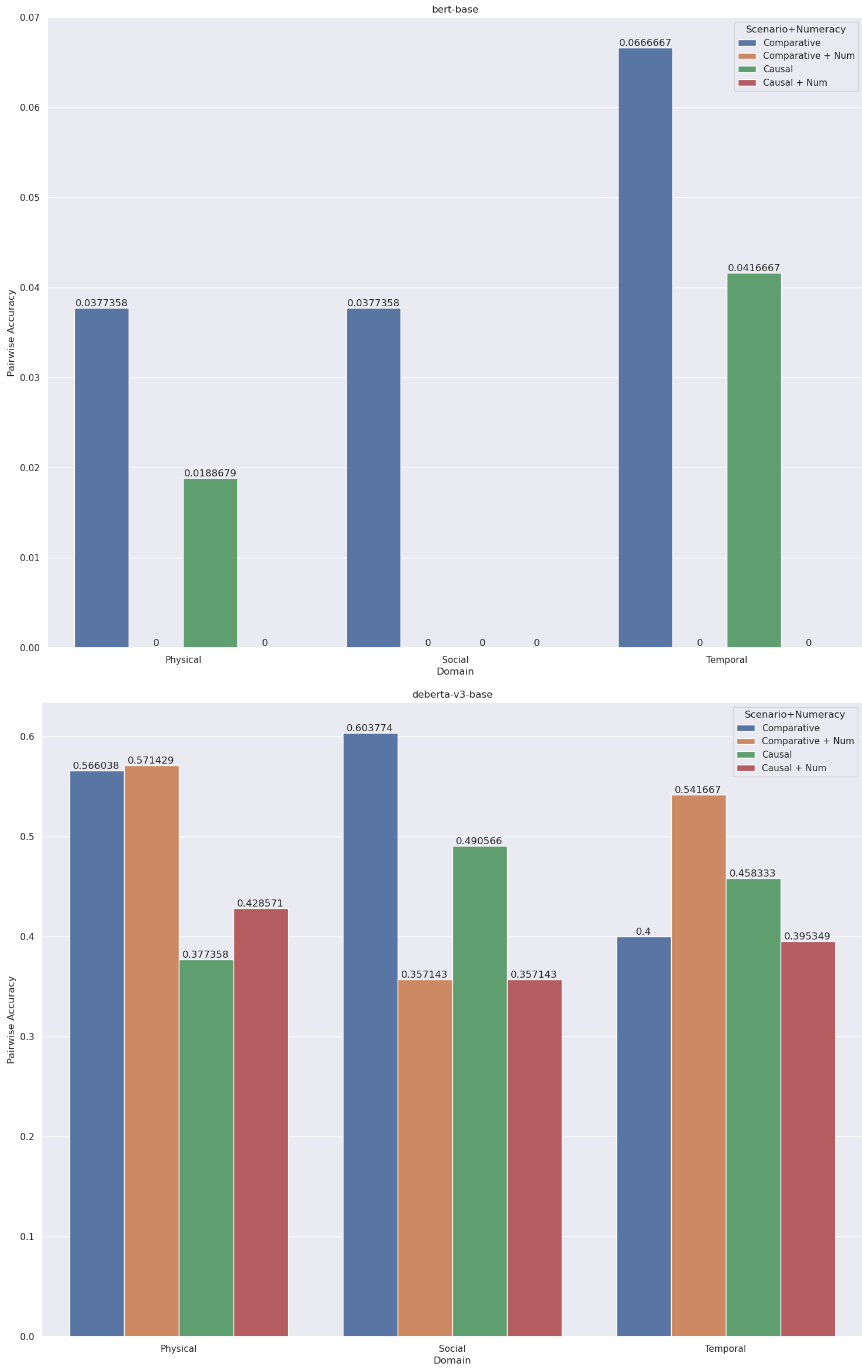

Figure 2: The top graph shows the pairwise accuracy of different dimension combinations for $\text{BERT}_{\text{base}}$, and the graph below shows the pairwise accuracy of different dimension combinations for $\text{DeBERTaV3}_{\text{base}}$.

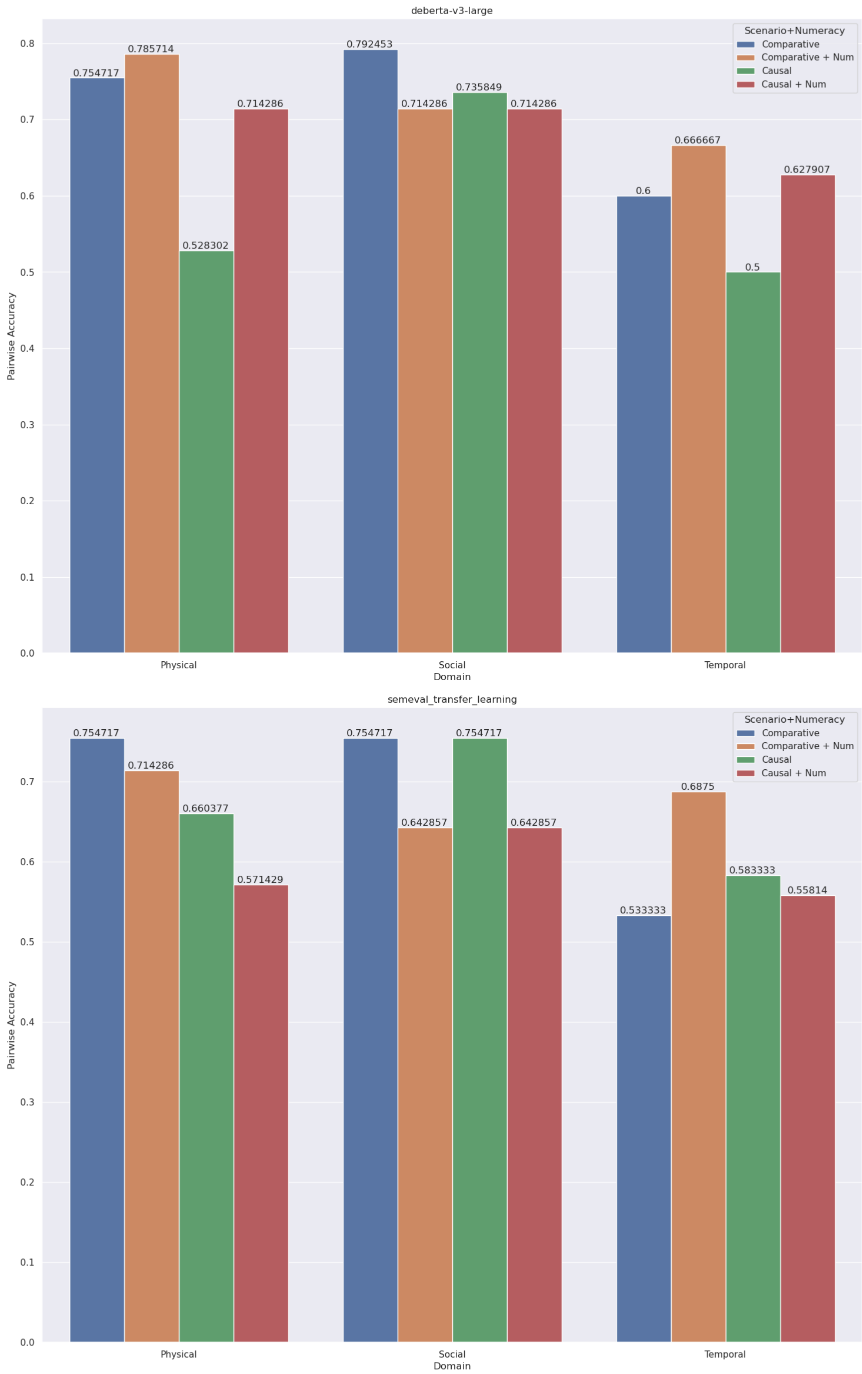

Figure 3: The top graph shows the pairwise accuracy of different dimension combinations for $\text{BERT}_{\text{large}}$, and the graph below shows the pairwise accuracy of different dimension combinations for $\text{DeBERTaV3}_{\text{large}}$ pretrained on SemEval dataset.

6.2 Table of Best Parameters

Line No.	Model	best step	batch size	lr	weight decay	adam ϵ	warmup step	Pairwise Acc %	F1 score
1	$\text{BERT}_{\text{base}}$	1020	32	1e-5	0	1e-8	0	3.01	0.4073
2	$\text{BERT}_{\text{large}}$	60	64	5e-5	0	1e-8	0	2.51	0.3720
3	$\text{RoBERTa}_{\text{base}}$	1040	64	1e-5	0.01	1e-8	0	18.84	0.5463
4	$\text{DeBERTa}_{\text{base}}$	6000	32	1e-5	0	1e-8	500	17.84	0.5302
5	$\text{DeBERTaV3}_{\text{base}}$	4500	48	1e-5	0.01	1e-6	500	48.74	0.7145
6	$\text{DeBERTaV3}_{\text{base}}$	1500	48	3e-5	0.01	1e-6	100	52.76	0.7219
7	$\text{DeBERTaV3}_{\text{base}}$	2500	48	3e-5	0.01	1e-6	500	49.00	0.7057
8	$\text{DeBERTaV3}_{\text{base}}$	1000	48	9e-6	0.01	1e-6	500	45.48	0.6767
9	$\text{DeBERTaV3}_{\text{large}}$	750	64	9e-6	0.01	1e-6	500	67.84	0.8090
10	**$\textbf{DeBERTaV3}_{\textbf{large}}$**	**1900**	**48**	**9e-6**	**0.01**	**1e-6**	**500**	**68.34**	**0.8103**
11	$\text{DeBERTaV3}_{\text{large}}$	1000	48	8.5e-6	0.01	1e-6	500	67.34	0.8111
12	$\text{DeBERTaV3}_{\text{large}}$	450	48	9.5e-6	0.01	1e-6	500	66.33	0.7990
13	$\text{DeBERTaV3}_{\text{large}}$	1000	48	9e-6	0.01	1e-6	300	67.59	0.8059
14	$\text{DeBERTaV3}_{\text{large}}$	1400	48	9e-6	0.01	1e-6	750	66.58	0.8029

Table 2: Summary of hyper-parameter tuning with results calculated on the dev dataset, the experiments are focused on finding the best model backbone, model size and ideal values for hyper-parameters. The best performing model and ideal hyper-parameter group is highlighted in bold.

6.3 Cross Validation Details

The training dataset contains 797 pairs of examples and the development set has 398 pairs. We hypothesize that leveraging both datasets for training would yield a more generalized model with a higher level of reliability. We thus employed k-fold cross validation on our best DeBERTaV3$_{\text{large}}$ model and tested for $k = 2$ and $k = 5$.

As shown in Table 1 lines 3-5, cross validation helped to improve test performance by 2.5% empirically by incorporating the dev set for finetuning, which added around 50% more training data. Consequently, the training time escalates with the number of folds: 2-fold cross validation took around 20 hours to train and 5-fold took more than 50 hours.