CS 188 Final Project Report

Pranav Charkupalli
UCLA

Nicholas Dean
UCLA

Patrick Ian Wildenhain
UCLA

Eric Zhang
UCLA

1 Introduction

In this project, we attempted to leverage a large-scale language corpora pretrained model to gain experience with the practical side of NLP and to acquire the knowledge and skills required to work on NLP projects in the future. As part of this, we were given a starter codebase written in PyTorch and needed to finish some incomplete parts to have it functioning, as well as to select the MLM model we wished to use in our implementation.

After getting our implementation working, we trained it on the downstream Com2Sense task which consisted of two parts: supervised learning on the Com2Sense dataset where we worked with different hyperparameters and found the best combination, and knowledge transfer from the Sem-Eval dataset to improve performance on the Com2Sense task where we took a checkpoint from fine-tuning the implementation on Sem-Eval dataset and used it at the start of our Com2Sense training pipeline.

Finally, we worked on an open-ended question meant to simulate actual research problems with no set answer. Of the two possible options, we chose to try and improve the overall performance of our model, primarily by leveraging external knowledge bases. The details of our approach will be expanded upon further in our report.

2 Referenced Related Works

As previously mentioned, we planned to leverage external knowledge bases in order to improve the performance of our model. And to that end, one of the most useful works we encountered was "UNICORN on RAINBOW: A Universal Commonsense Reasoning Model on a New Multitask Benchmark". The paper was focused around proposing new metrics for evaluating models that make use of multiple datasets.

But what was more important for our implementation was one of the conclusions of the paper (Lourie et al., 2021):

> ...intermediate-task transfer can always lead to better or equivalent performance if following a particular recipe, that QA-based commonsense datasets transfer well to each other, while commonsense knowledge graphs do not, and that perhaps counter-intuitively, larger models benefit much more from transfer learning compared to smaller ones. (p. 2)

In terms of our dataset choice, we had further evidence to use QA from another paper: "Intermediate-Task Transfer Learning with Pretrained Models for Natural Language Understanding: When and Why Does It Work?". The paper concluded that while intermediate-task training was often beneficial for a final target task, the greatest gains in performances were seen by choosing intermediate tasks that required high-level inference and reasoning abilities (Pruksachatkun et al., 2020), which a commonsense dataset would be able to provide.

Reading through the papers encouraged us to make modifications to our previous approach of leveraging datasets. Whereas we started out trying to use semantic knowledge bases like ConceptNet, we decided to instead switch to QA-based commonsense datasets like Physical IQA and Social IQA. We also made use of the paper's discovery that, when working with multiple datasets, sequential training beats out multitask training or finetuning.

3 Main Methods

This section details the methods we used to deal with the provided experimental tasks as well as the chosen open-ended question of improving model performance through additional means.

3.1 Experimental Task

We first conducted supervised learning by training using the SemEval dataset and the Com2Sense dataset. Using some hyperparameter tuning on a baseline model, we settled on using a learning rate of 1E-5 and trained the model for 100 epochs. The model was trained using a dataset reserved for training while evaluating on the dev dataset, reserved for training evaluation. We tested multiple training checkpoints and multiple parameter combinations with a testing dataset to pick out the model with the best performance. We choose to focus on training off of a pre-trained BERT model to get a baseline model for the Semeval, Com2Sense, and Physical IQA tasks.

Additionally, we experimented with transfer learning to investigate the effects of knowledge transfer between models. For this, we trained a commonsense model using a baseline trained with the SemEval dataset or Physical IQA for the open-ended investigation. We then compare the results of these models to determine whether knowledge gained from a different task would transfer over to improve the accuracy of the commonsense task. In analyzing the model's results, we can isolate each domain and scenario to understand how transfer learning affected each subcategory.

3.2 Open-Ended Question

As previously mentioned, we chose the open-ended question of trying to improve the overall performance of the model, and to that end, we tried utilizing the Physical and Social IQA commonsense datasets. In particular, we tried using the approach encouraged by Pruksachatkun (2020) of training on these datasets as intermediate-tasks before performing the supervised learning and knowledge transfer tasks on Com2Sense. We ended up choosing to experiment with the Physical IQA dataset and used it to train a baseline model. We could then use this Physical IQA trained model to investigate the effects of knowledge transfer by using the model to train the commonsense task.

We chose Physical IQA to provide our model with physical commonsense information. Physical IQA works as a binary choice task, with there being a given goal and two choices to go about achieving them (Bisk, 2019). We hypothesized that using the Physical IQA dataset for transfer learning would increase the accuracy for the Com2Sense entries in the physical domain. For future experimentation, we could attempt training with Social IQA for commonsense information about social interactions, motivations, and reactions, which works by providing "multiple choice questions for probing emotional and social intelligence in a variety of everyday situations." (Sap, 2019).

4 Main Results

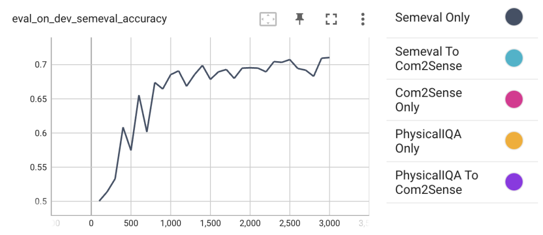

Figure 1: Semeval Baseline Accuracy

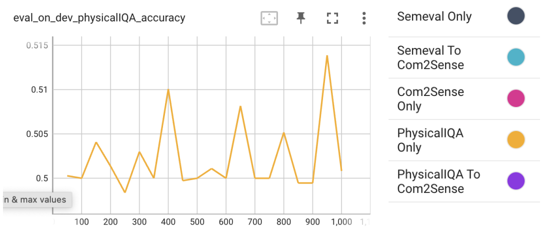

Figure 2: Physical IQA Baseline Accuracy

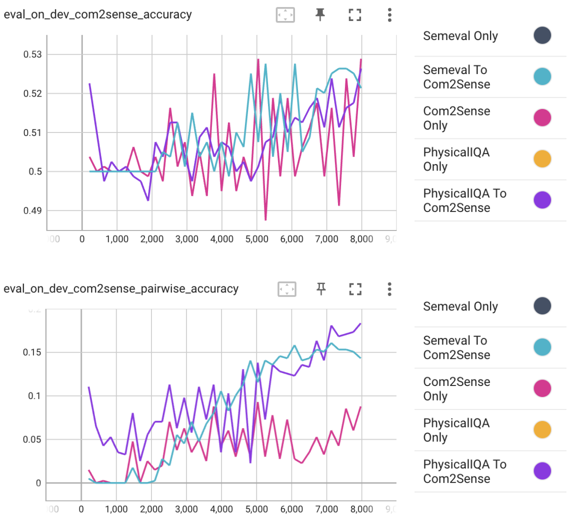

Figure 3: Com2Sense Accuracy and Pairwise Accuracy

Overall, we can see that our SemEval transfer learning trained model slightly outperformed our model trained purely on the Com2Sense task, supporting the idea that transfer learning can boost performance. Our Physical IQA transfer learning trained model also outperformed the base model, and despite having slightly a lower accuracy an F1 score compared to the SemEval transfer learning trained model, our PIQA trained model had a significantly higher pairwise accuracy.

Evaluation Results			
Model	Accuracy	Pairwise	F1
Com2Sense	0.5088	0.0799	0.6157
SemEval Transfer	0.5188	0.1330	0.5812
PIQA Transfer	0.5167	0.1570	0.5736

Examining our models' results on the test data set, we see the same trends. Both the SemEval and Physical IQA trained models outperformed the base Com2Sense model. The SemEval model performed the best in terms of standard accuracy, but the Physical IQA model again had the best pairwise accuracy.

4.1 Scenario and Domain Analysis

In taking a deeper look at the results of the various Com2Sense models we trained, we examined their performance on the validation set partitioned by domains or scenarios, to analyze whether they performed better within certain categories. This also allowed us to see whether the transfer learning had a more pronounced impact on certain categories of domain or scenario.

The validation dataset which predictions were recorded for and analyzed in this section consisted of a total of 796 examples. For this analysis, there are two attributes which we can partition these predictions by: scenario and domain. There are two different scenario labels: causal, which contains 402 examples, and comparison, which contains 394 examples. There are three different domain labels: physical, with 268 examples, social, with 268 examples, and time, with 260 examples. As we can see, within the categories of scenario and domain, each label has a roughly even number of examples, meaning that differing sample size should not be a major consideration in explaining differences between the performance within different partitions.

Com2Sense Only			
Category	Accuracy	Pairwise	F1
Overall	0.5251	0.0879	0.4019
Causal	0.5448	0.1095	0.4438
Comparison	0.5051	0.0660	0.3564
Physical	0.5075	0.0448	0.3714
Social	0.5410	0.1269	0.4332
Time	0.5269	0.0923	0.4000

SemEval Transfer Learning Com2Sense			
Category	Accuracy	Pairwise	F1
Overall	0.5276	0.1583	0.5922
Causal	0.5149	0.1592	0.5946
Comparison	0.5406	0.1574	0.5896
Physical	0.5187	0.1642	0.5743
Social	0.5187	0.1567	0.5981
Time	0.5462	0.1538	0.6040

In further examining and comparing the metrics of the partitioned results of these models to their respective overall results, we can take note of a few standout categories for each model.

For the model trained purely on the Com2Sense task, it performed far better on the causal scenario examples than the comparison scenario examples, having almost an additional .02 accuracy and .01 pairwise accuracy in the causal case, while the comparison case fell behind by .02 accuracy and .03 pairwise accuracy. With respect to domains, the model performed much better in the social domain, much worse in the physical domain, and about average in the time domain.

For the model trained with transfer learning from the SemEval task, there are some difference. Unlike the previous model, this model performed better accuracy-wise in the comparison scenario examples and worse in the causal scenario examples. However, in both scenarios, pairwise accuracy remained close to the overall metric. With respect to domains, this model performed better accuracy-wise in the time domain, but with a high pairwise accuracy in the frequency domain.

From these partitioned results, we can note that transfer learning had a significant impact on which scenarios and domains the model performed best at. While the SemEval transfer learning model had a better overall accuracy, it was inferior to the Com2Sense only model in causal scenarios and the social domain, but superior in every other category, standing out in comparison scenarios and the time domain.

PIQA Transfer Learning Com2Sense			
Category	Accuracy	Pairwise	F1
Overall	0.5264	0.1834	0.5662
Causal	0.5075	0.1393	0.5580
Comparison	0.5457	0.2284	0.5748
Physical	0.5075	0.1791	0.5541
Social	0.5373	0.1940	0.5724
Time	0.5346	0.1769	0.5724

Examining the partitioned results from our attempt at improving performance through transfer learning from the physical interaction question answering task, we can similarly determine scenarios and domains which stand out above the overall performance. In the table above, we can note that this model had much better accuracy and pairwise accuracy in comparison scenarios, with slightly better accuracy in social and time domains, while lagging behind in the causal scenarios and the physical domain. Interestingly, this improved performance in the comparison scenario is very similar to the SemEval transfer learned model, as both had a similar tendency, unlike the model trained purely on Com2Sense. Furthermore, while the model trained purely on Com2Sense did best in the time domain, and the SemEval model did best in the time domain, the Physical IQA model did well in both of these domains.

5 Conclusions and/or Discussions

In looking at the overall results of our three models, each models achieved similar levels of accuracy, but the models which took advantage of transfer learning had much better pairwise accuracy, particularly our model trained with Physical IQA. This demonstrated that the additional data we trained the model on was generally useful for discriminating between pairs of sentences.

Furthermore, additional analysis revealed that they each had a unique distribution of accuracy across different scenarios and domains. It would be interesting to try further transfer learning with more datasets like ConceptNet or other IQA datasets, with the goal of improvement for the scenarios and domains in which our models underperformed compared to the overall accuracy.

This project was a great way to introduce us to a practical NLP workflow. We were able to take our learnings from the class and use them to create a basic NLP model utilizing transformers. And in the process of completing the experimental task, we learned how to tune hyperparameters to improve performance while making use of knowledge transfer to further improve performance. Meanwhile, the open-ended question also gave us the opportunity to research different ways to go about improving model performance, and provided exposure to some extremely versatile QA commonsense knowledge bases.

With all this experience with using QA-based commonsense datasets, it would've also been interesting to try tackling the alternate open-ended question of generating plausible explanations. Since Physical and Social IQA deal with very high-level concepts that might be difficult to plainly explain in language, it would be worth considering whether using them in the implementation would possibly cause the explanations to end up being nonsensical, given how specific the reasoning can be behind some of the answers.

We'd also want to try this project using different NLP models and seeing how the results varied between them. Transformers are known to be one of the more powerful models thanks to the attention mechanism allowing the model to retain faraway context, and it'd be interesting trying to tweak other models in an effort to reach transformer performance levels.

6 References

Lourie, N., Bras, R. L., Bhagavatula, C., Choi, Y. (2021). (tech.). *UNICORN on RAINBOW: A Universal Commonsense Reasoning Model on a New Multitask Benchmark*. Retrieved March 17, 2022, from https://arxiv.org/pdf/2103.13009.pdf.

Pruksachatkun, Y., Phang, J., Liu, H., Htut, P. M., Zhang, X., Pang, R. Y., Vania, C., Kann, K., Bowman S. R. (2020). (tech.). *Intermediate-Task Transfer Learning with Pretrained Models for Natural Language Understanding: When and Why Does It Work?* Retrieved March 17, 2022, from https://arxiv.org/pdf/2005.00628.pdf.

Bisk, Y. (2019). *Physical Interaction: Question Answering*. PIQA: Physical interaction - Question answering. Retrieved March 17, 2022, from https://yonatanbisk.com/piqa/

Sap, M., Rashkin, H., Chen, D., Bras, R. L., Yejin Choi. (2019). (tech.). *SocialIQA: Commonsense Reasoning about Social Interactions*. Retrieved March 17, 2022, from https://arxiv.org/pdf/1904.09728.pdf.

7 Model Checkpoints

Semeval Only Model
Physical IQA Only Model
Com2Sense Only Model (Trial 21)
Semeval to Com2Sense Model (Trial 22)
Physical IQA to Com2Sense Model (Trial 23)