CS378: Final Project - Data Artifacts https://github.com/lulucopter/curriculum-learning-NLP

Lee Zheng Yao Daniel
dzl237

Abstract

Stanford Natural Language Inference (SNLI) dataset (Bowman et al., 2015) is a widely used benchmark for natural language inference (NLI). However, upon further inspection of the dataset, we discover some dataset artifacts, in which a model achieves high performance by learning spurious correlations. In this study, we tackle annotation artifacts introduced from the annotators' language used, and investigate the effectiveness of curriculum learning on mitigating these data artifacts.

1 Introduction

Natural Language Processing (NLP) has been on the rise due to its vast applications, especially recently with the introduction of Large Language Models (LLMs) such as ChatGPT. One common application is natural language understanding (NLU) tasks. One problem in NLU is natural language inference (NLI), where a model is tasked with identifying the logical relationship between two given texts: a premise and a hypothesis. The model then attempts to predict if the hypothesis can be inferred, contradicts, or is neutral in relation to the premise. One such dataset used to benchmark NLI tasks is the SNLI dataset. However, no dataset is perfect, and there are dataset artifacts which models can learn to exploit in NLI (Poliak et al., 2018). An issue with the SNLI dataset are the presence of annotation artifacts, which leads to models exploiting annotation biases instead of actually learning the underlying semantic relationships (Gururangan et al., 2018).

Introducing curriculum learning (Bengio et al., 2009). Curriculum learning is the formalization of training strategies in machine learning based on the difficulty of training. Curriculum training has been shown to lead to faster convergence (Krueger Dayan, 2009), and shown to achieve better generalizability (Bengio et al., 2009), and improves performance in domains such as computer vision and NLP (Graves et al., 2017; Platanios et al., 2019). However, in their research, they investigated curriculum learning in the context of language modelling. In this study we attempt to apply curriculum learning in the context of NLI, and investigate its' effectiveness.

2 Task/Dataset/Model Description

In the SNLI dataset, our objective is to predict the relationship between a given premise (P) and hypothesis (H). Their relationship can be one of the following possible classes: entailment, contradiction, or neutral.

Mathematically, let P denote the given premise, and H denote it's hypothesis. Our task then is to model the function f such that:

$$f(P, H) : y \in (entailment, contradiction, neutral) \tag{1}$$

In this work, we apply transfer learning and train the pretrained ELECTRA-small model (Clark et al., 2020), which model's architecture is based on BERT, on a total of 22,000 examples from the SNLI dataset, with and without curriculum learning during training.

The learning algorithm used is based on the transformer architecture (Vaswani et al., 2017), utilizing attention mechanisms and self-attention to capture contextual information related to the task. The loss function is given as follows:

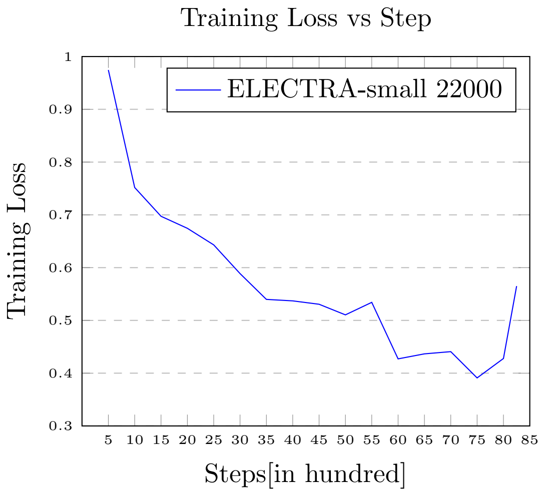
$$L(y, \hat{y}) = -\sum(y_i * log(\hat{y}_i)) \tag{2}$$

where y is the true label vector, $\hat{y}$ is the predicted probability vector, and $i \in (entailment, contradiction, neutral)$.

The stopping criteria is when we finish training the model over 3 epochs of its' 22,000 examples (2% of the SNLI dataset). This should be sufficient to train the model to a reasonably high accuracy, while preventing overfitting of the training data or its artifacts, while allowing us to observe which model learns concepts faster.

3 Performance Analysis

During training, we observe a general monotonic decrease in training loss fairly, indicating that the model learns the logical relationships of the NLI task relatively quickly.



After training the model, the overall performance is as follows:

Model	Loss	Accuracy
Untrained ELECTRA	1.0984	0.33753
Trained ELECTRA	0.58762	0.815891

Table 1: Baseline models

The model was then systematically analyzed for the specific errors and the general class of mistakes a model makes. Generally, it is identified that the model struggles with negation handling, and often incorrectly classifies examples containing negation, such as sentences with ”not” or ”never”. For example, the following is labelled wrongly:

- Premise: A man in a red shirt is sitting at his desk with one leg on the desk, using a computer with two screens set up side by side.
- Hypothesis: A male is not sitting outside.
- Model Prediction: Contradiction
- Ground Truth: Entailment

The model also sometimes struggled with ambiguous examples, where there wasn't a clear agreement on the ground truth by annotators, indicating potential annotation inconsistencies or artifacts. An example of ambiguous data:

- Premise: A person with dark hair is standing on the sidewalk in front of an orange and white truck.
- Hypothesis: The person gets in the white truck.
- Model Prediction: Neutral
- Ground Truth: Entailment

The above is one an ambiguous example in the dataset. On one hand, we can ”deduce” that in the future, the person will be getting in the white truck, as that is the most probable reason that person is currently stand right in front of the truck. But on the other hand, one can also argue that the person is just simply standing right in front of the truck, and that has no bearing on it's future actions.

4 Describing Your Fix

To mitigate these ”hard” and ”ambiguous” examples, we first have to identify what are easy to learn, ambiguous, and hard to learn. To do this, we will utilize dataset cartography (Swayamdipta et al., 2020). Data is classified based on their correctness, which is defined as the fraction of times the model correctly labels x_i across epochs. In this study, the difficulties are partitioned as follows:

DataMap	Difficulty	Correctness Score
easy-to-learn	easy	[5,6]
ambiguous	medium	[3,4]
hard-to-learn	hard	[0,1,2]

Table 2: Difficulty of problems

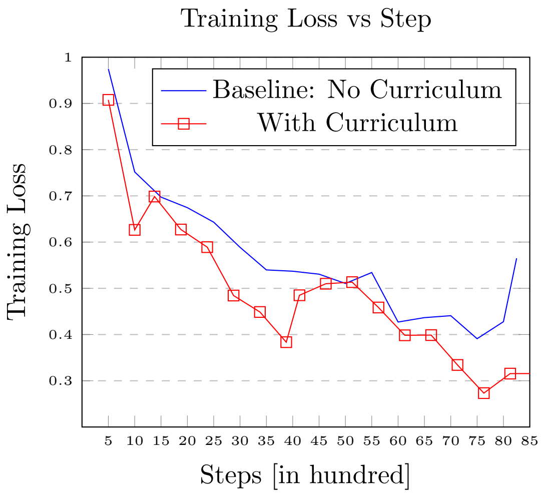
Now, curriculum learning (Bengio et al., 2009) is used onto the different difficulties of examples to guide the model's training. This involves training the model on easier examples first to first help the model understand basic logical inferences, before easing the model into increasing difficulty levels as the model progresses.

The same basic training hyperparameters are used as the baseline models, with the addition of 1 new hyperparamter, curriculum split. This is defined as the fraction of examples we increase a dataset's difficulty by in the curriculum. The same curriculum split is used as the Natural Language Model example in Bengio et al.'s paper. Our curriculum is split into 3 phases. First, the model is solely trained on easy examples, before introducing the same amount of medium examples, and finally with hard examples, for a total number of 22,000 examples as with the baseline model.

During training, the same loss function and stopping criteria is used as the baseline model, with the only difference being the curriculum learning strategy that we will employ. The models are then evaluated on the same test set, and then compared against each other to assess the effectiveness of curriculum learning on the NLI task.

5 Evaluating Your Fix

In this section, the effectiveness of curriculum learning on the NLI task on the SNLI dataset is analyzed. The training losses of the model implementing curriculum learning and it's baseline model are plotted as follows:



We then evaluated both models as presented below:

Model	Loss	Accuracy
Untrained	1.0984	0.33753
Normal Training	0.58762	0.815891
Curriculum Training	0.83165	0.80980

Table 3: Baseline models

Both baseline and curriculum models perform relatively well, achieving similar accuracies on the evaluation dataset. It is to be noted that the normal trained model achieves a slightly better accuracy than the curriculum trained model by around 0.6%, which is not too significant given the small example size (2% of SNLI dataset). However, it is to be noted that the training loss for normal training is significantly much lower than the training loss in curriculum training, suggesting that normal training may have learnt the dataset artifacts, and possibly overfitted to it.

The new curriculum model is then analyzed on the examples which the baseline model struggles to classify correctly to verify our hypothesis. It is then observed that the curriculum trained model performs much better on these examples, suggesting that the curriculum trained model, despite its much higher loss, performs significantly much better than the normal trained model. This shows that curriculum learning is more effective in teaching the model difficult concepts, especially on complex, ambiguous and hard to train examples. On the same examples which the baseline model struggled to classify:

Hard to learn (Negation handling)

- Premise: A man in a red shirt is sitting at his desk with one leg on the desk, using a computer with two screens set up side by side.
- Hypothesis: A male is not sitting outside.
- Baseline Model Prediction: Contradiction
- Curriculum Model Prediction: Entailment
- Ground Truth: Entailment

Ambiguous

- Premise: A person with dark hair is standing on the sidewalk in front of an orange and white truck.
- Hypothesis: The person gets in the white truck.
- Baseline Model Prediction: Neutral
- Curriculum Model Prediction: Entailment
- Ground Truth: Entailment

Hence, we can conclude that curriculum learning is useful and effective in reducing the likelihood of a model learning and fitting data artifacts, performing significantly better than the baseline model on medium and hard difficulty examples.

6 Related Work

Our approach utilizes the original idea of curriculum learning, first proposed by Bengio et al. (2009), and integrates this learning approach with dataset cartography (Swayamdipta et al., 2020). Swayamdipta et al. (2020) worked and was able to characterize data in the SNLI dataset into various difficulties accurately. This enables us to specifically split examples into more accurate partitions, and better curate a curriculum for our model to target and learn–challenging and ambiguous examples, which are more likely to be affected by annotation artifacts.

7 Conclusion

This study demonstrated the potential benefits of curriculum learning on NLI tasks on the SNLI dataset. Although the overall accuracy of our curriculum trained model was very slightly lower than the baseline model, it show significant improvement on ambiguous and hard examples, suggesting that that it may lead to more robust understanding of patterns in data and a better handling of challenging examples.

Our fix, incorporating both dataset cartography and curriculum learning, has much potential outside of just NLI. This approach can be beneficial in instances where a dataset contains examples with a large variance in difficulty, and the model is required to handle and capture these complex relationship and patterns.

However, Curriculum learning does come at a trade off. Curriculum learning increases the complexity of the training process, and in our case, the improvement in performance on challenging examples came as the cost of a very slight decrease in overall accuracy. It is important for one to carefully evaluate their requirements and goals to determine the suitability of this approach.

Overall, curriculum learning has been shown to have promising results in tackling challenging examples, highlighting its value in training models for complex tasks. However, it is essential for one to go over their specified requirements and its' trade-offs to determine the most suitable approach in training models.

AI Assistance

AI assistance was used to help fact check that I correctly interpreted sources, and to double check the grammar of sentences.