Replication: Distilling Task-Specific Knowledge from BERT into Simple Neural Networks

Vayunandhan Reddy, Muni Lohith, Maanusri and Shilpa Sweth
{vyekkaluri, mkonidala, mbalasubrama, ssweth}@umass.edu
https://github.com/Letzt/distilBERT_BiLSTM

1 Problem statement

In recent years, models used in natural language processing are becoming increasingly deeper and complex. Large and complex models like BERT and GPT are better at capturing the nuances and complexities of natural language (like syntax, semantics, and context). While fine-tuned for NLP tasks like text classification, named entity recognition, and machine translation these models achieve great performance. However, the high demand for computing resources in training such models hinders their application in practice.

Distillation helps us leverage the knowledge learned by the larger model (trained on massive amounts of data) and transfer it to a smaller model that can perform the same task with a lower computational cost, making it more practical to deploy the model in production environments. This also aids in improving the efficiency and speed of training, as smaller models can be trained more quickly than larger ones. This can be achieved by training the smaller model to mimic the outputs of the larger model, using the same input data.

In Tang et al. (2019) a simple yet effective approach has been proposed to compress the original teacher (e.g., BERT) into a lightweight student model without performance sacrifice. Our work investigates the efficacy of this approach by implementing and expanding on the proposed idea of training the student model and comparing the results with the teacher model on benchmark datasets.

2 What you proposed vs. what you accomplished

- ~~Data Augmentation:~~ A rule-based data augmentation on the existing datasets was carried out to generate a large knowledge dataset. We followed the folowing four heuristics to facilitate task-agnostic data augmentation: Masking, POS-guided word replacement, n-gram sampling, Random word swap. For a sentence we do the following
  - If $x < p_{mask}$ then we applied Masking.
  - If $p_{mask} \leq x \leq p_{mask} + p_{pos}$ then applied POS-guided swapping.
  - Else continue onto the next word.

  For pairwise sentence datasets we do the above by keeping one sentence fixed at a time and also manipulte both sentences at the sametime.
- ~~Distilled knowledge from BERT to the BiLTSM model for a single sentence dataset. The regular BiLSTM architecture.~~
- ~~Distilled knowledge from BERT to the BiLTSM model for a paired sentence dataset. The siamese BiLSTM arcitecture.~~
- ~~Implemented custom loss based on the teacher and student logits.~~ Tweaked the traditional cross-entropy loss function of the student model. Our final loss function of the student is a weighted sum of distillation loss(mean squared error loss with the teacher logits) and the cross-entropy loss(with the true logits).

$$\mathcal{L} = \alpha.\mathcal{L}_{CE} + (1 - \alpha).\mathcal{L}_{distill} \quad (1)$$

- *PKD-Last Implementation*: The Patient Knowledge Distillation (Sun et al. (2019)) paper deals with distilling from BERT to a smaller k-layer version of BERT. However, our student model is a different architecture altogether. So experiments trying to map the mismatched intermediate layers did not result in any favourable observations.

3 Related work

In recent years, knowledge distillation has gained popularity as a method for compressing large, complex models into smaller, more efficient ones. In the context of natural language processing (NLP), several studies have explored the use of distillation techniques to improve the performance of NLP models.

In one such study, Sanh et al. (2019) proposed DistilBERT, a distilled version of the popular BERT language model. DistilBERT was trained to replicate the behavior of the original BERT model while using fewer parameters, resulting in a model that is both faster and more memory-efficient. Similarly, Jiao et al. (2019) proposed TinyBERT, a technique for distilling BERT models that results in even smaller and more efficient models.

There were also multiple works that focused on making better use of soft labels to transfer more knowledge. WSLD (from Zhou et al. (2021)) takes into account the bias-variance trade-off when analyzing soft labels and assigns them different weights accordingly. In Zhao et al. (2022), the author proposes to modify the formulation of knowledge distillation according to the teacher's prediction. Forcing the output logits of the teacher's and student's features to be the same has proved to improve the student's performance in Yang et al. (2021).

Some other research focuses on transferring knowledge from intermediate features, not just the logits. Romero et al. (2014) distills the semantic information from intermediate features directly, while Heo et al. (2019) modifies the measurement for the distance between students and teachers using margin ReLU. Park et al. (2019) extracts the relation from the feature map, while Tian et al. (2019) applies contrastive learning to distillation successfully. Chen et al. (2021) transfers knowledge from multi-level features for distillation and, finally, Yang et al. (2022) proposes a new distillation method that makes the student generate the teacher's feature instead of mimicking it.

Other studies have explored the use of distillation for transfer learning in NLP. For example, Sun et al. (2019) proposed a method for transferring knowledge from a pre-trained language model to a smaller, task-specific model using distillation. They showed that this approach can significantly improve the performance of the task-specific model while reducing training time and computational resources.

Distillation has also been used as a form of model compression in NLP. In one study, Chen et al. (2020) proposed a method for compressing pre-trained language models by selectively pruning neurons and fine-tuning the remaining ones using distillation. They showed that this approach can result in models that are up to 30% smaller and 60% faster than the original models, with only a minimal loss in performance.

As in the aforementioned paper Tang et al. (2019), our approach leverages pre-trained language models and task-specific data to create smaller, more efficient models that are optimized for a specific task. We plan to evaluate the results from our model against the outcomes mentioned in the paper. We also aim to run this on several benchmark datasets and show that the small student model that learns from a large complex teacher model outperforms a similar-sized model that is directly trained for this task.

4 Your dataset

We plan on testing our approach with the datasets mentioned below. The teacher model used is a fine-tuned version of bert-large-cased originally released in "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" and trained on the corresponding datasets; part of the General Language Understanding Evaluation (GLUE) benchmark.

SST-2 Stanford Sentiment Treebank 2 (SST-2; Socher et al. (2013)) comprises single sentences extracted from movie reviews for binary sentiment classification (positive vs. negative). Sentences are complete and express opinions about some film. Labels are denoted by 0 (Negative sentiment) or 1 (Positive sentiment).

- After augmentation the dataset has 404066 training examples, 872 validation examples and 1821 test examples.
Sentence: goes to absurd lengths
Label: 0

MNLI The Multi-genre Natural Language Inference (MNLI; Williams et al. (2018)) corpus is a large-scale, crowdsourced entailment classification dataset. The objective is to predict the relationship between a pair of sentences as one of en-

tailment, neutrality, or contradiction. Labels are denoted by 0 (The pairing indicates entailment) or 1 (The pairing doesn't indicate anything in particular) or 2 (The pairing indicates contradiction). The matched set consists of sentence pairs that are drawn from the same genres as the training set while the mismatched set has sentence pairs of different genres than the training set. This gives us insights of the generalization and robustness of the student model.

- The dataset contains 392702 training examples, 9815 validation examples and 248 test examples. Premises and hypothesis are paired sentences.
Premise: I burst through a set of cabin doors, and fell to the ground
Hypo: I burst through the doors and fell down.
Label: 0

QQP Quora Question Pairs (QQP; Iyer and et al. (2017)) consists of pairs of potentially duplicate questions collected from Quora, a question-and-answer website. The binary label of each question pair indicates redundancy. The two sentences are questions asked on quora. Labels are denoted by 0 (The two questions don't have the same meaning) or 1 (The two questions have the same meaning).

- After augmentation the dataset has 363846 training examples, 40430 validation examples and 390965 test examples.
Question 1: What can one do after MBBS?
Question 2: What do i do after my MBBS ?
Label: 1

MRPC Microsoft Research Paraphrase Corpus (MRPC; Dolan et al. (2004)) contains pairs of sentences and corresponding labels, which indicate the semantic equivalence relationship between each pair. The two sentences that are to be judged whether are paraphrases of each other. Labels are denoted by 0 (There is no paraphrasing) or 1 (There is paraphrasing).

- After augmentation the dataset contains 58677 training examples, 408 validation examples and 1725 test examples.
Sentence1: Legislation making it harder for consumers to erase their debts in bankruptcy court won overwhelming House approval in March .
Sentence2: Legislation making it harder for consumers to erase their debts in bankruptcy court won speedy, House approval in March and was endorsed by the White House .
Label: 0

RTE Recognizing Textual Entailment is based on a series of textual entailment challenges, created by General Language Understanding Evaluation (GLUE) benchmark (Wang et al. (2018)). The two sentences that are to be judged whether the second one is an entailment of the first. Labels are denoted by 0 (There is no entailment) or 1 (There is entailment).

- After augmentation the dataset contains 21147 training examples, 277 validation examples and 3000 test examples.
Sentence1: Valero Energy Corp., on Monday, said it found "extensive" additional damage at its 250,000-barrel-per-day Port Arthur refinery.
Sentence2: Valero Energy Corp. produces 250,000 barrels per day.
Label: 0

Dataset	Train Set	Validation Set	Test Set
SST2		872	1821
Aug SST2	404066	872	1821
MNLI	392702	9815	248
QQP	363846	40430	390965
MRPC	3668	408	1725
Aug MRPC	58677	408	1725
RTE	2490	277	3000
Aug RTE	21147	277	3000

4.1 Data preprocessing

Firstly, we imported the Hugging Face datasets to get the train, test and validation splits for each dataset.

To estimate the distillation loss and train the student model, we required the teacher's logits. So we imported the teacher model, specifically the Hugging Face bert-base-cased model pre-trained for each dataset. We obtained the teacher's logits from this imported model for the train split. Additionally, we combined the teacher's logits and the true label logits as a zipped pair and appended this to the dataset.

Upon importing the Hugging Face SST-2 dataset, we discovered that the test set did not include sentiment labels, deliberately set to -1. Considering the large size of the train set, we made the decision to repurpose the validation set as the test set. Subsequently, we split the original train

set into separate train and validation sets for our purposes.

4.2 Data augmentation

In order to facilitate effective knowledge transfer, it is essential to have a substantial amount of data. We determined that the datasets we utilized, including SST2, MRPC, MNLI, QQP, and RTE, did not possess the necessary size for optimal knowledge transfer. As a result, we employed the rule-based augmentation techniques listed below to generate extensive datasets of knowledge:

1. **Masking** - We masked random words in a training example, where each word has a probability p_{mask} of being masked with the tag [MASK].

2. **POS-guided word replacement** - We substituted words within each example with alternative words belonging to the same part-of-speech category. In this process, each word had a probability (p_{pos}) of being replaced.

3. **n-gram sampling** - The value of 'n' was assigned randomly within the range of 1 to 5. Each training example had a probability of p_{ng} for being sampled. From the selected training examples, a subset of size 'n' was chosen.

The procedure first started with iterating over words of each training example. For every word, we assigned a random number x from a uniform distribution. If $x < p_{mask}$, then we did Masking. Similarly, as POS-guided replacement augmentations is mutually exclusive of masking we applied POS-guided swapping on this word if $p_{mask} \leq x \leq p_{mask} + p_{pos}$.

Once again, for each training example, we made a decision using the probability p_{ng} on whether to conduct n-gram sampling. If the decision was affirmative, we performed the sampling and added the modified example to the augmented dataset. We repeated this entire process for a total of n_{iter} iterations for each training example.

For sentence-pair datasets, MNLI, QQP, MRPC, and RTE, we explored three different approaches: we augmented only the first sentence, we augmented only the second sentence, we augmented both sentences simultaneously.

We also experimented with randomly swapping 'n' words within a sentence. Our expectation was that this approach would enhance the model's capability to handle diverse expressions in input text, allowing it to recognize the same meaning conveyed through different phrasings. However, upon evaluation, we observed inconsistent results in each scenario. As a result we decided to exclude this augmentation method.

5 Baselines

We are using Hugging Face BERT$_{base}$ models and Hugging Face BERT$_{base}$ models fine-tuned on the corresponding datasets, as our baseline models. BERT$_{base}$ comprises of 12 layers, 768 hidden units, 12 self-attention heads, and 110M parameters.

We're also comparing our results with those of a BiLSTM with 300 hidden units and fully connected layers with 400 hidden units, primarily trained on just the dataset.

6 Your approach

We primarily based our implementation on the guidelines presented in the paper, ensuring that we followed the recommended approach. However, due to our computational limitations, we had to make certain design choices. We considered the availability of pre-trained models and adapted our implementation accordingly to utilize the models that were readily accessible to us. These adjustments were made to accommodate our specific computational resources and constraints.

6.1 Design Decisions

Fine-tuning large language models for downstream tasks can be a demanding and resource-intensive process. Fortunately, much of this work has already been done and is readily available through Huggingface. Therefore, we opted to utilize a BERT-based model that has already been fine-tuned on the specific dataset we are working with.

To streamline the training process and improve runtime efficiency, we introduced a forward pass step when creating copies of the datasets from Huggingface. This involved adding a new column, called "combined logits," to our datasets. This column contains both the output logits generated by the corresponding teacher model and the true label for each sample. By including this information upfront, we were able to significantly reduce

S.No.	Model	SST2 Acc/F1	MNLI Acc/F1	QQP Acc/F1	MRPC Acc/F1	RTE Acc/F1
1	BERT_{BASE}	93.5	86.7	89.3	88.9	71.1
2	BERT_{PRE}	91.4	82.8	90.8	83.1	67.1
3	$\text{BiLSTM}_{(baseline)}$	85.9	70.3	81.7	74.3	56.5
4	$\text{BiLSTM}_{(student)}$	83.6/84	59/59	81.1/73.8	70.5/80.9	51.6

Table 1: Test results by baseline and student models on different datasets. All of our test results are obtained from the GLUE benchmark website.

the runtime during training and fine-tuning of the student model.

In summary, we leveraged pre-existing fine-tuned models from Huggingface as our teacher model, and we optimized the dataset creation process by incorporating the teacher model's logits alongside the true labels in the "combined logits" column. This approach allowed us to expedite training and improve efficiency in our project.

Instead of training our own embeddings, we opted to utilize traditional Glove embeddings to represent our input data.

6.2 Architectures

In our implementation, we focused on two main architectures: BiLSTM for single-sentence datasets and a siamese BiLSTM for datasets containing pairs of sentences.

For the regular BiLSTM architecture, the embedding layer is an internal component of the model. However, the trainable attribute of the embedding parameters is set to false, indicating that the embeddings are not updated during training.

In the siamese BiLSTM architecture, we separate the embeddings layer from the main model. Before passing the sentence embeddings into the model, we apply additional operations on the embeddings of the two sentences. These operations include concatenation, element-wise multiplication, and addition. The resulting output from these operations is then passed into the model.

The model consists of two main layers. The first layer is a Bidirectional LSTM with 300 hidden units, which captures contextual information from the input sentences. This layer is connected to a fully connected layer with 400 hidden units. Finally, the output layer of the model has the number of units equal to the number of classes in the dataset.

Overall, our architectures accommodate different types of datasets and employ variations in the embedding layer and additional operations for the siamese BiLSTM architecture to enhance model performance.

6.3 Implementation Details

We utilized the TensorFlow Sequential API to construct our model, as it offers a straightforward and intuitive approach to defining deep learning models. With this API, we can easily stack layers on top of each other in a sequential manner. The majority of our code is written in TensorFlow, which provided us with a powerful framework for implementing our models.

During our experiments, we primarily conducted our training on Google Colab. This platform allowed us to leverage its computational resources and GPU support to train our models efficiently.

For word embeddings, we employed Glove embeddings of various dimensions depending on the dataset size, complexity, and the available RAM. To ensure uniform input lengths, all embeddings were padded to the same length before being passed to the BiLSTM layer.

Initially, we started with the hyperparameter values mentioned in the paper and for the model's number of units. These values yielded good results across most datasets. However, we observed that for the SST-2, MNLI and QQP datasets, using the Adam optimizer with a learning rate of 1e-3 provided better results compared to the Adadelta optimizer mentioned in the paper. Our implementation was primarily guided by the paper, which served as the main and sole reference for our work.

6.4 Issues Faced

For larger datasets such as MNLI and QQP, we encountered difficulties in computing the embeddings for the entire dataset and passing it to the training loop. To overcome this challenge, we implemented a generator that facilitated lazy load-

ing of smaller chunks of the dataset. This generator applied the necessary preprocessing function to each chunk before passing it to the training loop. However, we experienced issues with the forward and backward passes of the BiLSTM due to the multithreading nature of the generator and the internally set worker thread count. Debugging this problem proved challenging as the failure points were random, and the multithreading aspect made it difficult to pinpoint the exact cause of failure. We also attempted to use the Keras Sequence approach for lazy loading, but the issue persisted and could not be resolved.

To tackle the resource limitations in running MNLI and QQP (without augmentation) on Google Colab, we initially tried utilizing Google Colab Pro which provides increased RAM, but even that did not provide the desired solution. Eventually, we resorted to Unity Cluster, a platform that offers larger RAM capacity, enabling us to run the experiments effectively.

6.5 Experimental Results

We successfully trained our model on all of the datasets mentioned earlier, and we obtained the following results.

The graphs illustrating accuracy variation over different alpha values provide insights into how the accuracy of the validation data changes as we adjust the weighted average between cross-entropy loss with the true labels and mean squared error with respect to teacher logits.

Additionally, the learning curve graph gives an indication of how effectively the model learns over the course of training, showing the relationship between the number of epochs and the model's performance.

SST-2

F1 Score: 0.8402234636871508

Accuracy: 0.8360091743119266

Best $\alpha = 0.0$	Actual Labels	
Predicted Labels	0	1
0	353	75
1	68	376

Table 2: Confusion Matrix on SST2 test split

MRPC

F1 Score: 0.8089171974522292

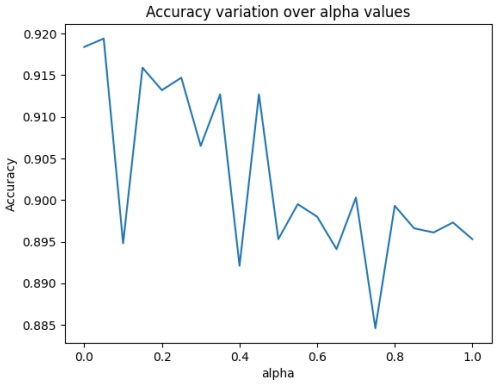

Figure 1: Plotting accuracy over α values on SST2 validation

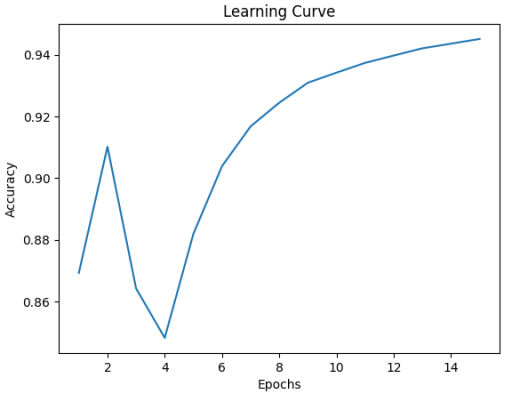

Figure 2: Plotting the learning curve for SST2 train

Accuracy: 0.7058823529411765

Best $\alpha = 0.2$	Actual Labels	
Predicted Labels	0	1
0	34	95
1	25	254

Table 3: Confusion Matrix on MRPC test split

RTE

F1 Score: 0.07142857142857142

Accuracy: 0.5163043478260869

MNLI

1. Matched Test Set:

F1 Score: 0.5907020441861486

Accuracy: 0.5895058583800306

2. Mismatched Test Split:

F1 Score: 0.5945964280376622

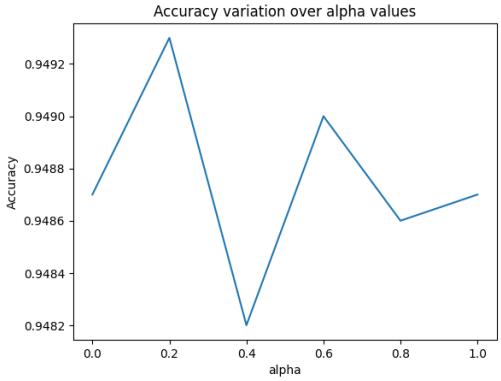

Figure 3: Plotting accuracy over α values on MRPC validation

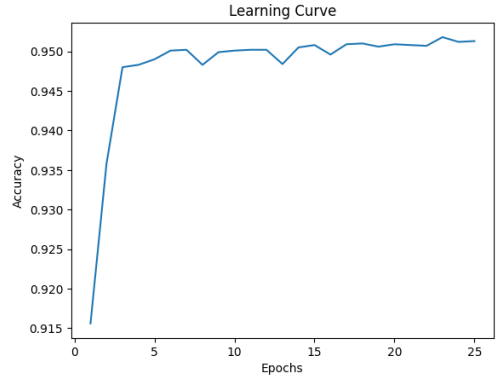

Figure 4: Plotting the learning curve for MRPC train

Accuracy: 0.5923515052888527

QQP

F1 Score: 0.7377872574652539

Accuracy: 0.8114766262676231

Based on the observations from the graphs, we can conclude that the model demonstrates rapid learning and stabilization within the first epoch for datasets like MRPC and RTE. However, for the SST-2 dataset, the model's performance continues to improve as the number of epochs increases. Due to computational limitations, we had to limit the number of epochs to 15 for MRPC and SST-2. Despite these constraints, the model shows promising progress and achieves satisfactory performance within the given epoch limits.

7 Error analysis

We performed manual error analysis by annotating 100 failed examples from each dataset based on their properties and have discussed each in detail below.

Best $\alpha = 0.0$	Actual Labels	
Predicted Labels	0	1
0	137	9
1	125	6

Table 4: Confusion Matrix on RTE test split

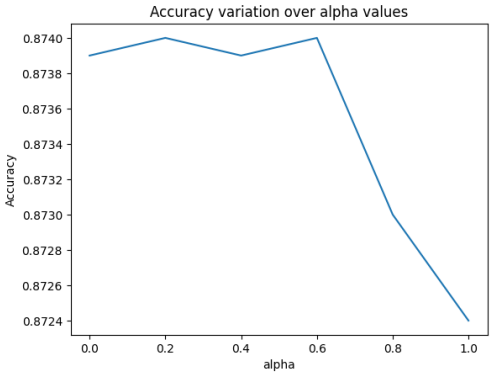

Figure 5: Plotting accuracy over α values on RTE validation

7.1 SST-2

In SST-2, we categorize movie reviews on whether the sentence is positive or negative.

We see from SST2's confusion matrix, that validation population generally lies within the positive diagonal of the confusion matrix, so it looks like the model is able to predict reasonably well.

As showcased in Table 8, we have two kinds of errors.

Negative connotation words: When the models incorrectly assigns a negative sentiment to a positive sentence, it usually because of the presence of word(s) that are heavily assigned a negative connotation. But in the english language a slight restructuring of the sentence would allow these negative words to impart positive meaning to the sentence. So the model is unable to identify these twists in context.

Positive connotation words: Like the error above, the inverse happens here. Positive words get twisted to give the sentence a negative tint, and so the model fails to correctly categorize the sentence.

7.2 MNLI

This dataset categorizes paired sentences on whether they both exhibit entailment, contradiction or no significant relationship.

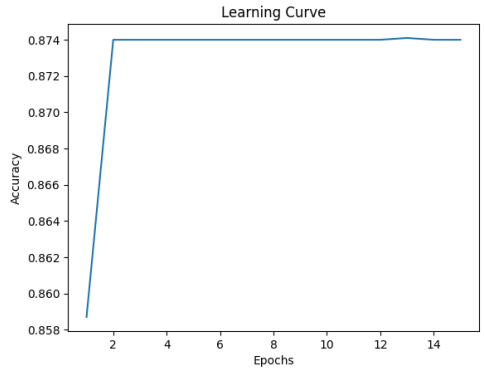

Figure 6: Plotting the learning curve for RTE train

Best $\alpha = 0.0$	Actual Labels		
Predicted Labels	0	1	2
0	1917	1209	353
1	611	2133	379
2	589	888	1736

Table 5: Confusion Matrix on MNLI matched test split

Matched version: From MNLI Matched's confusion matrix, it tends get nearly half of them wrong. But seeing as there 3 labels, the level up in complexity is expected.

Mismatched version: From MNLI Mismatched's confusion matrix, again get nearly half of them wrong. And like in matched, we find similar types of errors. As showcased in Table 9 and Table 10, we have six kinds of errors.

Unable to infer relation: One sentence contains the context of the other sentence, but uses different words and roundabout ways of conveying that meaning. This causes the model to fumble at identifying the crucial meaning that indicates the relation between the two.

Out of context names: One of the sentence is referring to same thing but under a different name, hence the model is unable to identify the relation.

Uncaught synonyms: Certain words or phrases that are similar in meaning are not identified as such by the model, so it erroneously indicates a wrong relation.

Best $\alpha = 0.0$	Actual Labels		
Predicted Labels	0	1	2
0	1919	1257	287
1	621	2152	356
2	540	947	1753

Table 6: Confusion Matrix on MNLI mismatched test split

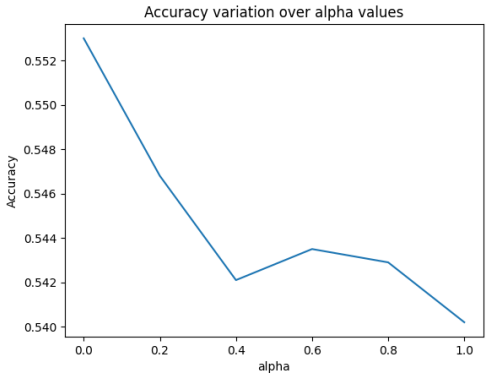

Figure 7: Plotting accuracy over α values on MNLI validation

Extraneous words: Although the sentences are similar, there are still words that are not common between the two and hence causes the model to identify a different relation.

Significant word overlap: Even though two sentences are not paraphrased, the existence of many common words and phrases causes the model to categorize them as nearly identical.

Annotation error: At times the original labelling is incorrect. Since the dataset is annotated by crowdfunding, it is expected that the quality of data depends on the person doing the annotations.

7.3 QQP

This dataset categorizes paired quora questions on whether they both are asking the same thing.

From QQP's confusion matrix, it seems to be performing decently.

We find similar types of errors as the other datasets, with a new addition, Synonym differences. As showcased in Table 11 , we have FIVE kinds of errors.

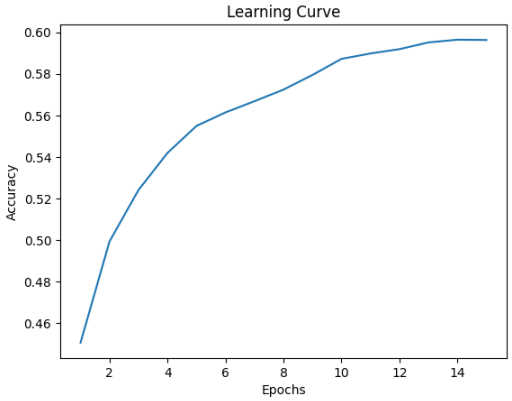

Figure 8: Plotting the learning curve for MNLI train

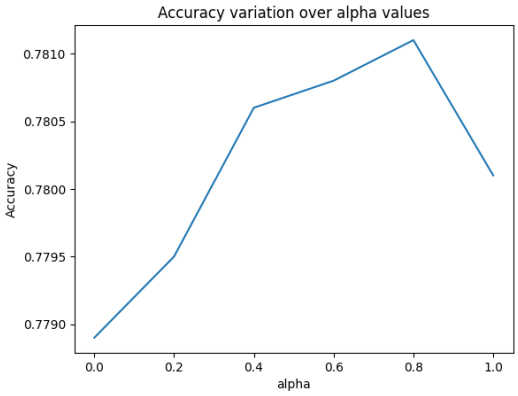

Figure 9: Plotting accuracy over α values on QQP validation

Best α = 0.8	Actual Labels	
Predicted Labels	0	1
0	22085	3460
1	4162	10723

Table 7: Confusion Matrix on QQP test split

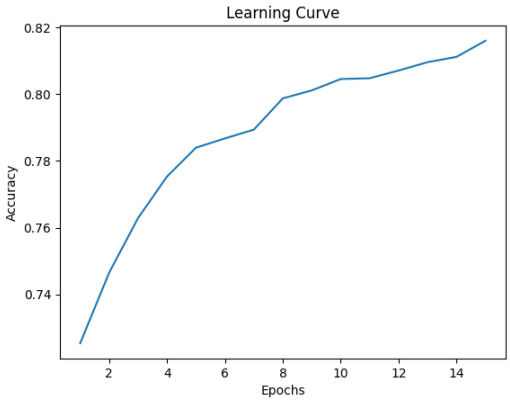

Figure 10: Plotting the learning curve for QQP train

Out of context names: One of the sentence is referring to same thing but under a different name, hence the model is unable to identify the relation.

Extraneous words: Although the sentences are similar, there are still words that are not common between the two and hence causes the model to identify a different relation.

Significant word overlap: Even though two sentences are not paraphrased, the existence of many common words and phrases causes the model to categorize them as nearly identical.

Annotation error: At times the original labelling is incorrect. Since the dataset is annotated by crowdfunding, it is expected that the quality of data depends on the person doing the annotations.

Synonym differences: Some questions have synonyms, but the slight context that differentiates the two words causes the two question to differ in meaning significantly. However, the model incorrectly attributes similar meaning to the questions on the basis of the synonyms.

7.4 MRPC

This dataset dealt with trying to find out if there was any paraphrasing between the two sentences. From MRPC's confusion matrix, we can see that our model finds it a bit more difficult to predict sentences with actual labels 0, meaning that the two sentences are not paraphrased.
Moreover, we are able to identify certain types of sentences that cause our model problems.
As showcased in Table 12, we have four kinds of errors.
Out of context names: One of the sentence is referring to same thing but under a different name, hence the model is unable to identify the similarity.

Uncaught synonyms: Certain words or phrases that are similar in meaning are not identified as such by the model, so it erroneously marks the pairing as not similar.

Extraneous words: Although the sentences

Table 8: SST2 Error Analysis

Sentences	Label	Error Type
we root for (clara and paul) , even like them , though perhaps it 's an emotion closer to pity .	1	Positive connotation words
if you 're hard up for raunchy college humor , this is your ticket right here .	1	Positive connotation words
or doing last year 's taxes with your ex-wife .	0	Negative connotation words
(w) hile long on amiable monkeys and worthy environmentalism , jane goodall 's wild chimpanzees is short on the thrills the oversize medium demands .	0	Negative connotation words

Table 9: MNLI Matched Error Analysis

Sentences	Label	Error Type
uh i don't know i i have mixed emotions about him uh sometimes i like him but at the same times i love to see somebody beat him	0	Annotation error
I like him for the most part, but would still enjoy seeing someone beat him.		
Calcutta seems to be the only other production center having any pretensions to artistic creativity at all, but ironically you're actually more likely to see the works of Satyajit Ray or Mrinal Sen shown in Europe or North America than in India itself.	1	Extraneous words
Most of Mrinal Sen's work can be found in European collections.		
Dare you rise to the occasion, like Raskolnikov, and reject the petty rules that govern lesser men?	1	Out of context names
Would you rise up and defeaat all evil lords in the town?		
The most important directions are simply up and up leads eventually to the cathedral and fortress commanding the hilltop, and down inevitably leads to one of three gates through the wall to the new town.	2	Unable to infer relation
Go downwards to one of the gates, all of which will lead you into the cathedral.		
Rouen is the ancient center of Normandy's thriving textile industry, and the place of Joan of Arc's martyrdom ' a national symbol of resistance to tyranny.	0	Uncaught synonyms
Joan of Arc sacrificed her life at Rouen, which became an enduring symbol of opposition to tyranny.		
While it's probably true that democracies are unlikely to go to war unless they're attacked, sometimes they are the first to take the offensive.	0	Significant word overlap
Democracies probably won't go to war unless someone attacks them on their soil		

Table 10: MNLI Mismatched Error Analysis

Sentences	Label	Error Type
The bankruptcy of secular, autocratic nationalism was evident across the Muslim world by the late 1970s.	1	Unable to infer relation
Muslims disliked autocratic nationalism by the late 1970s.		
That means we now have the opportunity to be a stable, positive and important part of each child's life for an entire decade.	1	Uncaught synonyms
Providing stability and positivity for each child has been made possible from continued support.		
The forecasting challenges retailers confront have been amplified in recent years by product proliferation in almost every category.	2	Annotation error
Forecasting has been easier recently due to the updated process we have today.		

Table 11: QQP Error Analysis

Sentences	Label	Error Type
'I am 25 year old guy and never had a girlfriend. Is this weird?	1	Annotation error
I am 25 years old. I have never had a girlfriend. Is something wrong with me?		
Why was the Roman Empire so successful?	0	Annotation error
What are some of the rarely known facts about the Roman Empire?		
Why is my life getting so complicated?	0	Extraneous words
Why is my life so complicated?		
Can you TRANSLATE these to English language?	0	Out of context names
Can you translate this from Bengali to English language?		
What was the deadliest battle in history?	1	Synonym differences
What was the bloodiest battle in history?		
Is it a bad time to buy a condo or a house in the Bay Area in 2017?	1	Significant word overlap
Would 2017 be a good time to buy a house in Bay Area?		

are similar, there are still words that are not common between the two and hence causes the model to identify them as different.

Significant word overlap: Even though two sentences are not paraphrased, the existence of many common words and phrases causes the model to categorize them as nearly identical.

7.5 RTE

For RTE, we have to check if the second sentence is entailed within the first sentence. In RTE's confusion matrix, there aren't any 0 labels in this validation, due to the data being very skewed, so we only deal with errors where the model incorrectly assumes that no entailment occurs when in actuality there is entailment.
As showcased in Table 13, we have three kinds of errors.

Unable to infer relation: The first sentence contains the context of second sentence, but uses different words and roundabout ways of conveying that meaning. This causes the model to fumble at identifying the entailment.

Uncaught synonyms: Certain words or phrases that are similar in meaning are not identified as such by the model, so it is unable find the second sentence within the first.

Extraneous words: Although the sentences are similar, there are still words that are not common between the two and hence causes the model to identify them as different.

8 Contributions of group members

List what each member of the group contributed to this project here. For example:

- Vayunandhan Reddy: Dataset processing, built and trained models
- Muni Lohith: Dataset processing, built and trained models
- Maanusri : Dataset processing, error analysis and annotations
- Shilpa Sweth : PKD-Last implementation proof of concept and error analysis

9 Conclusion

We were amazed to observe that smaller models, like the ones we are using, can achieve similar performance compared to larger models such as BERT Base that are specifically fine-tuned for the task. This was achieved by augmenting the training dataset, even though there is a substantial difference in the number of parameters between these models. This led us to believe that improving the performance of language models doesn't necessarily require scaling them to such high parameter counts. Instead, focusing on enhancing the training objective and methodologies can also lead to significant improvements.

Furthermore, during our project, we extensively explored various libraries available for implementing, training, and fine-tuning these models. This allowed us to gain a comprehensive understanding of the different tools and resources at our disposal, enabling us to make informed decisions and optimize our approach.

10 AI Disclosure

- Did you use any AI assistance to complete this proposal? If so, please also specify what AI you used.
  - Yes

If you answered yes to the above question, please complete the following as well:

- If you used a large language model to assist you, please paste *all* of the prompts that you used below. Add a separate bullet for each prompt, and specify which part of the proposal is associated with which prompt.
  - Used prompts to get bibtex and format the proposal etc
  - Used ChatGPT to parphase written context
- **Free response:** For each section or paragraph for which you used assistance, describe your overall experience with the AI. How helpful was it? Did it just directly give you a good output, or did you have to edit it? Was its output ever obviously wrong or irrelevant? Did you use it to generate new text, check your own ideas, or rewrite text?
  - your response here

Table 12: MRPC Error Analysis

Sentences	Label	Error Type
Magnarelli said Racicot hated the Iraqi regime and looked forward to using his long years of training in the war .	0	Out of context names
His wife said he was " 100 percent behind George Bush " and looked forward to using his years of training in the war .		
Cooley said he expects Muhammad will similarly be called as a witness at a pretrial hearing for Malvo .	0	Out of context names
Lee Boyd Malvo will be called as a witness Wednesday in a pretrial hearing for fellow sniper suspect John Allen Muhammad .		
The pound also made progress against the dollar , reached fresh three-year highs at $ 1.6789 .	0	Uncaught synonyms
The British pound flexed its muscle against the dollar , last up 1 percent at $ 1.6672 .		
No dates have been set for the civil or the criminal trial .	0	Uncaught synonyms
No dates have been set for the criminal or civil cases , but Shanley has pleaded not guilty .		
" Sanitation is poor ... there could be typhoid and cholera , " he said .	0	Extraneous words
" Sanitation is poor , drinking water is generally left behind . . . there could be typhoid and cholera . "		
The driver , Eugene Rogers , helped to remove children from the bus , Wood said .	0	Extraneous words
At the accident scene , the driver was " covered in blood " but helped to remove children , Wood said .		
At community colleges , tuition will jump to $2,800 from$ 2,500 .	1	Significant word overlap
Community college students will see their tuition rise by $300 to$ 2,800 or 12 percent .		
HP 's shipments increased 48 percent year-over-year , compared to an increase of 31 percent for Dell .	1	Significant word overlap
HPs shipments increased 48 per cent year-on-year , compared to an increase of 31 per cent for Dell .		

Table 13: RTE Error Analysis

Sentences	Label	Error Type
Yet, we now are discovering that antibiotics are losing their effectiveness against illness. Disease-causing bacteria are mutating faster than we can come up with new antibiotics to fight the new variations.	0	Unable to infer relation
Bacteria is winning the war against antibiotics.		
Security forces were on high alert after an election campaign in which more than 1,000 people, including seven election candidates, have been killed.	0	Unable to infer relation
Security forces were on high alert after a campaign marred by violence.		
As spacecraft commander for Apollo XI, the first manned lunar landing mission, Armstrong was the first man to walk on the Moon. "Thatś one small step for a man, one giant leap for mankind." With these historic words, manś dream of the ages was fulfilled.	0	Uncaught synonyms
Neil Armstrong was the first man who landed on the Moon.		
Russian cosmonaut Valery Polyakov set the record for the longest continuous amount of time spent in space, a staggering 438 days, between 1994 and 1995.	0	Uncaught synonyms
Russians hold record for longest stay in space.		
Charles de Gaulle died in 1970 at the age of eighty. He was thus fifty years old when, as an unknown officer recently promoted to the (temporary) rank of brigadier general, he made his famous broadcast from London rejecting the capitulation of France to the Nazis after the debacle of May-June 1940.	0	Extraneous words
Charles de Gaulle died in 1970.		
Pibul Songgram was the pro-Japanese military dictator of Thailand during World War 2.	0	Extraneous words
Pibul was the dictator of Thailand.		

References

Chen, P., Liu, S., Zhao, H., and Jia, J. (2021). Distilling knowledge via knowledge review.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). A simple framework for contrastive learning of visual representations.

Dolan, B., Quirk, C., and Brockett, C. (2004). Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, pages 350–356. Association for Computational Linguistics.

Heo, B., Kim, J., Yun, S., Park, H., Kwak, N., and Choi, J. Y. (2019). A comprehensive overhaul of feature distillation.

Iyer, S. and et al. (2017). First quora dataset release: Question pairs. *Quora*.

Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., and Liu, Q. (2019). Tinybert: Distilling bert for natural language understanding.

Park, W., Kim, D., Lu, Y., and Cho, M. (2019). Relational knowledge distillation.

Romero, A., Ballas, N., Kahou, S. E., Chassang, A., Gatta, C., and Bengio, Y. (2014). Fitnets: Hints for thin deep nets.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Sun, S., Cheng, Y., Gan, Z., and Liu, J. (2019). Patient knowledge distillation for bert model compression.

Tang, R., Lu, Y., Liu, L., Mou, L., Vechtomova, O., and Lin, J. (2019). Distilling task-specific knowledge from BERT into simple neural networks.

Tian, Y., Krishnan, D., and Isola, P. (2019). Contrastive representation distillation.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. (2018). Glue: A multi-task benchmark and analysis platform for natural language understanding.

Williams, A., Nangia, N., and Bowman, S. (2018). A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Yang, J., Martinez, B., Bulat, A., and Tzimiropoulos, G. (2021). Knowledge distillation via softmax regression representation learning. In *International Conference on Learning Representations*.

Yang, Z., Li, Z., Shao, M., Shi, D., Yuan, Z., and Yuan, C. (2022). Masked generative distillation.

Zhao, B., Cui, Q., Song, R., Qiu, Y., and Liang, J. (2022). Decoupled knowledge distillation.

Zhou, H., Song, L., Chen, J., Zhou, Y., Wang, G., Yuan, J., and Zhang, Q. (2021). Rethinking soft labels for knowledge distillation: A bias-variance tradeoff perspective.