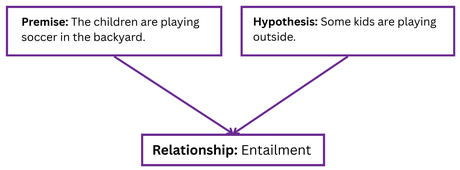
COMP34812 Natural Language Understanding Coursework: State-of-the-Art Approaches to Solving a Natural Language Inference Task

Group 15: Hala Alsaffarini (t94363ha) and Kareem Seifo (g64462ks)

01 Introduction

A Natural Language Inference (NLI) task is one where a relationship needs to be determined between two pieces of text, a premise and a hypothesis. There exist 3 relationships: entailment, contradiction, and neutral. For this NLI task, we are only concerned whether the hypothesis entails the premise or not. An example is shown below:



Here the relationship is entailment since the hypothesis can be inferred from the premise.

02 Dataset

The dataset provided by the COMP34812 Team consists of 24,432 training pairs and 6,737 validation pairs. Each example contains a premise, a hypothesis, and a binary prediction label indicating whether the premise entails the hypothesis. A sample from the dataset is shown below to illustrate its structure.

premise	hypothesis	label
I know that many of you are interested in addressing these issues through legislation.	The problems must be addressed.	1
The mountains shield us and our swords are the torrents.	The mountains offer no protection.	0

03 Methodology Solution B

Category B: Attention Decomposition + Pre-trained Transformer Embedding Layer

This model is based upon the attention decomposition model described in [1].

The attention decomposition pipeline can be described in three steps:

1. Attend: The model uses neural attention to create a soft alignment between subphrases of the two input sentences.
2. Compare: Each aligned subphrase is then separately compared using feed-forward networks to generate comparison vectors.
3. Aggregate: The comparison vectors are aggregated by summation and fed into a final classifier to predict the relationship between the sentences.

The model we submitted for category B appends an additional layer to the attention decomposition model, a pre-trained transformer embedding layer. Our model leverages transformer based embeddings from DeBERTa-v3, a transformer based model.

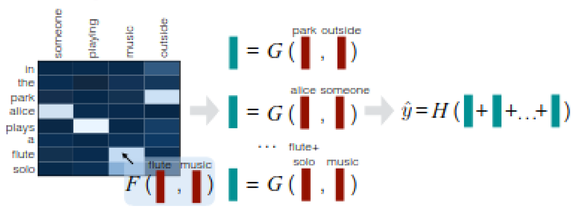


Figure 1: Showing Attend (left), Compare (center), and aggregate (right) steps [1].

04 Methodology Solution C

Category C: ULMFiT-Inspired Fine-Tuning of RoBERTa with Stochastic Weight Averaging (SWA)

We adopt RoBERTa [2], a transformer-based language model pretrained on large-scale corpora using a masked language modeling objective. Its architecture, shown in Figure 2, serves as the foundation for our downstream ULMFiT-inspired fine-tuning pipeline, providing rich contextualized representations for textual inputs.

Our fine-tuning strategy is inspired by the ULMFiT framework [3] for text classification, with the addition of Stochastic Weight Averaging (SWA) to improve generalization [4]. The process consists of three phases, as illustrated in Figure 3.

- **Phase 1 Language Model Pretraining**: — This step is implicitly handled by leveraging the pretrained RoBERTa large model, which has already been trained on large scale corpora.
- **Phase 2 Gradual Unfreezing**: We progressively unfreeze the 24 transformer layers of RoBERTa from top to bottom. During this phase, we apply discriminative fine-tuning by using adaptive learning rates per layer, decaying the rate as we unfreeze deeper layers.
- **Phase 3 Classifier Fine-Tuning with SWA**: After all layers are unfrozen, we further fine-tune the entire model with a focus on the classification head. To enhance stability and performance, we integrate Stochastic Weight Averaging (SWA), which averages model weights across multiple training steps to converge to flatter minima and reduce generalization error.

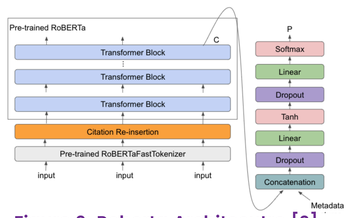


Figure 2: Roberta Architecutre [2]

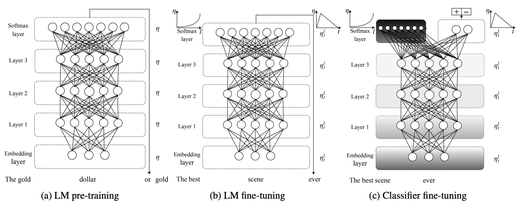
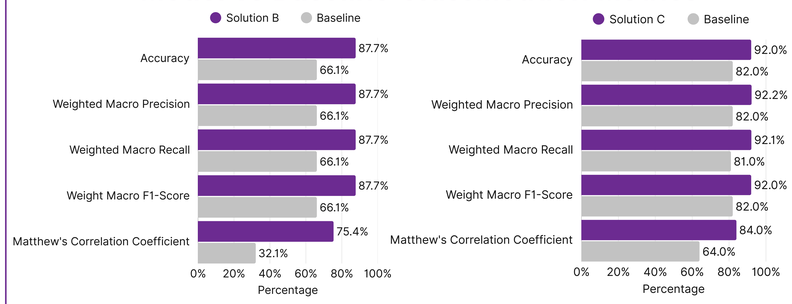


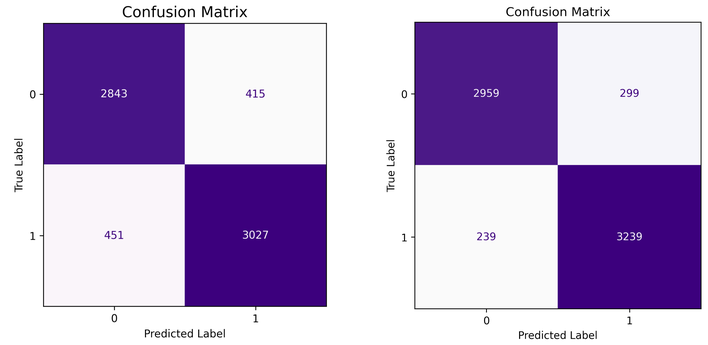
Figure 3: ULMFiT Fine-tuning Phases [3]

05 Results

Model vs Baseline Classification Metrics



Confusion Matrix Analysis



Solution B **Solution C**

06 Conclusions

- The experiments demonstrate the extent to which these solutions can be used to solve an NLI task
- Solution B shows a higher number of misclassifications overall (415 false positives, 451 false negatives) compared to Solution C (399 false positives, 239 false negatives)
- Solution C has more correct predictions for both classes (2959 + 3239) than Solution B (2843 + 3027)
- On the development set, the Decomposition model outperformed the baseline by 21%, while the Transformer-based solution showed an average improvement of 10% across all metrics.
- Fine-tuning the models on a bigger and more complex datasets can accomodate precise user needs

References:

[1]: Parikh, Ankur P., et al. "A Decomposable Attention Model for Natural Language Inference." ArXiv:1606.01933 [Cs], 25 Sept. 2016, arxiv.org/abs/1606.01933.

[2]:Huang, Zihan, et al. "Context-Aware Legal Citation Recommendation Using Deep Learning." Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law, 21 June 2021, arxiv.org/pdf/2106.10776.pdf, https://doi.org/10.1145/3462757.3466066. Accessed 29 Mar. 2022.

[3]: Howard, Jeremy, and Sebastian Ruder. "Universal Language Model Fine-Tuning for Text Classification." ArXiv.org, 2018, arxiv.org/abs/1801.06146.

[4]: Izmailov, Pavel, et al. "Averaging Weights Leads to Wider Optima and Better Generalization." ArXiv:1803.05407 [Cs, Stat], 25 Feb. 2019, arxiv.org/abs/1803.05407. Accessed 13 Apr. 2023.