NLI Solutions: Ensemble and Bi-GRU Models

GROUP28 - Renhao Bao & Shiqi Yang

Abstract

This study contrasts two approaches for Natural Language Inference (NLI): an Ensemble Machine Learning Model and a Bi-GRU Deep Learning Model with attention. Performance metrics show the Deep Learning Model slightly outperforms the ensemble, suggesting attention mechanisms paired with Bi-GRUs could be more effective for NLI tasks.

Introduction

- Natural Language Inference (NLI) is a critical task in natural language processing where the goal is to determine if a hypothesis is true or false based on a given premise.
- Our aim is to evaluate and compare the ensemble machine learning model and the neural network model featuring attention and Bi-GRU in NLI tasks.
- Our findings indicates the neural approach works better than stacking multiple traditional models, hinting at the future directions for improving NLI systems.

Methodology

Models and Architecture:

- **Ensemble Machine Learning Model:**
  - **Base Learners:** Random Forest, SVM (Support Vector Machine), and Logistic Regression, integrated using a StackingClassifier framework.
  - **Final Estimator:** Gradient Boosting Classifier.
- **Deep Learning Model:**
  - **Architecture:** The neural model integrates a bidirectional GRU (Bi-GRU) with an attention mechanism to weigh different parts of the text sequence. Layer normalization and dropout are employed to stabilize and regularize the network.

Training Procedure:

- **Embeddings:** Utilizes Sentence-BERT model ('all-MiniLM-L6-v2') to generate embeddings for both premises and hypotheses.
- **Optimization:** Employed Adam optimizer with a learning rate reduction on plateau.

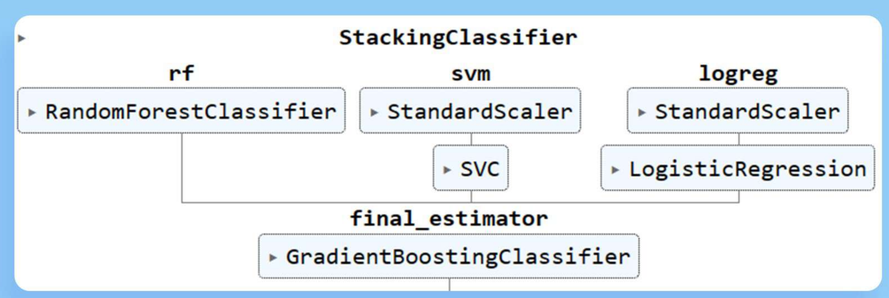

Figure 1 – A Diagram of the Ensembled Model Architecture

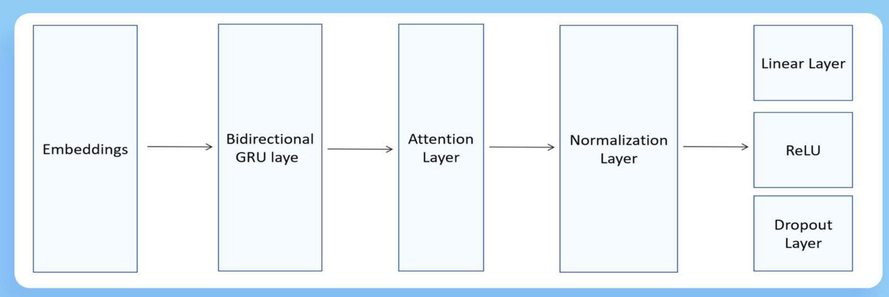

Figure 2 - The Deep Learning Model Architecture

Table 1: Validation Metrics on Deep Learning Model

Class	Precision	Recall	F1-Score
0	0.73	0.64	0.68
1	0.70	0.77	0.73
Overall Accuracy: 0.71			
Macro Avg	0.71	0.71	0.71
Weighted Avg	0.71	0.71	0.71

Result

Our evaluation showed the Deep Learning Model achieved a 71% accuracy, slightly outperforming the Ensemble Model's 69%. In comparison to the typical machine learning stack (Table 2), the attention-augmented Bi-GRU performed better, as evidenced by the precision and recall metrics (Table 1).

Table 2: Validation Metrics on Ensemble Model

Class	Precision	Recall	F1-Score
0	0.71	0.63	0.66
1	0.68	0.76	0.72
Overall Accuracy: 0.69			
Macro Avg	0.70	0.69	0.69
Weighted Avg	0.69	0.69	0.69

Conclusion

The comparative analysis showed that the Deep Learning Model with Bi-GRU and attention mechanisms surpasses the Ensemble Machine Learning Model in NLI tasks, indicating that deep learning could enhance language understanding. These results suggest that developing hybrid models that merge ensemble robustness with neural network sophistication would benefit NLI systems.

Reference

- Reimers, Nils, and Iryna Gurevych. "Sentence-bert: Sentence embeddings using siamese bert-networks." arXiv preprint arXiv:1908.10084 (2019).
- Chen, Qian, et al. "Enhanced LSTM for natural language inference." arXiv preprint arXiv:1609.06038 (2016).
- Dietterich, Thomas G. "Ensemble methods in machine learning." International workshop on multiple classifier systems. Berlin, Heidelberg: Springer Berlin Heidelberg, 2000.