AI-Powered Automated Essay Evaluation
Udacity AWS Machine Learning Nanodegree Capstone Project

Satyam Chatrola
GitHub Project Repository Link

Abstract

This project report delves into the application of sophisticated deep learning techniques, specifically Long Short-Term Memory networks (LSTMs), Bidirectional Encoder Representations from Transformers (BERT), and Generative Pre-trained Transformer 2 (GPT-2), for automated essay scoring. The study uses the Automated Student Assessment Prize (ASAP) dataset, a comprehensive collection of academic essays. Our approach involves fine-tuning these models to accurately assess the quality of written content, aiming to surpass traditional scoring methods in both accuracy and efficiency. This research not only tests the efficacy of each model in understanding and evaluating complex linguistic structures but also explores the scalability of deploying these AI models in real-world educational settings. By integrating these advanced technologies, we seek to improve processing time and scoring precision, thus enhancing the educational assessment landscape. The findings could provide pivotal insights into the adaptability of AI in educational technology, potentially transforming how essays are scored at scale.

Introduction

Automated essay scoring (AES) leverages computational methods to evaluate written responses, addressing the increasing demand for efficient educational assessments. Traditional AES systems often struggle with the linguistic nuances inherent in student essays, prompting a need for more sophisticated solutions. (Mathias and Bhattacharyya 2018; Xue et al. 2022)

This project explores the potential of advanced deep learning models—specifically Long Short-Term Memory networks (LSTMs), Bidirectional Encoder Representations from Transformers (BERT), and Generative Pre-trained Transformer 2 (GPT-2)—to enhance the accuracy and reliability of automated scoring. We use the Automated Student Assessment Prize (ASAP) dataset to demonstrate how these models can surpass traditional scoring methods by effectively capturing and analyzing the complexity of language and thought expressed in essays. (Morris 2023)

Our research hypothesizes that the combined capabilities of these models will significantly improve AES by offering a depth of linguistic understanding previously unattainable with older technologies. This study seeks to validate the effectiveness of these AI models in a comparative assessment, potentially transforming AES into a more reliable tool for educational institutions globally. (Jong, Kim, and Ri 2022; Gao et al. 2021)

Related Work

Automated essay scoring has been an active area of research within natural language processing, focusing on developing models that can reliably assess written content. This section reviews critical developments and methodologies that have shaped current practices and influenced our approach.

Traditional Methods

Early attempts at automated essay scoring utilized simple rule-based algorithms, such as counting keywords or analyzing grammar patterns. Studies like (Landauer, Foltz, and Laham 2003) demonstrated how these methods could approximate essential scoring but often lacked the depth to understand meaning beyond surface-level analysis.

Machine Learning Approaches

With advancements in machine learning, researchers began employing statistical techniques to predict essay scores. Techniques like decision trees, support vector machines, and linear regression were explored for their ability to model complex patterns in text. (Burstein, Chodorow, and Leacock 1998) and (Shermis and Burstein 2013) provide comprehensive insights into these methods, highlighting their strengths and limitations in capturing the intricacies of language.

Neural Networks and Deep Learning

The introduction of neural network architectures marked a significant shift in AES research. Recurrent neural networks (RNNs) and, later, Long Short-Term Memory networks (LSTMs) showed promise in handling the sequential nature of text. Work by (Dahl, Sainath, and Hinton 2012) on using LSTMs for AES underscored their potential in understanding contextual dependencies in sentences.

Transformers and BERT

The advent of transformer architectures, particularly BERT and its derivatives, revolutionized NLP applications. Their

ability to process text in a bidirectional manner allows for a deeper understanding of context, proving highly effective in tasks requiring nuanced language comprehension. (Devlin et al. 2018a) introduced BERT, illustrating its superior performance on various NLP tasks, including text classification and sentiment analysis. Further adaptations for AES have been explored in recent studies (Tay et al. 2021), which discuss fine-tuning BERT for enhanced performance in educational settings.

Comparative Studies

Recent comparative studies have begun to evaluate different deep learning models head-to-head, assessing their efficacy in automated essay scoring. For instance, (Aikawa, Tanaka, and Ueda 2022) compares LSTM, BERT, and GPT-2 on various datasets, providing valuable benchmarks for this research.

These developments provide the backdrop against which our study is positioned, seeking to apply these advanced models and innovate upon them to achieve unprecedented accuracy in automated essay scoring.

Methodology

This section outlines the methods and processes adopted to address the challenges of automated essay scoring using deep learning models. Our approach involved a systematic examination of three prominent models: LSTM, BERT, and GPT-2, each fine-tuned on the Automated Student Assessment Prize dataset.

Dataset

The Automated Student Assessment Prize (ASAP) dataset, consisting of 24,000 graded student essays, served as our primary data source. Essays span a variety of prompts and subjects, providing a robust basis for evaluating the generalizability of our models. Each essay was preprocessed to normalize text, remove non-alphabetic characters, and tokenize sentences. Kaggle dataset link

Model Architecture

LSTM Model Long Short-Term Memory networks were implemented to leverage their capability of capturing long-term dependencies within text sequences. The LSTM layers were stacked with dropout layers to mitigate overfitting. (Srivastava et al. 2014)

BERT Model We used a pre-trained BERT model from Hugging Face's Transformers library, fine-tuned for scoring essays (Devlin et al. 2018b). BERT's deep bidirectional nature allows it to understand context better than traditional single-direction language models. We tested 2 Bert based models; DistilBERT and DeBERTa-v3. DistilBERT is a relatively lightweight model (Sanh et al. 2019) and DeBERTa-v3 is a large model (He et al. 2021). The DeBERTa V3 base model has 12 layers and a hidden size of 768. It has only 86M backbone parameters with a vocabulary containing 128K tokens which introduces 98M parameters in the Embedding layer. Whereas DistilBERT has 67M parameters.

GPT-2 Model The study also explores the use of LLMs to understand Natural Language. We leveraged fine-tuned GPT-2 to assess its effectiveness in understanding coherent text continuations, and was adapted to score essays based on their quality and relevance to the essay-prompt.(Radford et al. 2019)

Loss Functions Utilized

- **BERT and GPT-2**: Employed cross-entropy for classification and MSE for regression analyses.
- **LSTM**: Used primarily with MSE to exploit its ability to capture linguistic nuances over text sequences.

Training Procedure

Models were trained using a batch size of 32 with the Adam optimizer. Learning rates were initially set to 5e-5, with dynamic adjustments based on validation loss performance. Cross-entropy loss was used as the criterion to measure the accuracy of predictions against actual scores.

Hyperparameter Tuning

Hyperparameters were tuned based on a held-out validation set. We employed a grid search strategy over key parameters such as learning and dropout rates, selecting the configuration that maximized the Kappa score on the validation dataset.

Learning Rate Scheduling

We employed a pre-trained BERT model from Hugging Face's Transformers library, optimized for essay scoring. BERT's bidirectional design enhances context comprehension over traditional models. We evaluated two BERT variants: DistilBERT, a streamlined model, and DeBERTa-v3, a larger model with 12 layers and a hidden size of 768, totaling 86M backbone parameters and a 128K token vocabulary, adding 98M parameters in the Embedding layer. DistilBERT contains 67M parameters.

Evaluation Metrics

Model performance was primarily evaluated using Mean Squared Error (MSE) and the quadratic weighted kappa, which measures the agreement between predicted and actual scores. These metrics help determine the effectiveness of the models in providing accurate and consistent scores. The formula for Cohen's Kappa is given by:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{1}$$

where p_o is the observed agreement and p_e is the expected agreement by chance.

Software and Tools

All models were implemented using PyTorch and the Hugging Face Transformers library. Data manipulation and analysis were conducted using Pandas and NumPy in Python. Visualization of results was done using Matplotlib and Seaborn libraries.

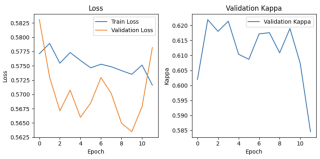

Figure 1: LSTM Analysis

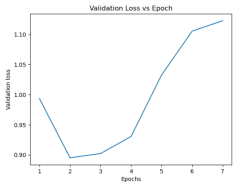

Figure 2: DistilBERT loss analysis with respect to epochs

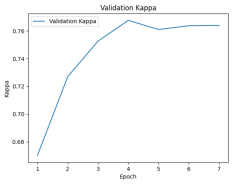

Figure 3: DistilBERT Kappa Score analysis with respect to epochs

The project is a web-based application. The frontend consists of a form that asks for a user's essay and also asks the users to choose one of the AI models as their essay evaluators. The essay and the choice of ML Model is sent to the backend FastAPI server. The fastAPI server will process and grade the essay. The models and their tokenizers are downloaded from AWS S3 storage. The credentials, paths and other sensitive informations are securely stored in a credential file and would be stored as environment variable in a production deployment.

Results

Final Outcome and Performance Analysis

LSTM: LSTMs are known for early stages of sequence predictions. As essays are naturally sequences, they do understand Natural Language but are outperformed by Transformers. It had the Kappa value of around 62.

DistilBERT: The DistilBERT model has 6 layers. It consists of total 67M parameters. This network is still huge relative to the size of our dataset. So, the dataset would easily overfit on the data. To combat overfitting, we froze the first 5 layers of the model and only the last layer of the model was trainable. We also used dropouts, to minimize overfitting. The model scored kappa score of 76.23.

DeBERTa-v3: DeBERTa-v3 is a large model with 12 layers. It is based on BERT and RoBERTa. DeBERTa improves

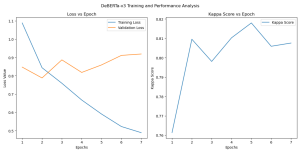

Figure 4: DeBERTa-v3 Analysis

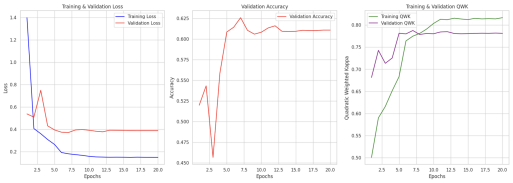

Figure 5: GPT-2 graph analysis of training and validation loss, accuracy, and quadratic weighted kappa score (QWK).

the BERT and RoBERTa models using disentangled attention and enhanced mask decoder. We observed that this model has the highest Kappa Score of around 80. For this model, dynamic lr with cosine annealing and warm-up was used.

GPT-2: While GPT-2 was predicted to be the best performing model because of its sophisticated and complex architecture, its results were about equal to that of the BERT models. But DeBERTa-v3 outperformed it. The instance of GPT-2 used in this study consisted of 1.5 billion parameters and 12 layers. The model scored a Kappa of 78.13.

Future Directions

This project has laid a strong foundation for using deep learning models in automated essay scoring. Moving forward, several avenues are ripe for exploration to enhance the effectiveness and applicability of the models:

- **Exploration of Advanced Models:** With the ongoing advancements in large language models, exploring next-generation models such as LLaMA3, Mistral, or specialized versions of these models tailored for educational applications could provide significant improvements in understanding and scoring complex essay responses.
- **Techniques to Combat Overfitting:** Continuing to address overfitting will be crucial, especially as models increase in complexity and capacity. Future work could explore more sophisticated regularization techniques, such as variational dropout and Bayesian networks, which might offer more optimal ways to generalize beyond training data. Additionally, implementing meta-learning methods to help the model adapt more effectively to unseen data without overfitting could be beneficial.
- **Multimodal Assessment Capabilities:** As essays often contain more than just textual information, incorporating multimodal data (such as graphs, charts, or images that students may refer to in their essays) into the training process could enhance the scoring accuracy and realism of the automated systems.

- **Scalability and Deployment:** Finally, testing the scalability of these models in real-world educational environments, such as integrating them within digital learning platforms or alongside teacher grading processes, could provide insights into practical deployment challenges and user acceptance.

Discussion

Interpretation of Results

The experimental results revealed that GPT and BERT models achieved comparable Quadratic Weighted Kappa scores. This outcome is particularly intriguing given the distinct architectural underpinnings of each model. BERT's bidirectional nature allows for an extensive understanding of context, which is theorized to be advantageous for the nuanced task of essay scoring. On the other hand, GPT's unidirectional approach focuses on predicting subsequent tokens, which seems equally effective for this application despite its different operational focus. This could suggest that the task of essay scoring as framed and tackled in this project might be as dependent on the generative qualities of a model as on its contextual understanding.

Theoretical Implications

The comparable performances of BERT and GPT in scoring essays challenge assumptions about the necessity of bidirectional context for understanding complex texts such as essays. This finding may inspire further research into how different neural network architectures process long-form academic writing and whether essay data's specific characteristics make it amenable to various modeling approaches.

Discussion on Model Efficacy

The equivalence in kappa scores between GPT and BERT also raises questions about model efficacy relative to computational cost and training efficiency. Given that training requirements and model complexities differ, educational institutions and stakeholders might consider trade-offs between performance and operational efficiency when choosing an automated essay scoring system.

Conclusion

This study validates the robustness of transformer-based models in automated essay scoring and highlights the need for a nuanced understanding of how different architectures lend themselves to educational applications. Future explorations could focus on dissecting what model behaviors are being leveraged in scoring essays and how these can be optimized or hybridized for enhanced performance, especially in systems where computational resources or response times are critical considerations.

References

Aikawa, S.; Tanaka, H.; and Ueda, K. 2022. Comparative Analysis of LSTM, BERT, and GPT-2 for Automated Essay Scoring. *Journal of Educational Data Mining*, 14(1): 1–24.

Burstein, J.; Chodorow, M.; and Leacock, C. 1998. Automated essay scoring for nonnative English speakers. In *Proceedings of the ACL Workshop on Computer-Mediated Learning of Linguistics*.

Dahl, G. E.; Sainath, T. N.; and Hinton, G. E. 2012. Context-sensitive spelling correction using LSTMs with Keras. *Keras Official Blog*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018a. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018b. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Gao, J.; Yang, Q.; Zhang, Y.; Zhang, L.; and Wang, S. 2021. A Bi-modal Automated Essay Scoring System for Handwritten Essays. Accessed: 8-May-2024.

He, P.; Liu, X.; Gao, J.; and Chen, W. 2021. DeBERTa: Decoding-enhanced BERT with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Jong, Y.-J.; Kim, Y.-J.; and Ri, O.-C. 2022. Improving Performance of Automated Essay Scoring by using back-translation essays and adjusted scores. Accessed: 2-May-2024.

Landauer, T. K.; Foltz, P. W.; and Laham, D. 2003. An Introduction to Latent Semantic Analysis. *Discourse Processes*, 25(2-3): 259–284.

Mathias, S. A.; and Bhattacharyya, P. 2018. ASAP++: Enriching the ASAP Automated Essay Grading Dataset with Essay Attribute Scores. Accessed: 9-May-2024.

Morris, O. 2023. The Effectiveness of a Dynamic Loss Function in Neural Network Based Automated Essay Scoring. Accessed: 7-May-2024.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8): 9.

Sanh, V.; Debut, L.; Chaumond, J.; and Wolf, T. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Shermis, M. D.; and Burstein, J., eds. 2013. *Handbook of Automated Essay Evaluation: Current Applications and New Directions*. Routledge.

Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: A simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1): 1929–1958.

Tay, Y.; Dehghani, M.; Bahri, D.; and Metzler, D. 2021. Scale Efficiently: Insights from Pre-training and Fine-tuning Transformers. *arXiv preprint arXiv:2102.01293*.

Xue, S.; Zhang, J.; Zhou, J.; and Ren, F. 2022. Robust Automated Essay Scoring by Using Attentive Capsule. Accessed: 8-May-2024.