Legacy NLP vs Modern Transformer Architecture Using DistilBERT

Chandler Grote, Nick Brady, Jakob Wickham

May 7, 2025

0.0.1 Abstract

Revisiting and improving upon prior research is how we advance scientific understanding. We aimed to reproduce the Natural Language Processing (NLP) methods described in Chen et al.(2016) with an updated architecture. This paper explores practical advantages and benchmark accuracy between current transformer NLP methods and legacy neural network-based approaches. To create a controlled comparison, we recreated the original data set using the methods described in the Git repository (Chen, n.d.). We compared those results to modern transformer architecture in NLP with Distilled Bidirectional Encoder Representations from Transformers (DistilBERT), which would improve the accuracy of previous NLP methods. The main challenge is that modern transformer-based NLP has high hardware demands which require workarounds, including dataset limits and infrastructure tradeoffs. However, we achieved a peak accuracy score of 89.44% using DistilBERT while the legacy framework only achieved 23.09% at its peak.

0.0.2 Key Terms

- **Gated Recurrent Units (GRU)**: Lightweight neural networks that process sequences of words while retaining memory of previous inputs. Efficient and commonly used in earlier NLP systems. (Anishama, 2023)
- **Transformer**: Modern architecture that simultaneously processes the entire input using self-attention, but has higher computational demands. (Aguilera, 2025)
- **Attention**: A mechanism that allows the model to focus on the most essential parts of the context, improving decision-making by weighing its importance. (Winland, 2025)
- **Bidirectional Encoder Representations from Transformers (BERT)**: A deep learning model using transformer-based architecture to understand sentence context by parsing left-to-right and right-to-left.
- **Tokenization**: The process of breaking text into smaller units called tokens. Some models use a fixed token limit, which can force the shortening of long inputs, a process called truncation.

1 Introduction

Natural Language Processing is a branch of machine learning focused on enabling computers to understand and interpret human language. Both methods used in this paper rely on a cloze-style question preprocessor, shown in Figure 1. A cloze-style question is a fill-in-the-blank question with one masked word, typically a named entity such as a person, place, or thing. The model must predict the correct answer based on the surrounding context. The two models we use in this paper differ in executing this process. The legacy GRU framework masks the word, constructs the cloze-style question, and ranks possible entity candidates (Anishnama, 2023). The entity marker with the highest score is selected as the predicted answer. The modern transformer framework relies on token-level prediction based on a fixed vocabulary (Aguilera, 2025). It turns the task into a multiple-choice problem based on token likelihoods. The versatility achieved by doing this comes at the cost of significantly higher computational resources.

Figure 1: An example of a cloze-style format

```
Question: celebrity politician and former wrestler @placeholder had posed a
question concerning extraterrestrials to a government committee
Context: ( @entity2 ) the classic video game " @entity1 " was developed in
@entity3 back in the late 1970 's -- and now their real - life counterparts are
the topic of an earnest political discussion in @entity3 's corridors of power .
Label: @entity18
```

Figure description: *This figure illustrates the cloze-style question format used in both the GRU and DistilBERT models. A named entity is masked within the question, and the model must infer the correct answer based on the surrounding context. This format is foundational to the CNN/Daily Mail datasets.*

To reduce computational demands, we used a distilled model. Distilled models are smaller models, known as students, that are trained to mimic behavior from a larger teacher model. BERT is trained by masking some words in the text and then using the context in the text to the left and right of the masked word to predict it. DistilBERT is a distillation of BERT, which utilizes the same underlying process but runs on a smaller and more efficient basis. By reducing the number of parameters, DistilBERT can achieve similar goals on less intensive hardware while maintaining comparable levels of accuracy.

Chen et al.(2016) introduced large-scale cloze-style question answering tasks using new *CNN* and *Daily Mail* articles. Each article included bullet point summaries with named entities. The model aims to predict the masked entity by leveraging the article's context. They implemented a recurrent neural network with GRU and an attention mechanism to help align relevant parts of the article with the questions, which was optimized with computational demand and interoperability. Modern models like BERT and its distillations use transformer architecture, which applies parallel self-attention rather than sequential (Aguilera, 2025). These models are structurally different and have different resource demands. Both models aim to map a cloze-style question using the article context to a predicted entity.

2 Methodology

Theano and Lasagne were implemented in the legacy model to reproduce the original architecture by Chen et al.(2016). It could process full-length articles without truncation, requiring minimal processing. This allowed the model to operate on variable-length inputs without the need to filter or constrain articles. Training was conducted using a batch size of 96 on consumer-grade hardware, a Quadro RTX 3000 GPU. Due to modern compatibility issues with Therano and Lasagne, reproducing the model required a virtual Ubuntu machine with a legacy CUDA driver and libraries to ensure compatibility and GPU acceleration. This GRU model was trained from scratch without pretrained embedding or external knowledge, making it a benchmark for evaluating models with minimal dependencies.

Hugging Face Transformers and Pytorch were our baseline for the modern model architecture. We selected distilbert-base-unccased as the pretrained model architecture, and set floating-point-16 precision to reduce memory usage. The preprocessing pipeline truncated each article around the answer and filtered out samples with fewer than four entities present. To prevent token overflow, we used an approximation of final token counts. These adjustments excluded over 70% of the data due to token length constraints or a lack of entities. Even with this significant data loss, the retained samples still reflected messy data of real-world environments, as seen in Figure 2. DistilBert was trained on an A100 GPU through Google Colab Pro using a batch size of 1 and no gradient accumulation. The model was evaluated on a separate development subset after each epoch.

Both models were trained using hyperparameters over three epochs to maintain fairness. Evaluation relied on accuracy to focus on answer correctness and aligned with the outputs of the legacy implementation. Both models' data sets were constructed using Google DeepMind's question story files from the NYU Archive (Cho, n.d.). The original train/dev/test splits were preserved. By stratified random sampling, 80,000 training samples were selected from *CNN* and *Daily Mail* articles to

maintain article diversity and reduce computational constraints on the transformer architecture.

Figure 2: Entity count distribution

	Dataset	Split	Filtered Count	Raw Max	Raw Mean	Raw Median
0	CNN	Train	25256	527	26.39	23.0
1	CNN	Validation	1376	187	26.19	22.0
2	CNN	Test	1007	394	24.22	21.0
3	DailyMail	Train	16019	329	26.25	22.0
4	DailyMail	Validation	16424	230	25.02	21.0
5	DailyMail	Test	12731	245	25.48	21.0

Figure description: *Even with filtering out over 70% of the original dataset to meet the tokenizer length requirements, the splits show decent variability. This shows that even after aggressive preprocessing, the data simulates how noisy real-world data is, and even with the smaller split sizes, the models will still be challenged.*

3 Results

During the training, there was a clear divide in each model's convergence characteristics. DistilBERT showed a rapid improvement in accuracy within the first thousand steps and increased over the subsequent epochs. This is expected, as the model had an advantage from extensive pretraining on the general domain. The GRU reached an early plateau, showing limited gains after epoch one, even with stable behavior during training. Training loss mirrored this trend, as DistilBERT consistently reduced training loss over time while the GRU model stabilized at a higher loss. DistilBERT's pretrained initialization provided a significant head start, where the GRU architecture training from scratch was limited by the limited cloze-style data it was provided, see figures 3 and 4 for final accuracy, training time, and hardware differences.

Figure 3: Accuracy comparisons between GRU and DistilBERT

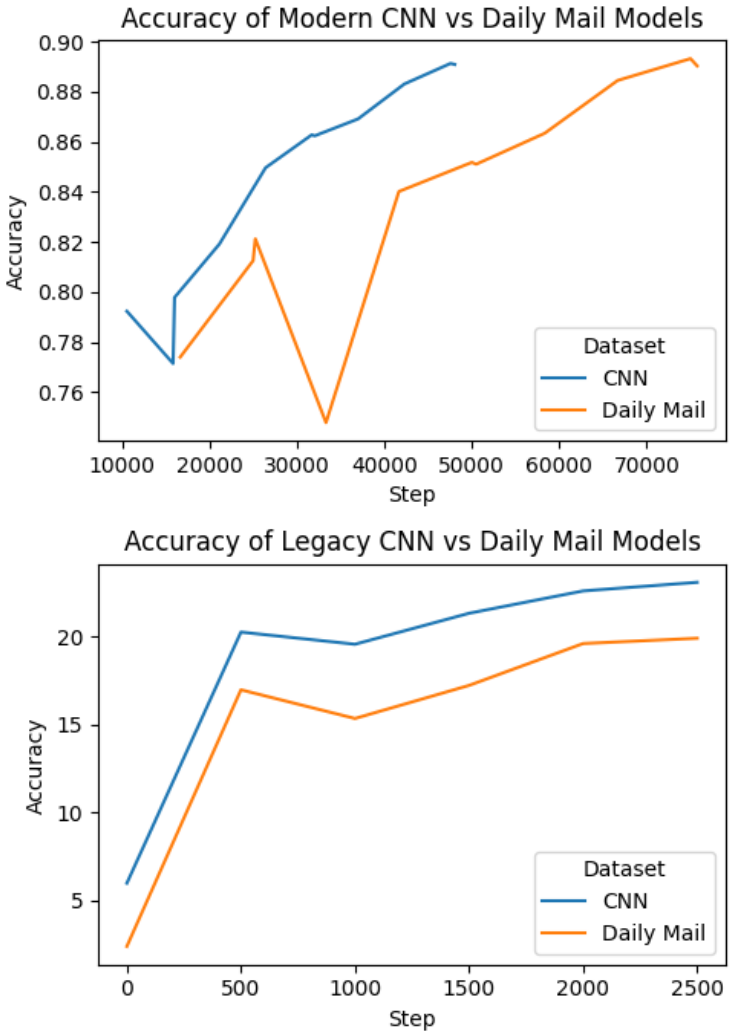

Figure 4: Table of performance and accuracies of GRU and DistilBERT

Model	Dataset	Best Accuracy	Training Time	Notes
GRU	CNN	23.09%	5h 21m	Full context, consumer GPU
GRU	Daily Mail	19.91%	10h 24m	Full context, consumer GPU
DistilBERT	CNN	89.44%	4h 11m	Truncated input, A100 GPU
DistilBERT	Daily Mail	86.79%	1h 51m	Truncated input, A100 GPU

Figure descriptions: *DistilBERT, as shown in both the table and the graph, achieves a much higher accuracy due to pretraining and having a smaller dataset to work with as well as dealing with smaller batch sizing, but at the cost of requiring more complicated preprocessing and intensive computational power, requiring an A100 GPU to be able to train the models.*

Superior performance was achieved with DistilBERT in both datasets, but it came at significant tradeoffs. This model required high-end hardware, aggressive preprocessing, and exclusion of most of the dataset to meet memory constraints. While the GRU model ran entirely on consumer-grade hardware, it had minimal processing and could use full-length article context. The GRU model was less accurate, but its ability to be used and its interpretability make it a consideration with limited resources or technical requirements

4 Conclusion

This exploration shows one of the central debates in NLP: the newest model isn't always the right tool. DistilBERT's performance is four times greater than the GRU model's, but this comes at the cost of extensive infrastructure, massive data pruning, and reliance on pre-trained language, showing the power and limitations of modern transformer-based architecture. Still, it shows the growing barrier for anyone without access to high-end computing. Despite being made nine years ago, the GRU model has proven capable of handling long-form input and offering relative ease of deployment. It is compatible with unfitted full article text while operating in resource-constrained environments. These traits, simplicity, transparency, and compatibility, offer meaningful advantages for resource-limited settings. These often get lost due to the ever-increasing need for benchmark improvements. Ultimately, these tools are only as good as the context in which they are used. Data scientists must understand when and why to use them, choosing tools that align with the problem, its constraints, and intended impact. Revisiting older models isn't for nostalgia. It's about expanding our perspective and re-evaluating which solution meets our goals. As NLP continues to evolve, looking back can still show us paths forward.

5 References

Chen, D., Bolton, J., & Manning, C. D. (2016). A thorough examination of the CNN/Daily Mail Reading Comprehension Task. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* https://doi.org/10.18653/v1/p16-1223

Chen, D. (n.d.). *Danqi/RC-CNN-dailymail: CNN/Daily Mail Reading Comprehension task.*

GitHub. https://github.com/danqi/rc-cnn-dailymail

Aguilera, F. M. (2025, April 18). *Transformer-based multiple choice question answering: Implementation, output analysis, and bias...* Medium. https://medium.com/ai-simplified-in-plain-english/transformer-based-multiple-choice-question-answering-implementation-output-analysis-and-bias-e04d3d6d9c03

Winland, V. (2025, March 11). *What is self-attention?*. IBM. https://www.ibm.com/think/topics/self-attention

Anishnama. (2023, May 4). *Understanding gated recurrent unit (GRU) in deep learning.* Medium. https://medium.com/@anishnama20/understanding-gated-recurrent-unit-gru-in-deep-learning-2e54923f3e2

Cho, K. (n.d.). *DeepMind Q&A Dataset.* DMQA. https://cs.nyu.edu/~kcho/DMQA/