Comparing Performance of different Transformer models on non auto-regressive NLP tasks

Barath Ramashankar
NYU Tandon
New York University
br2543@nyu.edu

Sampreeth Avvari
NYU Tandon
New York University
sp9659@nyu.edu

Abstract

This project evaluates the performance of different transformer architectures—encoder-only (e.g., RoBERTa), decoder-only (e.g., GPT-2), and encoder-decoder (e.g., T5)— on typically non-autoregressive NLP tasks such as text classification, NER, sentence similarity, summarization, and translation. Using models of comparable size, fine-tuned on standard datasets, we evaluate various standard metrics based on the task, inference time, and resource efficiency. This analysis aims to guide model selection by identifying the most effective architectures for specific non-autoregressive applications.

1 Introduction

This project evaluates the effectiveness of transformer architectures—encoder-only (DistilBERT), decoder-only (GPT2), and encoder-decoder (T5)—on non-autoregressive NLP tasks, including text classification, NER, and sentence similarity. We compare Small (60–80M), Medium (200–250M), and Large (700–900M) models to identify their strengths and weaknesses across performance, speed, and resource efficiency. The goal is to provide insights for developing optimized NLP solutions for real-world applications.

2 Related Work

One foundational work in comparing transformer architectures is BERT by Devlin et al. (2019), which introduced a bidirectional transformer excelling at non-autoregressive tasks like text classification and NER, setting new benchmarks on GLUE and SQuAD. T5 by Raffel et al. (2020) extended this with an encoder-decoder architecture, unifying NLP tasks into a sequence-to-sequence format and outperforming BERT on both autoregressive and non-autoregressive tasks. However, a concise comparison of these models' performance on non-autoregressive tasks remains lacking.

3 Approach

This project evaluates transformer models of varying sizes—Small, Medium, and Large—on non-autoregressive NLP tasks (classification, NER, sentence similarity, and summarization) to assess the impact of model size and architecture. Each tier includes an encoder-only, decoder-only, and encoder-decoder model, enabling a comprehensive analysis of how these architectures perform across tasks. By fine-tuning and comparing their effectiveness and efficiency, we aim to identify the optimal configurations for each task.

3.1 Model Tiers

Our model selection covers three parameter scales:

- **Small (60-90M):** DistillGPT2 (80M, Decoder), T5 Small (60.5M, Encoder/Decoder), DistilBERT (66M/88M, Encoder).
- **Medium (200-350M):** GPT-Medium (354, Decoder), T5-base (223M, Encoder/Decoder), RoBERTa Medium (210M, Encoder).

- **Large (700-800M):** GPT2 Large (774M, Decoder), T5 Large (738M, Encoder/Decoder), DeBERTa XLarge (750M, Encoder).

4 Experiments

4.1 Data

To evaluate transformer models across non-autoregressive NLP tasks, we utilize several established datasets, each designed for a specific task: text classification, named entity recognition (NER), sentence similarity, and summarization. Each dataset offers a unique format and output structure, enabling a comprehensive evaluation of model performance.

Text Classification: AG News [3], Named Entity Recognition (NER): CoNLL-2003 [4], Sentence Similarity: SICK [5], Summarization: CNN/Daily Mail [6]

4.2 Evaluation method

We use standard evaluation metrics for non-autoregressive tasks, including accuracy, F1 score, precision, and recall. For sentence similarity, we add Pearson and Spearman correlations, Mean Squared Error (MSE), and Mean Absolute Error (MAE) to capture semantic similarity more effectively. We are maintaining the same hyper-parameters across all models to have a fair comparison. For Summarisation, We used ROUGE, BLEU, and METEOR metrics for evaluation.

4.3 Results

NER

We have the results of the all the models sizes below. We will start with named entity recognition.

Model/Metric	Accuracy	F1	Recall	Precision
Small				
DistillBERT	**0.98028**	**0.92204**	**0.9315**	**0.91276**
T5-Small	0.6506	0.72759	0.65206	0.82611
Distill-GPT2	0.97171	0.79264	0.82178	0.7677
Medium				
Xlmroberta	**0.9881**	**0.94033**	**0.95086**	**0.93004**
T5-Base	0.98969	0.61643	0.49302	0.82193
GPT2-Medium	0.97173	0.7969	0.82935	0.76688
Large				
DeBERTa-Xlarge	**0.99247**	**0.96567**	**0.97038**	**0.9610**
T5-Large	0.99726	0.61895	0.5278	0.8423
GPT2-Large	0.97572	0.82418	0.84769	0.80562

Table 1: Performance comparison of transformer models for Named Entity Recognition.

The results show that encoder-only models consistently outperform others on the NER task, while decoder-only models remain competitive but fall short. Encoder-decoder models lag significantly, indicating potential inefficiencies or architectural misalignments for NER, which favors models optimized for feature extraction and sequence-level understanding.

Sentence Similarity

Model/Metric	MSE	MAE	Pearson	Spearman
Small				
DistillBERT	0.3534	0.44229	0.82184	0.75825
T5-Small	**0.3427**	**0.42147**	**0.8567**	**0.8088**
Medium				
Xlmroberta	0.3512	0.4322	0.8378	0.7756
T5-Base	**0.2360**	**0.3535**	**0.8854**	**0.8343**
Large				
DeBERTa-Xlarge	0.3466	0.4289	0.8394	0.7814
T5-Large	**0.257**	**0.3664**	**0.8717**	**0.8234**

Table 2: Performance comparison of transformer models for sentence similarity.

Model/Metric	Accuracy	F1	Recall	Precision
Small				
Distill-GPT2	0.2020	0.2124	0.2020	0.2741
Medium				
GPT2-Medium	0.3980	0.3945	0.3980	0.4376
Large				
GPT2-Large	**0.6853**	**0.5243**	**0.5142**	**0.6795**

Table 3: Performance comparison of GPT models for sentence similarity.

The task of generating sentence similarity scores in the range of 0 to 5 (including decimal values) using decoder-only models like GPT-2 proved to be highly ineffective. Decoder-only architectures are not particularly well-suited for numerical tasks, and this limitation was evident even when employing few-shot instruction-based fine-tuning methods. The models struggled to generate meaningful scores and often produced the same number repeatedly for all sentence pairs, failing to capture the variability in the data. As a result, it was not feasible to compute evaluation metrics, making the approach unviable for the task. so the task was reformulated into a classification problem by rounding scores to the nearest integers. Smaller models performed poorly, showing low accuracy and evaluation metrics. Larger models performed slightly better, achieving 0.6853 accuracy, 0.5243 F1, 0.5142 recall, and 0.6795 precision, highlighting scale's limited ability to mitigate these issues. Despite the reformulation, decoder-only models remained suboptimal compared to stronger results from encoder-decoder and models.

Classification

Model/Metric	Accuracy	F1	Recall	Precision
Small				
DistillBERT	0.93645	0.93632	0.93645	0.93623
T5-Small	**0.94408**	**0.94409**	**0.94408**	**0.9441**
Distill-GPT2	0.94079	0.94075	0.94079	0.94075
Medium				
Xlmroberta	0.94263	0.94271	0.94263	0.9429
T5-Base	**0.95197**	**0.95198**	**0.95197**	**0.95201**
GPT2-Medium	0.94329	0.94331	0.94329	0.94334
Large				
DeBERTa-Xlarge	0.94871	0.94705	0.9487	0.94569
T5-Large	**0.95316**	**0.95315**	**0.95316**	**0.95315**
GPT2-Large	0.93737	0.93734	0.95737	0.93734

Table 4: Performance comparison of transformer models across sizes and metrics for Classification.

The classification results show that T5 consistently outperforms other architectures across all model sizes. T5-Small achieves the best results among Small models (Accuracy, F1, Recall: 0.94408; Precision: 0.9441). T5-Base leads in the Medium category (Accuracy: 0.95197, F1: 0.95198), and

T5-Large excels among Large models (Accuracy, Recall: 0.95316; F1, Precision: 0.95315). This highlights the superior performance of encoder-decoder models for classification tasks.

Summarization

Model/Metric	**BLEU**	**ROUGE-L**	**METEOR**	**Train Loss**
Small				
T5-Small	0.10734	0.21860	0.26657	1.94713
Distill-GPT2	**0.89843**	**0.92800**	**0.92539**	**2.85213**
Medium				
T5-Base	0.18971	0.34406	0.38305	1.13754
GPT2-Medium	**0.88485**	**0.92275**	**0.88554**	**2.09448**
Large				
T5-Large	0.31149	0.47222	0.49424	0.96410
GPT2-Large	**0.88695**	**0.92327**	**0.89083**	**1.33120**

Table 5: Performance comparison of transformer models for the summarization task.

The experiments were conducted on the CNN/DailyMail dataset, which contains news articles paired with highlights for text summarization tasks. Encoder-decoder models, such as T5, process the input by encoding the full context and generating concise outputs through a decoder. In contrast, decoder-only models like GPT generate summaries autoregressively, producing one token at a time based on the left-context.

From the results, decoder-only models outperform encoder-decoder models across all evaluation metrics. Notably, **GPT2-Large** achieves the highest BLEU, METEOR, and ROUGE scores, showcasing its superior ability to generate fluent and human-readable summaries. While T5-Large performs better than its smaller variants, it still lags behind GPT models, suggesting that decoder-only architectures are more effective for fine-tuning tasks in text summarization.

Encoder-only models like DistilBERT, designed for classification and token-level tasks, proved unsuitable for summarization tasks requiring coherent text generation. Even after adding a decoder head to transform them into encoder-decoder architectures (effectively defeating the purpose of this experiment), the results remained suboptimal. Although this modification increased the model size by 30% due to additional parameters for autoregressive generation, encoder-only models still struggled with learning long-range dependencies and sequential generation. This shows that encoder-only transformers are fundamentally limited for tasks requiring fluent and cohesive text generation.

Power Consumption

This analysis evaluates the GPU performance and efficiency of **Encoder-Only**, **Encoder-Decoder**, and **Decoder-Only** transformer architectures across NER, sentence similarity, classification, and summarization tasks. **Decoder-Only models** were the most power-hungry, reaching up to ~225–250W, with high GPU utilization (~90%) and clock speeds (~1600 MHz). **Encoder-Only models** excelled in power efficiency, consuming ~111–150W while achieving the best results for NER and sentence similarity tasks. **Encoder-Decoder models** balanced power consumption and performance, performing best for classification tasks. For summarization, **GPT models** delivered superior results but at significant power costs, highlighting their trade-off between performance and resource usage.

5 Conclusion

This study evaluated transformer architectures—encoder-only, decoder-only, and encoder-decoder—across non-autoregressive NLP tasks like NER, sentence similarity, classification, and summarization. Encoder-only models excelled at NER due to their strong feature extraction capabilities. Encoder-decoder models demonstrated versatility, achieving the best results for sentence similarity and classification. Decoder-only models outperformed others in summarization, benefiting from their autoregressive generation ability.

In terms of resource efficiency, encoder-only models were the most power-efficient, while decoder-only models delivered superior performance at higher power costs. Overall, aligning model architecture with task requirements is key to achieving optimal results and balancing performance with resource constraints. The entire codebase can be found at this link.

References

[1] J. Devlin, M. Chang, K. Lee, and K. Toutanova. "BERT: Pre-training of deep bidirectional transformers for language understanding." Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019.

[2] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. "Exploring the limits of transfer learning with a unified text-to-text transformer." Journal of Machine Learning Research, 2020.

[3] X. Zhang, J. Zhao, and Y. LeCun. "Character-level convolutional networks for text classification." Advances in Neural Information Processing Systems, 2015.

[4] E. Tjong Kim Sang and F. De Meulder. "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition." Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, 2003.

[5] M. Marelli, S. Menini, M. Baroni, L. Bentivogli, R. Bernardi, and R. Zamparelli. "A SICK cure for the evaluation of compositional distributional semantic models." Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC), 2014.

[6] K. M. Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom. "Teaching machines to read and comprehend." Advances in Neural Information Processing Systems, 2015.

[7] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. "RoBERTa: A robustly optimized BERT pretraining approach." arXiv preprint arXiv:1907.11692, 2019.

[8] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. "Language models are unsupervised multitask learners." OpenAI technical report, 2019.

[9] P. He, X. Liu, J. Gao, and W. Chen. "DeBERTa: Decoding-enhanced BERT with disentangled attention." International Conference on Learning Representations (ICLR), 2021.