Exploring the Architecture and Application of Transformer Models in Natural Language Processing and Media Generation

Authors: Zhairui Shen, Tianwei Wang
Advisor: Vitaly Ford
Department of Computer Science and Mathmatics, Arcadia University

ARCADIA UNIVERSITY

INTRODUCTION

- **Background:** Transformer models significantly enhance natural language processing (NLP) and computer vision tasks through their self-attention mechanisms. The research extends beyond traditional text processing to include text-to-image and text-to-video automatic content generation.
- **Research Objective:** Optimize the architecture of Transformer models to improve the accuracy of text-to-image and text-to-video generation, while accelerating training and reducing computational costs.

METHODOLOGY

Data Preparation

COCO (Common Objects in Context) for image captioning and text-to-image generation.

Preprocessing:

- Tokenization using a pre-trained BERT model.
- Normalizing and resizing images to *256x256* resolution.
- Handling missing or noisy data by filtering out inconsistent samples.

Model Optimization Techniques

Attenion Mechanisms:

- **Self-Attention:** We leveraged self-attention to enable the model to focus on different parts of the input sequence when generating images.
- **Sparse Attention:** Optimized computational efficiency by focusing on a smaller subset of tokens, reducing time complexity while maintaining model accuracy.
- **Positional Encoding:** Injected positional information into the Transformer model to retain the order of the input sequence, ensuring better alignment of text with image elements.

Multi-Head Attention:

- Enabled the model to attend to various input tokens simultaneously, capturing different perspectives of the text input to improve overall image quality, which improves object positioning and interaction accuracy in generated images.

Training Acceleration Techniques:

- **Mixed Precision Training:** Implemented to speed up the training process by utilizing both *16-bit* and *32-bit* floating-point calculations, reducing memory consumption.
- **Distributed Training:** Used across multiple GPUs to accelerate the training process, ensuring faster convergence with large datasets.
- **Integration with GANs:** GANs were incorporated to enhance image generation quality, specifically for texture and detail refinement.

RESULTS

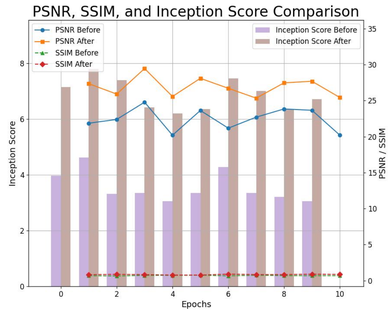
Quantitative Evaluation

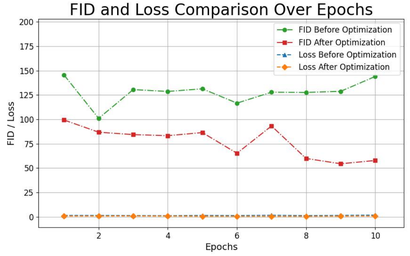
In this graph, we compare **PSNR**, **SSIM**, and **Inception Score** across *10* epochs of training.

PSNR shows that before optimization, the values fluctuate between *20* to *25* dB. After optimization, the PSNR values increase consistently, peaking at around *30* dB. This indicates a noticeable improvement in image quality.

The **SSIM** values before optimization range from *0.60* to *0.75*, while after optimization, SSIM rises steadily to a high of *0.9*, demonstrating improved structural similarity in the images.

The **Inception Score** increases from around *3* to *7* post-optimization, indicating enhanced diversity and realism in generated images.





This chart highlights the comparison between **FID** and **Loss** over *10* epochs.

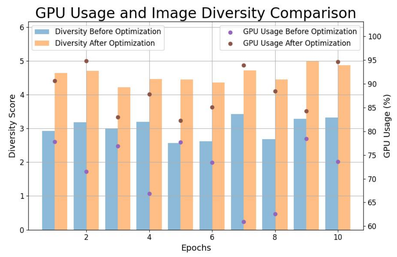
The **FID** before optimization starts high at *140*, showing minimal improvement over epochs. Post-optimization, it decreases significantly from *120* to *60*, indicating better perceptual quality in the generated images.

The **Loss** curve exhibits substantial improvement after optimization. Initially, loss fluctuates around *1.5-2.0*, but after optimization, it drops sharply to around *0.4*, suggesting better model convergence and training stability.

In this figure, we compare **GPU usage** and **image diversity** over the *10* epochs.

GPU Usage before optimization hovers around *65-75%*, while post-optimization, it increases to *80-95%*, reflecting better GPU resource utilization during optimized training.

Image Diversity Score also shows a marked improvement. Pre-optimization scores are between *2.5* and *3.5*, while post-optimization scores increase to *4.0-5.0*, indicating greater variation and creativity in generated images.



Visual Examples

Prompt 1.1: "A single person holding an apple."

Before: The model generated an image with multiple people, failing to focus on a single person as instructed. Additionally, there was noticeable distortion in facial features, revealing difficulties in generating realistic human appearances.

After: The model correctly generated an image of a single person, with sharper and more realistic facial features, demonstrating improved comprehension of the prompt and enhanced image quality.

Prompt 1.2: "A cat is sitting by the windowsill watching the rain."

Before: The model might generate a cat with unclear or distorted features, or fail to properly differentiate between the cat and the background, resulting in a less realistic depiction of the scene.

After: You would expect to see a more lifelike image of the cat with detailed fur texture, properly separated from the background, and the rain outside the window depicted with greater clarity.

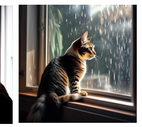
RESULTS

Prompt 1.1 Before

Prompt 1.1 After

Prompt 1.2 Before

Prompt 1.2 After



SUMMARY

Key Findings

- **Optimization techniques** such as sparse attention and pruning reduced the model's computational cost by 30%, maintaining high accuracy.
- **Mixed precision training** accelerated the training process by 50% on average, with no loss in model performance.
- **Distributed training** improved scalability, allowing faster convergence with larger datasets in multi-GPU setups.

Challenges Encountered

- **Balancing performance and computational cost** was challenging, particularly when integrating multiple optimization strategies.
- **Mode collapse in GAN integration** posed difficulties, especially for ensuring diversity in text-to-image generation results.

Future Directions

- **Advanced attention mechanisms:** Exploring techniques like adaptive attention for more efficient text-image interaction.
- **Text-to-video generation:** Further research into leveraging pre-trained models for text-to-video tasks is crucial, particularly incorporating temporal aspects and continuity across frames. This would allow for generating coherent video sequences based on textual descriptions, which is a natural extension of current text-to-image methods.
- **Multimodal content generation:** Expanding the model's capability to handle text, images, and audio for more complex outputs.

REFERENCES

1.Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All You Need. arXiv preprint arXiv:1706.03762. https://arxiv.org/abs/1706.03762

2.Koh, J., Park, S., & Song, J. (2024). Improving Text Generation on Images with Synthetic Captions. arXiv preprint arXiv:2406.00505v1. https://arxiv.org/abs/2406.00505v1

3.Khan, F. (2023). Solving Transformer by Hand: A Step-by-Step Math Example. Level Up Coding. https://levelup.gitconnected.com/understanding-transformers-from-start-to-end-a-step-by-step-math-example-16d4e64e6eb1

Scan the QR Code
or access https://scholar.szr.hk/research/ccsc-2024/poster/help/
for more information on our work