UNIVERSITY OF TORONTO

University of Toronto

STA414 Assignment 4

Student :
Student: Shiva Chandrachary
Student ID: vishwak7
Student Number: 1009674461
Email:
shiva.chandrachary@mail.utoronto.ca

Teacher :
Murat Erdogdu

March 31, 2025

Contents

1 Part I - Relevant Course Papers

Table 1: Summary of Research Papers on Attention, Embedding, and Transformers

Paper Title	Description	Relevance to Course
Attention Is All You Need (https://arxiv.org/abs/1706.03762)	This paper introduces the Transformer architecture, which eliminates the need for recurrence in sequence-to-sequence models by relying entirely on self-attention mechanisms. The model significantly improves parallelization, reducing training time while achieving state-of-the-art results in machine translation tasks. The introduction of multi-head self-attention and positional encodings makes it a foundation for modern NLP architectures.	This paper is foundational to understanding attention mechanisms and embeddings. It connects the concepts learned on generalized representation of skip-grams, autoencoders and attention. It provides a practical application of the attention concepts discussed in the lectures, showcasing how self-attention can replace traditional recurrent models to enhance performance and efficiency in sequence-to-sequence tasks.
All Word Embeddings from One Embedding (https://arxiv.org/abs/2004.12073)	This paper proposes a novel approach where all word embeddings are derived from a single base embedding using a transformation function. It challenges traditional word embedding models like Word2Vec and GloVe by demonstrating that a single embedding can reconstruct multiple word embeddings with high accuracy. The method significantly reduces storage costs while maintaining competitive performance on NLP benchmarks.	This paper aligns with the course's exploration of word embeddings and their applications in NLP tasks. It offers an alternative perspective on embedding techniques, emphasizing efficiency and scalability, which are key considerations discussed in the lectures. This paper is relevant to the study of embeddings by offering an alternative to storing large word vector models. It connects to concepts of Word2Vec, BoW, bBoW and tokenizers.
Unlimiformer: Long-Range Transformers with Unlimited Length Input (https://arxiv.org/abs/2305.01625)	This paper introduces Unlimiformer, a transformer model that allows for unlimited-length input using a memory-efficient key-value retrieval mechanism. Instead of truncating input or using fixed-length memory, Unlimiformer dynamically retrieves relevant information as needed, achieving state-of-the-art results in long-context tasks such as document summarization and long-range question answering.	The paper expands on transformer models and attention mechanisms covered in the course. It addresses the quadratic complexity of self-attention by introducing an efficient retrieval-based method, making transformers more applicable to long-text processing and real-world applications such as knowledge retrieval and document understanding.

An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (https://arxiv.org/abs/2010.11929)	This paper introduces the Vision Transformer (ViT), which applies transformer-based architectures to image recognition by treating an image as a sequence of non-overlapping patches. Instead of using convolutional layers like CNNs, ViT flattens image patches into tokens and applies self-attention. The model achieves state-of-the-art accuracy on image classification tasks, outperforming CNNs when trained on large datasets.	The paper extends transformer architectures from NLP to computer vision, demonstrating how self-attention can replace convolutions. It highlights the concept of embedding non-text data into a transformer-friendly format and reinforces ideas from the course on self-attention, sequence modeling, and positional encoding.
Not All Images are Worth 16x16 Words: Dynamic Transformers for Efficient Image Recognition (https://arxiv.org/abs/2105.15075)	This paper introduces DynamicViT, an improved Vision Transformer (ViT) that dynamically prunes unimportant image tokens to improve computational efficiency while maintaining accuracy. Instead of treating all image patches equally, DynamicViT learns to drop less informative tokens at different layers, significantly reducing the number of processed tokens. This leads to faster inference and lower computational costs without sacrificing performance. The method achieves competitive results on ImageNet while being significantly more efficient than standard ViT models.	The paper builds on transformer architectures discussed in the course and extends the Vision Transformer (ViT) approach by introducing dynamic token pruning. This aligns with course topics on efficient model design, self-attention mechanisms, and computational trade-offs. The work also highlights the importance of adaptively processing different inputs, a key idea in optimizing transformer-based architectures.

2 Part II - Table of hypotheses

Table 2: Table of Hypotheses for Improving Transformer Models

Hypothesis	**Justification**
Combining Dynamic Token Pruning with Long-Range Attention	Unlimiformer [6] enables long-document processing, while DynamicViT [8] prunes image tokens. Applying dynamic pruning to long-text attention could reduce unnecessary computations while maintaining global context.
Adaptive Patch Sizing for Vision Transformers	Vision Transformers (ViT) [1] use fixed 16x16 patches, but DynamicViT [8] suggests importance-based token pruning. Using adaptive patch sizes—larger for smooth regions and smaller for detailed textures—could enhance both accuracy and efficiency.
Using Unlimiformer-Style Memory Retrieval in Image Transformers	Unlimiformer [6] retrieves long-text memory dynamically. Applying similar retrieval mechanisms to image recognition could reduce redundant computations while retaining key visual information.
Integrating Local and Global Attention for Efficient Text Processing	The Transformer [5] uses full self-attention, but hybrid models could **dynamically switch between local and global attention** depending on sentence importance, improving efficiency while maintaining long-range dependencies.
Integrating Image Tokenization with Text Embedding Strategies for Cross-Modal Image Generation	By combining image tokenization from ViT [1] and text embedding strategies from "All Word Embeddings from One Embedding" [4], cross-modal attention mechanisms can be used to enhance both image generation and reconstruction with fewer tokens. The paper "An Image is Worth 32 Tokens for Reconstruction and Generation" [7] demonstrates that reducing token size leads to more efficient image generation.

3 Part III - Hypothesis verification and explanation

3.1 Introduction

The paper *"An Image is Worth 32 Tokens for Reconstruction and Generation"* [7] introduces an advanced tokenization approach for images, reducing the number of tokens required to effectively represent and reconstruct an image. This significantly improves efficiency in image generation tasks while maintaining high-quality output. The paper builds on the ideas of Vision Transformers (ViT) by proposing a learned tokenizer that enables image reconstruction with far fewer tokens than standard approaches.

3.2 Hypothesis Explanation

The hypothesis, **Integrating Image Tokenization with Text Embedding Strategies for Cross-Modal Image Generation**, proposes integrating image tokenization techniques with text embedding strategies to enhance cross-modal image generation. In multi-modal models like DALL·E [3] or CLIP [2], images and text need to be represented in a unified latent space. Existing models use separate embedding spaces for images and text, which can lead to inefficiencies and inconsistencies in cross-modal understanding. By leveraging efficient tokenization from the paper [7], we hypothesize that a shared embedding space for both image and text representations could enhance model performance in text-to-image tasks.

3.3 Key Results of the Paper

The paper [7] presents the following significant findings: - A learned image tokenizer is introduced, which reduces image token count from hundreds (as in ViT) to just **32 tokens** per image while preserving key visual details. The tokenized representation maintains high fidelity in image generation and reconstruction tasks. The method outperforms traditional patch-based Vision Transformers (ViT) by achieving better efficiency and generation quality with fewer computational resources. This approach is particularly useful for **generative tasks**, where efficiency is critical in handling large-scale datasets.

These findings suggest that if similar tokenization techniques are applied to multi-modal transformers, they could streamline the representation of images in a cross-modal setting, making image-text interactions more efficient and coherent.

3.4 Next Steps

To extend this research in the direction of our hypothesis, the following next steps could be taken: **Develop a shared latent space**: Investigate how learned image tokenization can be aligned with text embeddings from models like CLIP [2] to create a unified embedding space. **Cross-modal attention mechanisms**: Introduce mechanisms that allow image tokens and text tokens to interact dynamically, enhancing multi-modal understanding. **Fine-tuning on diverse datasets**: Test the integration of efficient image tokenization in text-to-image generation models on datasets like MS-COCO and LAION-5B. **Evaluating multi-modal learning efficiency**: Measure improvements in computational efficiency and generation quality compared to existing approaches.

References

[1] A Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

[2] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

[3] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*, 2021.

[4] S Ruder et al. All word embeddings from one embedding. *NeurIPS*, 2020.

[5] A Vaswani et al. Attention is all you need. *NeurIPS*, 2017.

[6] Y Xu et al. Unlimiformer: Long-range transformers with unlimited length input. *NeurIPS*, 2023.

[7] J Zhu et al. An image is worth 32 tokens for reconstruction and generation. *NeurIPS*, 2024.

[8] X Zhu et al. Not all images are worth 16x16 words: Dynamic transformers for efficient image recognition. *NeurIPS*, 2021.