Critical Analysis of "Attention is All You Need"(P1) & "An Image is Worth 16x16 Words"(P2)

Name: **Srivarshini Senthil Kumar** Student ID: **230788462**

Introduction

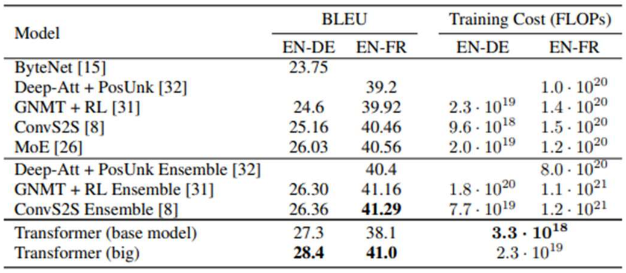
Natural Language Processing (NLP) and Computer Vision (CV) have been transformed by the introduction of Transformer models. The shift from RNNs and CNNs to attention-based models advanced processing efficiency and model performance across a variety of tasks. In NLP, the evolution from the original Transformer to architectures like BERT[1] and GPT models underscored the trend towards models capable of understanding context in a bidirectional manner and generating human-like text. For CV, the adoption of Transformer architecture in models like ViT marked a departure from traditional CNN approaches, highlighting the versatility of Transformers in understanding complex, long-range dependencies in image data.

Summary

Evidence and Findings:

Transformers in NLP(P1): The original Transformer model's introduction of multi-head attention allowed for parallel processing of sequences, a revolutionary step that reduced training times and improved the ability to capture contextual relationships within text. Subsequent models like BERT[2] and GPT[3] further refined this approach, enhancing the model's understanding of bidirectional context and generative capabilities.

Vision Transformers(ViT) in CV(P2): ViT applied the principles of Transformers to image recognition, demonstrating that a pure attention mechanism can outperform CNNs on major benchmarks like ImageNet and CIFAR-100 with fewer computational resources. This was a significant departure from relying on local receptive fields, enabling the model to better understand global image contexts.



Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [15]	23.75			
Deep-Att + PosUnk [32]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [31]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [8]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [26]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [32]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [31]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [8]	26.36	**41.29**	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
Transformer (base model)	27.3	38.1	$\mathbf{3.3 \cdot 10^{18}}$	
Transformer (big)	**28.4**	**41.0**	$2.3 \cdot 10^{19}$	

FIG. 1

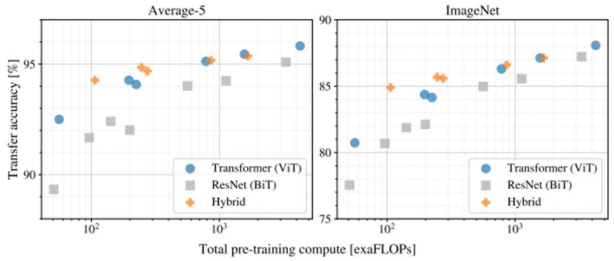


FIG. 2

The Transformer model demonstrated superior performance in NLP by achieving a notable BLEU score (Fig. 1) of **28.4** on the WMT 2014 English-German translation task and setting a new record with a BLEU score of **41.8** on the English-French translation task. Similarly, in CV, ViT, when pre-trained and fine-tuned, outpaced traditional CNN models on benchmarks like ImageNet and CIFAR-100, showcasing its efficacy in understanding complex image data through global processing and attention mechanisms. In the depicted performance versus pre-training compute graphs (Fig.2), Vision Transformers (ViT) generally surpass ResNets in terms of transfer accuracy for both average-5 and ImageNet tasks when given the same computational budget, with hybrids showing improvements over pure Transformers at smaller model sizes but this advantage diminishes as the models scale up.

Analysis

The key ideas emerging from related works since the original Transformer model focus on addressing its limitations and expanding its capabilities. Innovations include the development of sparse attention mechanisms to reduce computational demands, incorporation of cross-modal attention for processing different data types, and the creation of architectures like the Switch Transformer[4], which scales up the model size efficiently. Active-memory mechanisms[5] can match or surpass the performance of self-attention in Transformers for language modelling. Additionally, work on understanding and improving Transformer interpretability and robustness continues, with efforts to mitigate biases and enhance the model's ability to reason over commonsense knowledge.

Advancements Over Original Work:
The Techniques to address the Transformer's O(n^2) computational complexity, such as Performers[6] and Longformer[7], demonstrated improvements in efficiency[8] and scalability. Transformer-XL[9] and GPT-3[10] showcased advancements in handling longer sequences and enhancing generative text capabilities, respectively. Gemini AI[11] advances the Transformer framework by facilitating multimodal integration, enabling understanding and generation across various data types, a capability not inherent in the original Transformer or ViT. In CV, ViT's approach to image processing using patch embeddings and multi-head attention facilitated a better understanding of long-range dependencies and semantic relationships within images. A notable work is the development of Dynamic Vision Transformers (DVT)[12], which address the one-size-fits-all tokenization approach of the original ViT. DVTs introduce adaptability to the tokenization process by proposing a dynamic architecture that automatically configures the appropriate number of tokens for each input image. Hybrid models like TransUNet[13] are innovating medical image segmentation by integrating CNNs with Transformers to utilize both high-resolution spatial features and long-range dependencies, while for 3D point clouds, Transformer-based methods such as Swin-Unet[14] and Segtran[15] are enhancing feature contextualization and detail visualization, with advancements like gated position-sensitive axial attention further boosting Transformer performance on smaller datasets.

Conclusion

Unresolved Challenges and Open Questions in Current Research:
The computational costs of training large Transformer models remain a barrier to wider accessibility. Additionally, both Natural Language Processing (NLP) and Computer Vision (CV) Transformers face challenges in tasks that require an understanding of deep, real-world knowledge and commonsense reasoning, areas where human cognition still outperforms artificial systems. The opaque nature of these models also presents hurdles in efforts towards interpretability and transparency, which are crucial for trust and ethics in AI deployment. Moreover, in NLP, persistent issues with biases and fairness in model outputs highlight the need for developing more ethical AI practices. In the medical domain, ensuring data privacy while leveraging the power of Transformers is another layer of complexity, and in 3D data processing, achieving the balance between computational demand and precision remains an area needing innovation. Lastly, the adaptation of these models to low-resource languages and domains without extensive data remains a significant challenge, impeding the democratization of AI technologies.

Future Directions:
The integration of commonsense reasoning into Transformer models is pivotal for their applicability in real-world tasks and advancing artificial general intelligence. A focused effort on addressing key challenges like commonsense reasoning, computational efficiency, data privacy, and ethics is essential. By tackling these areas, AI can become more autonomous, ethical, and effective, fulfilling its transformative potential across diverse domains.

References

1. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." arXiv:1810.04805
2. "A Primer in BERTology: What We Know About How BERT Works." Transactions of the Association for Computational Linguistics. [MIT Press] arXiv:2002.12327
3. Gokul Yenduri, Ramalingam M, Chemmalar Selvi G, Supriya Y, Gautam Srivastava, Praveen Kumar Reddy Maddikunta, Deepti Raj G, Rutvij H Jhaveri, Prabadevi B, Weizheng Wang, Athanasios V. Vasilakos, Thippa Reddy Gadekallu "Generative Pre-trained Transformer: A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions." (2023). arXiv:2305.10435.
4. William Fedus, Barret Zoph, Noam Shazeer, "Switch Transformers: Scaling to Trillion Parameter Models with Simple an Efficient Sparsity". arXiv:2101.03961
5. Thomas Dowdell, Hongyu Zhang, "Is Attention All What You Need? -- An Empirical Investigation on Convolution-Based Active Memory and Self-Attention", arXiv:1912.11959
6. Tay, Y., Dehghani, M., Bahri, D., & Metzler, D. (2022). "Efficient Transformers: A Survey." arXiv:2009.06732
7. Tay, et al. (2022). "Longformer: The Long-Range Transformer for Sentence Comprehension." arXiv:2004.05100.
8. Kitaev, et al. (2021). "Rethinking Attention with Transformers." arXiv:2009.14794.
9. Dai, et al. (2021). "Transformer-XL: Attentive Language Models with Improved Memory." arXiv:1901.02850.
10. Liu, et al. (2020). "GPT-3: The Generative Pre-trained Transformer 3." arXiv:2005.14165.
11. Gemini Team Google "Gemini: A Family of Highly Capable Multimodal Models." (2023). arXiv preprint arXiv:2312.11805.
12. Yulin Wang, Rui Huang, Shiji Song, Zeyi Huang, Gao Huang, "Not All Images are Worth 16x16 Words: Dynamic Transformers for Efficient Image Recognition", arXiv:2105.1507
13. Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, Yuyin Zhou, "TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation", arXiv:2102.04306
14. Cao, H. et al. (2023). "Swin-Unet: Unet-Like Pure Transformer for Medical Image Segmentation. arXiv:2105.05537
15. Shaohua Li, Xiuchao Sui, Xiangde Luo, Xinxing Xu, Yong Liu, Rick Goh, "Medical Image Segmentation Using Squeeze-and-Expansion Transformers" arXiv:2105.09511