Arquitecturas Eficientes para LLMs: Estudo de Mixture of Experts

Artur Gil[1]*, Verónica Bolón-Canedo[1], Amparo Alonso-Betanzos[1], e Brais Cancela[1]

Universidade de A Coruña, España
a.gtorres@udc.es, veronica.bolon@udc.es, amparo.alonso.betanzos@udc.es, brais.cancela@udc.es

Resumo

A intelixencia artificial consolidouse como a tecnoloxía líder desta década, e o uso de ferramentas de xeración de texto baseadas en Large Language Models (LLMs) non deixa de medrar. Porén, o seu elevado consumo enerxético e impacto ambiental poñen de manifesto a necesidade de facelas máis eficientes. Ata agora, o progreso baseouse principalmente no escalado mediante hardware cada vez máis potente, mais esta tendencia está a cambiar. A optimización a través do software, como o é a computación condicional, está a xurdir como unha alternativa sostible para mellorar o rendemento e reducir custos ao mesmo tempo. Este traballo explora a técnica Mixture of Experts (MoE), que distribúe o cómputo activando unicamente as partes do modelo que mellor se adaptan a cada entrada. Na presente analízase a súa base teórica, desenvólvense diferentes implementacións e lévanse a cabo experimentos para avaliar as súas vantaxes e limitacións.

1. Metodoloxía

Para o estudo empírico destas ideas implementáronse e avaliáronse diversas arquitecturas MoE no contexto dos LLMs, co obxectivo de mellorar o rendemento e reducir a pegada ecolóxica. En MoE, unha rede de *gating* selecciona dinámicamente os expertos máis axeitados dun conxunto predefinido, activando só parte dos parámetros e logrando así un cómputo máis eficiente. Estes expertos impleméntanse como redes *feed-forward* (FFN) dentro dos bloques Transformer. Desenvolvéronse catro variantes de MoE e cinco mecanismos auxiliares para optimizar a técnica[1]. Todas as variantes foron integradas en nanoGPT , un modelo simplificado de GPT-2 empregado como liña base. Tres proceden da literatura [4, 3, 1], mentres que a cuarta, CondMLP, propúxose como un deseño condicional lixeiro orientado a reducir parámetros e custo computacional. O número de expertos mantívose fixo en 4, variando o número de expertos activos (*top-k*) segundo a estratexia de *gating*. Os modelos adestráronse de forma distribuída en dúas RTX 4090 co dataset FineWeb-Edu. O rendemento avaliouse sobre HellaSwag [6], e o impacto enerxético e ambiental monitorizouse con Zeus e Green Algorithms Calculator [5, 2].

2. Resultados

Os experimentos realizáronse a dúas escalas para avaliar as variantes MoE e os mecanismos auxiliares. A pequena escala, as variantes baseadas na literatura superaron moderadamente o modelo base de nanoGPT, mentres que a versión lixeira CondMLP acadou unha precisión

*Este traballo é unha adaptación do proxecto de fin de grao do autor.
[1]Todo o material está dispoñible en https://github.com/abrasi/enhanced-nanoGPT

similar cun consumo enerxético e unhas emisións de carbono moito menores. Ao aumentar a escala, os modelos da literatura melloraron a precisión pero a costa dun notable incremento no impacto ambiental, mentres que CondMLP mantivo a súa eficiencia sen superar o rendemento do modelo base. As perdas auxiliares amosaron efectos inconsistentes, engadindo nalgúns casos sobrecarga computacional. En conxunto, os resultados indican que o rendemento das arquitecturas MoE depende dunha interacción complexa entre o tamaño do modelo, a estratexia de enrutamento e os mecanismos auxiliares; o escalado por si só non garante melloras, e enfoques lixeiros como CondMLP poñen de manifesto compromisos relevantes baixo os principios da IA Verde. Na Figura 1 pódese observar un resumo dos resultados obtidos, comparando precisións HellaSwag e emisións de carbono entre modelos.

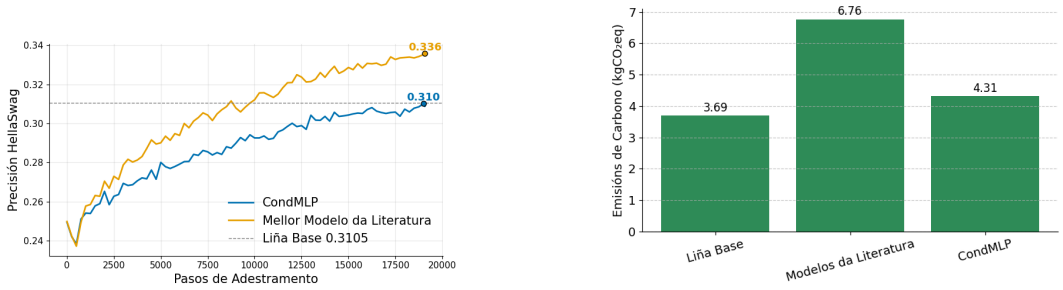

Figura 1: Comparativa de valores HellaSwag e emisións de carbono.

3. Conclusións

Este estudo demostra que as arquitecturas MoE poden superar os modelos Transformer estándar en termos de precisión, pero tamén cuestiona a idea de que reduzan de maneira inherente o custo computacional. A variante proposta CondMLP preséntase como unha alternativa lixeira prometedora, capaz de ofrecer unha eficiencia enerxética notablemente maior á das demais implementacións mantendo unha precisión comparable. Os resultados resaltan a importancia de atopar un equilibrio entre capacidade do modelo, complexidade do enrutamento e impacto ambiental ao deseñar sistemas MoE escalables.

Referencias

[1] W. Fedus, B. Zoph, and N. Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. 2022.

[2] L. Lannelongue, J. Grealey, and M. Inouye. Green algorithms: Quantifying the carbon footprint of computation. 2021.

[3] D. Lepikhin, H. Lee, Y. Xu, D. Chen, O. Firat, Y. Huang, M. Krikun, N. Shazeer, and Z. Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. 2020.

[4] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. 2017.

[5] J. You, J.-W. Chung, and M. Chowdhury. Zeus: Understanding and optimizing GPU energy consumption of DNN training. 2023.

[6] R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, and Y. Choi. Hellaswag: Can a machine really finish your sentence? 2019.