Resumo do artigo "OLMo: Accelerating the Science of Language Models"

Diogo Matos, Gabriel Lima, Pedro Gomes

13 de março de 2024

• Language Modeling • NLP • OLMo

Resumo

Este documento apresenta um resumo do artigo intitulado "OLMo: Accelerating the Science of Language Models" [1]. Este artigo técnico apresenta o OLMo como um avanço significativo, representando um modelo de linguagem. Diferente de abordagens anteriores, o OLMo não apenas fornece os pesos do modelo e código de inferência, mas disponibiliza também o conjunto completo de dados de treino, juntamente com códigos para treino e avaliação. Esta abordagem visa promover a transparência na pesquisa de processamento de linguagem natural.

Análise do artigo

O artigo [1] está dividido em 7 secções. Na secção 1, a *Introdução*, destaca a importância dos modelos de linguagem na área de Processamento de Linguagem Natural (PLN) ao longo dos anos. Ela aponta para a crescente comercialização desses modelos, que muitas vezes são limitados por interfaces proprietárias, deixando detalhes cruciais ocultos. Nesse contexto, o projeto OLMo é apresentado como uma iniciativa crucial, oferecendo acesso aberto completo a modelos de linguagem para a comunidade de pesquisa. O OLMo fornece uma estrutura abrangente, desde dados de treino até ferramentas de avaliação, com ênfase na transparência e abertura. Comparado a outros lançamentos recentes, destaca-se por disponibilizar não apenas os pesos do modelo, mas também o código de treino, logs e conjuntos de dados usados. O projeto visa encurtar a distância entre modelos menos acessíveis e aqueles de ponta, promovendo avanços científicos e melhorias práticas nos modelos de linguagem. O primeiro lançamento inclui várias variantes do modelo em diferentes escalas, além de recursos para construção, análise e avaliação de conjuntos de dados. O OLMo é apresentado como o primeiro passo em uma série planeada de lançamentos, com o objetivo de estimular a pesquisa em áreas ainda pouco compreendidas desses modelos.

A segunda secção *OLMo Framework*, compreende os modelos OLMo, o conjunto de dados de pré-treino chamado Dolma e a estrutura de avaliação. O modelo adota uma arquitetura de transformer, com variantes de 1B e 7B, e uma versão de 65B em desenvolvimento. Algumas melhorias incluem a exclusão de viés, uso de norma de camada não paramétrica, função de ativação SwiGLU, embeddings posicionais rotativos (RoPE) e um vocabulário modificado. O conjunto de dados Dolma é apresentado como um esforço para disponibilizar conjuntos de treino, composto por 3 trilhões de tokens de cinco bilhões de documentos de sete fontes acessíveis ao público. A avaliação do modelo ocorre em dois estágios: online para decisões de design e offline para avaliação de checkpoints. A ferramenta de avaliação Catwalk é utilizada para avaliação downstream e perplexidade, com ablações durante o treino para ajustes contínuos. O texto destaca a importância de modelos abertos para avançar na compreensão e aprimoramento de modelos de linguagem, prometendo futuros lançamentos com modelos maiores e mais funcionalidades.

A terceira secção, *Training OLMo*, são detalhados os elementos do processo de pré-treino do framework OLMo, abordada a estrutura de treino distribuído, configurações do otimizador, preparação de dados e hardware utilizado. O treino dos modelos é realizado utilizando a estratégia otimizadora ZeRO através do framework FSDP do PyTorch, que reduz o consumo de memória ao distribuir os pesos do modelo e o estado do otimizador pelos GPUs. Isso permite um tamanho de micro-lote de 4096 tokens por GPU em hardware específico (3.1). O otimizador escolhido foi o AdamW, com hiperparâmetros específicos para cada tamanho de modelo. A aprendizagem é iniciada gradualmente nos primeiros 5000 passos e, em seguida, decai linearmente (3.2). O conjunto de treino é derivado de uma amostra de 2T tokens do conjunto de dados aberto Dolma, apresentado na Secção 2.2. Os tokens são concatenados em instâncias de treino, que são baralhadas da mesma forma em cada execução de treino (3.3). O treino foi conduzido em dois clusters distintos, um com GPUs AMD e outro com GPUs NVIDIA, garantindo compatibilidade sem perda de desempenho. Os detalhes do hardware incluem o supercomputador LUMI, fornecendo até 256 nós com GPUs AMD MI250X, e o cluster MosaicML, composto por 27 nós com GPUs NVIDIA A100 (3.4). O treino é também otimizado para mistura de precisão, equilibrando a estabilidade e eficiência computacional. O texto destaca a consistência de desempenho entre as execuções em hardware diferente.

A quarta secção, *Results*, o ponto de controlo utilizado para avaliar o OLMo-7B é treinado até atingir 2,46 trilhões de tokens no conjunto de dados Dolma (Soldaini et al., 2024), seguindo um cronograma de decaimento linear da taxa de aprendizagem mencionado na Secção 3.2. Nas experiências, foi observado que ajustar ainda mais o ponto de controlo no conjunto de dados Dolma por 1000 passos, com uma decaída linear da taxa de aprendizagem até atingir 0, melhora o desempenho do modelo em perplexidade e nas avaliações de tarefas finais descritas na Secção 2.3. Foram realizadas comparações entre o OLMo e outros modelos publicamente disponíveis, incluindo LLaMA-7B, LLaMA2-7B, MPT-7B, Pythia-6.9B, Falcon-7B e RPJ-INCITE-7B.

A quinta secção, *Artifacts Released*, foi disponibilizado uma variedade de recursos abertos em todas as fases do processo, visando estimular a pesquisa aberta e reduzir esforços duplicados. Isso inclui o código de treino e modelagem, os pesos dos modelos treinados (7B, 7B-twin-2T e 1B), os dados de treino Dolma, ferramentas para construção de novos conjuntos de dados (Dolma Toolkit e WIMBD para análise), e códigos de avaliação (Catwalk para avaliação downstream e Paloma para avaliação baseada em perplexidade). Foi planeado futuras divulgações com registros de treino, ablações, descobertas, registros Weights Biases, e uma versão adaptada do OLMo com ajuste de instruções e RLHF, incluindo código e dados de treino e avaliação utilizando a biblioteca Open Instruct.

A sexta secção *License*, foi decidido adotar licenças permissivas, como a Licença Apache 2.0, para proporcionar flexibilidade aos utilizadores na utilização dos recursos e artefactos desenvolvidos. O objetivo é facilitar o progresso científico e capacitar a comunidade científica. Todo o código e os pesos dos modelos foram disponibilizados sob a licença, permitindo aos utilizadores adaptar e utilizar os resultados dos modelos para treinar sistemas de inteligência artificial ou aprendizagem de máquina. Acredita-se que esta abordagem mais aberta é a melhor opção para promover a colaboração e o avanço na compreensão dos modelos de linguagem. O risco de uso indevido dos modelos é considerado relativamente baixo, dada a utilização predominante como artefactos científicos, não como produtos amplamente adotados pelo público em geral. Essa decisão é respaldada pela observação de que outros modelos semelhantes foram lançados recentemente com licenças permissivas, contribuindo para um ambiente de pesquisa mais aberto.

O artigo conclui com a secção *Conclusion and Future Work*, uma pequena síntese do que foi previamente apresentado. É reiterado que OLMo é um modelo de linguagem de última geração, acompanhado de seu framework para estudo e construção da modelagem linguística. Algo que o diferencia de abordagens anteriores, que apenas disponibilizavam os pesos do modelo e o código de inferência, pois o modelo em questão é disponibilizado juntamente com todo o framework, incluindo dados de treino, código de treino e avaliação. É informada a existência de planos para a disponibilização de recursos adicionais, como logs de treino e novos achados. Além disso, estão a ser exploradas adaptações do OLMo, com a intenção de disponibilizar modelos adaptados, bem como também o código correspondente. O objetivo principal é apoiar a comunidade de pesquisa aberta, oferecendo diferentes tamanhos de modelo, conjuntos de dados, medidas de segurança e avaliações para o OLMo, com o objetivo de fortalecer a inovação na comunidade de pesquisa aberta.

Considerações finais

Ao analisar este artigo, chegou-se a conclusão de que o mesmo representa uma contribuição significativa para o campo dos modelos de linguagem. Ao disponibilizar não apenas os pesos do modelo, mas também o conjunto completo de dados de treino, juntamente com códigos para treino e avaliação, o OLMo se destaca como um passo crucial na promoção da transparência na pesquisa de processamento de linguagem natural. A abordagem adotada neste projeto é especialmente relevante, pois oferece acesso aberto completo à tecnologia, preenchendo uma lacuna onde muitas vezes detalhes cruciais são ocultados. A disponibilidade de variantes do modelo demonstra um esforço notável para encurtar a distância entre modelos menos acessíveis e modelos de ponta, estimulando avanços científicos e melhorias práticas. O OLMo, com sua estrutura abrangente e foco na transparência, certamente impulsionará a pesquisa em áreas ainda pouco compreendidas dos modelos de linguagem, fortalecendo assim a inovação tecnológica.

Contribuição do grupo

O Trabalho foi elaborado de maneira equiparada pelos membros do grupo.

Referências

[1] Groeneveld, D., Beltagy, I., Walsh, P., Bhagia, P., Kinney, R., Tafjord, O., Jha, A.H., Ivison, H., Magnusson, I., Wang, Y., Arora, S., Authur, D.A.R., Chandu, K.R., Cohan, A., Dumas, J., Elazar, Y., Gu, Y., Hessel, J., Khot, T., Merrill, W., Morrison, J., Naik, N.M.A., Nam, C., Peters, M.E., Pyatkin, V., Ravichander, A., Schwenk, D., Shah, S., Smith, W., Strubell, E., Subramani, N., Wortsman, M., Dasigi, P., Lambert, N., Richardson, K., Zettlemoyer, L., Dodge, J., Lo, K., Soldaini, L., Smith, S.A., Hajishirzi, H., "OLMo: Accelerating the Science of Language Models" *arxiv* https://arxiv.org/html/2402.00838v1. Last accessed 10 March 2024.