Processamento de Linguagem Natural com Transformers

Arquitetura Unified Transformer (UniT)

Unified Transformer (UniT) é uma arquitetura de modelo que combina as capacidades dos Transformers voltados para texto e para imagem em uma única arquitetura unificada. Isso permite que o modelo processe e entenda conjuntamente o texto e a imagem, que é útil para uma variedade de tarefas, como geração de legenda de imagem, perguntas e respostas visuais, e muito mais.

A arquitetura UniT foi apresentada em um trabalho chamado "UniT: Multimodal Multitask Learning with a Unified Transformer" pelos pesquisadores da Microsoft em 2021.

https://arxiv.org/abs/2102.10772

Nesta arquitetura, tanto o texto quanto a imagem são codificados em tokens e alimentados em uma única rede Transformer. O modelo é treinado em várias tarefas de processamento de texto e imagem para aprender a entender conjuntamente o texto e a imagem.

Aqui estão algumas características-chave do UniT:

Entrada unificada de texto e imagem: A UniT aceita tanto texto quanto imagem como entrada. As imagens são divididas em regiões (normalmente usando algum tipo de rede convolucional como um extrator de recursos) e cada região é tratada como um token, da mesma forma que as palavras no texto são tratadas como tokens.

Codificação de posição unificada: Assim como os Transformers padrão, a UniT usa codificações de posição para manter a informação sobre a ordem dos tokens. No entanto, a UniT também inclui uma codificação de tipo de entrada, que distingue entre tokens de imagem e tokens de texto.

Cabeças de atenção multi-tarefa: A UniT usa um mecanismo de atenção multi-tarefa, que permite que o modelo aprenda diferentes tipos de atenção para diferentes tarefas. Isso significa que o modelo pode aprender a prestar atenção a diferentes partes da entrada, dependendo da tarefa em questão.

Treinamento multi-tarefa: A UniT é treinada em várias tarefas diferentes ao mesmo tempo. Isso inclui tarefas baseadas em texto, tarefas baseadas em imagem e tarefas que envolvem tanto texto quanto imagem.

A arquitetura UniT é um exemplo de como os Transformers podem ser estendidos para lidar com diferentes tipos de dados e várias tarefas ao mesmo tempo, mostrando a flexibilidade e o poder desses modelos.