Processamento de Linguagem Natural com Transformers

Arquitetura Timesformer Para Processamento de Texto em Vídeo

Timesformer é uma arquitetura de rede neural desenvolvida para lidar com o processamento de vídeo. Como o nome sugere, esta arquitetura aplica o paradigma dos Transformers ao domínio do vídeo.

O trabalho "Is Space-Time Attention All You Need for Video Understanding?" apresentado por pesquisadores do Facebook AI introduziu o Timesformer em 2021. A ideia básica é tratar os frames de um vídeo como uma sequência no tempo (semelhante a uma sequência de palavras em PLN), e aplicar a atenção auto-regressiva do Transformer a esta sequência.

https://arxiv.org/abs/2102.05095

Em termos mais detalhados, o Timesformer pega uma série de frames de um vídeo e aplica atenção tanto ao espaço (cada frame individual) quanto ao tempo (a sequência de frames). No entanto, em vez de aplicar a atenção conjuntamente ao espaço-tempo (o que seria computacionalmente caro), o Timesformer divide a atenção em duas etapas separadas: atenção no espaço e atenção no tempo. Isso permite que o modelo lide com vídeos de forma eficiente, mesmo quando o número de frames é grande.

Isso significa que o Timesformer pode aprender relações complexas entre diferentes partes de um frame individual (usando atenção espacial) e entre diferentes frames em uma sequência de vídeo (usando atenção temporal). Isso é útil para uma variedade de tarefas, como classificação de vídeos, detecção de ação e muito mais.

É importante notar que, como em outros Transformers, o Timesformer não depende de convoluções e, portanto, não impõe uma noção predefinida de proximidade espacial ou temporal, ao contrário das arquiteturas baseadas em CNN para vídeos, como 3D-CNNs. Em vez disso, o modelo aprende a decidir por si só quais partes do vídeo são relevantes para a tarefa em questão.