Elton Cardoso do Nascimento 233840

IA024 - Redes Neurais Profundas para Processamento de Linguagem Natural - 1s2024

Leitura do Artigo "Attention is all you need, NIPS 2017"

O artigo apresenta uma nova arquitetura para redes neurais chamada "Transformer". A arquitetura foi desenvolvida se focando em tarefas de processamento textual, tentando superar as arquiteturas baseadas em redes recorrentes ou convolucionais utilizadas na época. A arquitetura é composta de um elemento de encoder, que recebe a entrada da rede; e um de decoder, que recebe a entrada codificada e a saída até o momento, sendo então um modelo auto-regressivo.

Como aspecto central da arquitetura está o mecanismo de auto-atenção, em que a saída é computada a partir de valores ponderados pela compatibilidade entre suas chaves e um elemento de consulta. No caso do trabalho é utilizado uma atenção baseado em produtos internos e com um mecanismo de multi-atenção, onde várias camadas de atenção são executadas em paralelo e os resultados concatenados, permitindo ao modelo explorar diferentes representações possíveis do mesmo dado.

Um outro aspecto implementado é o mecanismo de encoding posicional, onde a posição dos elementos da sentença são indicados adicionando aos embeddings um vetor representando essa a posição dele na sentença. Este vetor pode ser aprendido, ou representado por uma função fixa, no caso uma senóide.

O modelo se mostra mais eficiente computacionalmente do que os modelos recorrentes e convolucionais, tendo também uma capacidade de relacionar elementos em qualquer posição da sentença. Os resultados experimentos mostram que a arquitetura também é mais rápida de ser treinada, encontrando melhores resultados. Entre os fatores que influenciam a perfomance estão a quantidade de cabeças de atenção, o tamanho da dimensão do espaço das chaves, o tamanho em si do modelo, e o drop-out. O uso de aprendizado para obter os vetores posicionais não gerou diferenças significativas.

Por fim, o trabalho aponta como possíveis desdobramentos o uso de multi modalidades no modelo (o que foi de fato realizado em trabalhos posteriores), assim como outros mecanismos de atenção e tornar a geração menos sequencial.