Ciência da Computação / Sistemas de Informação

Utilização de um Modelo Imagem-para-sequência Pré-Treinado Para Extração de Tabelas

João Paulo Paiva Lima - 9° módulo de Ciência da Computação, UFLA, bolsista PIBIC/CNPq

Denilson Alves Perreira - Orientador DCC, UFLA - Orientador(a)

Resumo

Apesar de ser uma opção popular em documentos científicos e técnicos, tabelas geralmente apresentam um problema de compreensão por sistemas automatizados. O grande número de variações em estilo, topologia e conteúdo dificulta a criação de extrações baseadas em regras. Entretanto, a evolução dos modelos de linguagem baseados em aprendizado de máquina na última década abriram novas possibilidades para extração automatizada de tabelas. O objetivo deste trabalho é a utilização de modelos de linguagem pré-treinados para uma nova tarefa, a extração de informações tabulares. Para este fim, utilizamos um modelo imagem-para-sequência pré-treinado para Entendimento de Documentos Visuais e fizemos o ajuste fino no conjunto de dados PubTabNet (que utiliza a codificação em linguagem HTML). Visando extrair o máximo da arquitetura Transformer, criamos também uma nova codificação intermediária, que pode ser convertida de e para HTML sem perda de dados, que reduz o tamanho de sequência, uniformiza a topologia da tabela e permite a utilização de modelos Transformers para tabelas ainda maiores que as tratadas no conjunto de dados. Ao utilizar um modelo pré-treinado, conseguimos resultados competitivos de assertividade em uma fração do tempo de treinamento e, com a utilizando a nova codificação, conseguimos com que o acerto para estrutura de tabelas complexas após uma única época de treinamento passasse de 88,9% para 91,6% TEDS (Tree Edit Distance Score). Podemos, então, concluir que modelos pré-treinados podem ser utilizados com eficácia para extração de tabelas e que a nova codificação proposta, além de prover melhor generalização, pode também auxiliar na convergência e acerto do modelo estatístico.

Palavras-Chave: Extração de Dados, Extração de Tabelas, Processamento de Linguagem

Natural.
Instituição de Fomento: CNPq

Link do pitch: https://youtu.be/l6sBfwafQI4