LayoutLMv2 and LayoutLMv3

LayoutLMv2

LayoutLMv2 adds image information to the model, not just text and where the text is on the page. It uses a special image reader called ResNet to look at the whole document picture and create image features. It also learns to connect the text with the right parts of the image, which helps the model understand documents better than v1, which only used text and layout.

LayoutLMv3

LayoutLMv3 is even simpler and better. It does not use ResNet or any CNN for image features. Instead, it breaks the document image into small patches and learns from those directly, along with the text and layout. It also learns to match words with the right image patches. This makes v3 faster and able to handle both text and images more easily than v1.