TASK 3

01.07.2025

LayoutLM v2 and Layout LM v3

Layout LMv2

1. Multi-Modal Fusion from the Starting itself

- Unlike v1, LayoutLMv2 fuses text + layout + image during pretraining itself.
- This helps the model learn better joint representations early on.

2. Relative Positional Encoding

- v1 just had absolute 2D position embeddings (x0, y0, x1, y1).
- v2 adds a more relational understanding through spatial-aware self-attention so it can understand things like
    - *"Token B is to the right of Token A."*
    - *"This line is below that header."*

3. New Pretraining Objectives

They kept **Masked Visual Language Modeling (MVLM)** from v1, but added two new ones:

- **Text-Image Alignment (TIA):**
  Matches each line of text to its corresponding region in the image. Helps the model learn how text physically aligns on the page. It basically makes a match between the text line tokens and the corresponding visual region

- **Text-Image Matching (TIM):**
  basically text image matching which is a binary task to check if the text is paired with correct image.

Layout LMv3

1. Unified Embedding Space

- All modalities (text, layout, and image) are projected into the same space from the beginning.

- Compared to v2, v3 version uses a shared transformer encoder, so the model doesn't separate branches for each modality.

2. Image Input usign Patch Embeddings

- Instead of using ResNet or similar CNNs like in v2, v3 adopts a **Vision Transformer** approach to process the image.

- The image is divided into patches, just like tokens, and those patches are embedded and passed into the Transformer.

3. Simpler non multi branch Architecture

- v3 makes the model more uniform by image, layout, and text going through the same encoder.