Clean your desk! Transformers for unsupervised clustering of document images

Pooja Sethi [1]

[1]Department of Computer Science, Stanford University

Introduction

Suppose a busy scientist had a mixture of machine learning papers, handwritten notes, and receipts she last ordered takeout from on her desk. Could we organize her documents without any supervision? I explore the use of Transformers that jointly model text, layout (position), and visual features for **unsupervised document clustering**.

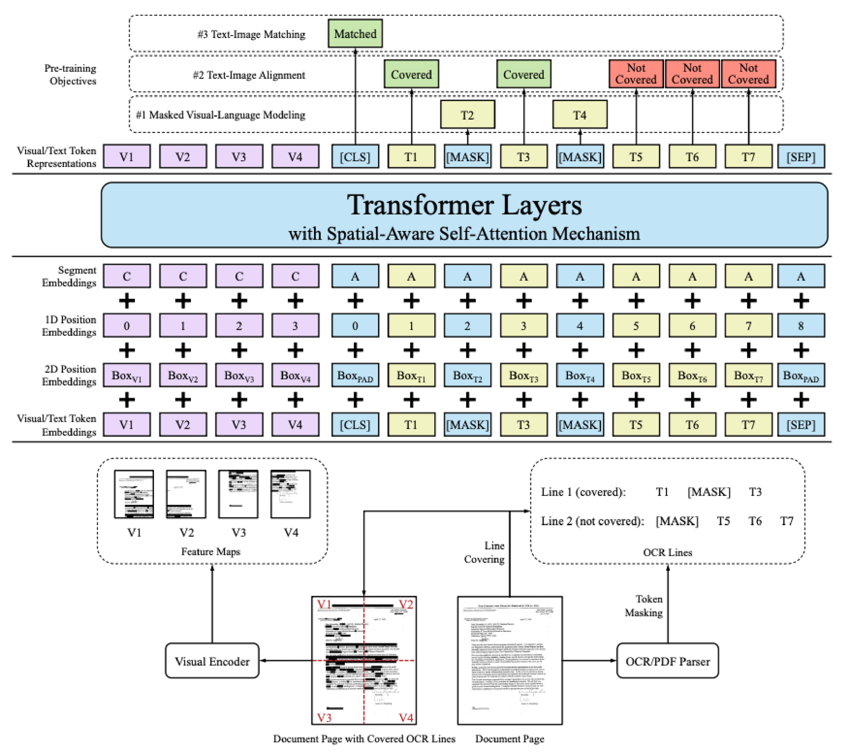


Figure 1. The LayoutLMv2 architecture, taken from [2].

Problem Statement

1. **Document Embedding** Given a document d_i, preprocess it into its constituent text, bounding boxes, and images. Then, using a model m_{encoder}, return its embedding e_{di}.
2. **Document Clustering** Given the embeddings e_{di} for N documents, use a model m_{cluster} to divide them into k clusters, such that each cluster is at least size 1 but no larger than N.

Datasets: Receipts (SROIE), Documents (RVL-CDIP), and Machine Learning Papers

Figure 2. Examples of document images from each of the three datasets. SROIE had 626 images, RVL-CDIP had 1000, and ML Papers had 30. Preprocessing (OCR) via Impira [1] returns words and bounding boxes.

Methods: What's different about BERT, LayoutLM , and LayoutLMv2?

The BERT, LayoutLM [3], and LayoutLMv2 [2] encoders all use the Transformer architecture, with slight differences.

Table 1. The similarities and differences between these encoders are summarized here.

	BERT	LayoutLM	LayoutLMv2
Inputs			
Text Embeddings	✓	✓	✓
Segment Embeddings	✓	✓	✓
1-D Position Embeddings	✓	✓	✓
2-D Position Embeddings		✓	✓
Visual Token Embeddings			✓
Attention			
Self-Attention	✓	✓	
Spatial-Aware Self-Attention			✓
Pretraining Objectives			
Masked Language Modeling (MLM)	✓		
Next Sentence Prediction (NSP)	✓		
Masked Visual Language Modeling (MVLM)		✓	✓
Multi-label Document Classification (MDC)		✓	
Text-Image Alignment (TIA)			✓
Text-Image Matching (TIM)			✓

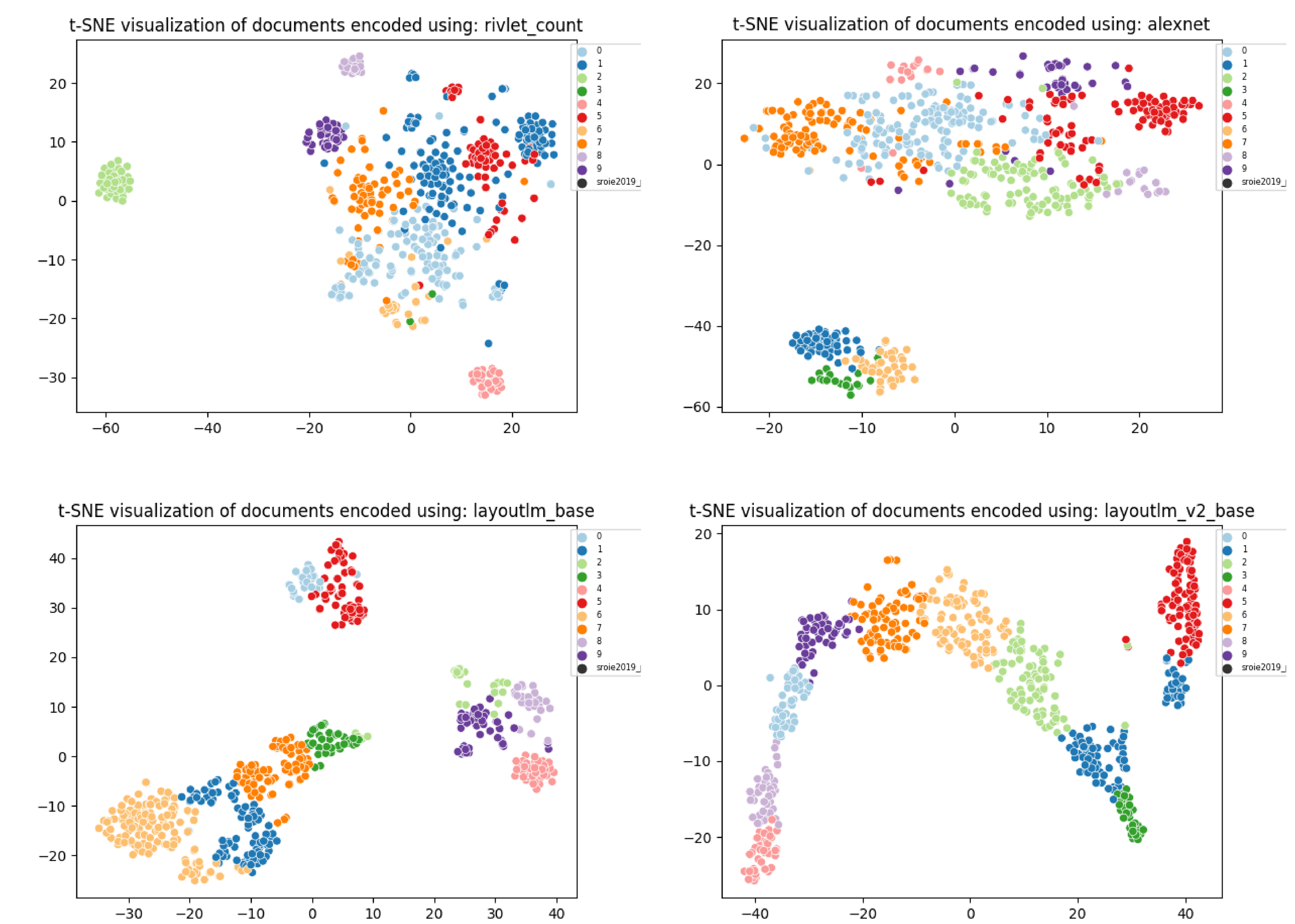
Results: SROIE and RVL-CDIP

Table 2. Unsupervised Clustering Results with No Labels
Metrics: (Silhouette Coefficient / Calinski-Harabasx (CH) score)

Method	SROIE ($n = 626$)	RVL-CDIP ($n = 1000$)
Baselines		
Bag-of-Words (BoW)	0.093 / 27.5	0.134 / 90.3
ResNet-18	0.101 / 69.888	0.047 / 57.977
AlexNet	0.116 / 75.764	0.070 / 69.955
LayoutLM Base		
`[CLS]` token	0.155 / 88.406	0.155 / 107.911
`[SEP]` token	0.194 / 378.588	**0.248** / 118.642
Average all tokens	0.128 / 41.326	0.055 / 42.778
LayoutLM Large		
`[SEP]` token	0.156 / 44.062	0.057 / 36.746
LayoutLMv2 Base		
`[CLS]` token	0.127 / 212.757	0.131 / 223.778
`[SEP]` token	0.150 / 84.804	0.091 / 126.566
Average image tokens	**0.281** / **713.625**	0.187 / **666.915**
Average all tokens	0.177 / 366.174	0.130 / 262.591
LayoutLMv2 Large		
Average image tokens	0.079 / 78.228	0.056 / 96.490

LayoutLMv2 was *typically* the best performing model, although text-heavy documents may still be better off with LayoutLM. The `[CLS]` output was not the best representation.

Analysis: Visualizing Cluster Predictions



These t-SNE plots show document embeddings from four m_{encoder} models on the SROIE dataset. In quadrant order: Bag of Words (BoW), AlexNet, and LayoutLM (Base), and LayoutLMv2 (Base).

Conclusions

- Multi-modal learning of document representations, using text, positional, and visual features, is beneficial. LayoutLMv2 typically outperforms LayoutLM. However, LayoutLM may still be a better choice for text-heavy documents.
- The `[CLS]` token output may not always be the best choice of document representation. The `[SEP]` token output and the average of the image token outputs performed better on SROIE and RVL-CDIP.

Future Work

- Does finetuning on domain-specific datasets improve the `[CLS]` representation?
- Can we learn to mask out less important details of the document?

References

[1] Impira.
Available at `https://impira.com`.

[2] Yang Xu et al.
Layoutlmv2: Multi-modal pre-training for visually-rich document understanding.
In *Association for Computational Linguistics (ACL)*, 2021.

[3] Yiheng Xu et al.
Layoutlm: Pre-training of text and layout for document image understanding.
In *Association for Computing Machinery (ACM)*, 2020.