Learning to Cluster:

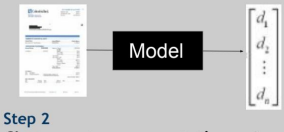
A Comparison of Document Vector Representations for Layout Identification

Bryan Chia & Pooja Sethi

Problem

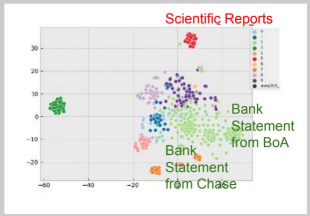
Step 1

Represent documents as ***vector representations***



Step 2

Cluster vector representations using k-means unsupervised clustering. Documents with the same category and origin should be most tightly clustered.



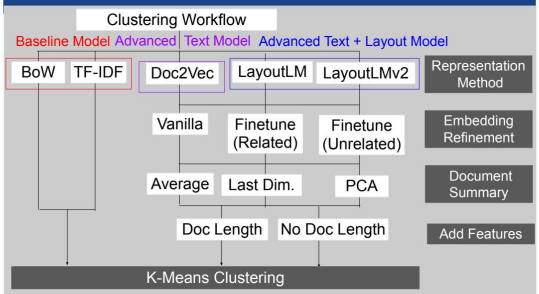
Goal

Optimize ***Silhouette coefficient*** *[-1,1]* or ***Caliski-Harabasx coefficient*** *[0, ∞)* which reward within-cluster tightness and between-cluster distance

Data for Clustering

- ***RVL-CDIP***: 16 types of documents including emails, advertisements, etc.
- ***SROIE***: Various types of receipts

Methods



Results

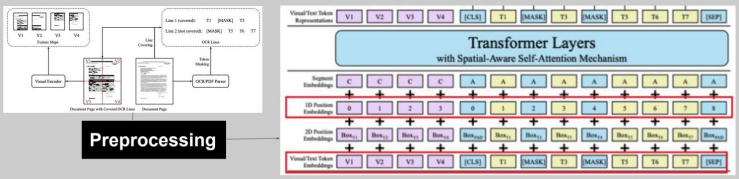
Silhouette Coefficient / Caliski-Harabasx Coefficient Scores

Table 2: Layout Identification Results (Silhouette / Caliski-Harabasx Coefficients)

Method	SROIE ($n_t = 626$)	RVL-CDIF ($n_t = 1000$)
Baselines		
Bag of Words (BoW)	0.09 / **27.53**	**0.134 / 90.309**
TF-IDF	**0.103** / 16.900	0.003 / 5.656
Vanilla LayoutLM		
Average all words	0.186 / 406.4	0.371 / 1214.2
Average all words, mask pads	0.436 / 2262.7	0.497 / 8525.4
Average all words, mask pads, append length	**0.536** / 3172.3	0.664 / **20513.6**
Average all words, mask pads, append & normalize length	0.436 / 2265.0	0.497 / 8542.2
Last word, append length	**0.536 / 3172.9**	**0.665** / 20442.5
PCA on all words, mask pads, append length	0.460 / 2412.6	0.576 / 18493.6
Finetuned LayoutLM on Related Task (RVL-CDIP)		
Average all words	0.253 / 272.0	0.229 / 388.2
Average all words, mask pads, append length	**0.534 / 3165.2**	**0.660 / 20405.5**
Last word, append length	0.524 / 3094.6	0.654 / 20093.0
Finetuned LayoutLM on Unrelated Task (FUNSD)		
Average all words	0.229 / 388.19	0.253 / 272.02
Average all words, mask pads, append length	**0.660 / 20405.5**	**0.534 / 3165.2**
Last word, append length	0.654 / 20093.0	0.524 / 3094.6
Vanilla LayoutLMv2		
Average all words	0.163 / 115.83	0.113 / 177.1
Average all words, mask pads, append length	**0.524 / 2887.8**	**0.652 / 20779.4**
Last word, append length	0.517 / 2858.7	0.646 / 20615.7

Bold numbers indicate the best performance in each category

Background



Neural Architecture of ***LayoutLM*** vs. ***LayoutLMv2*** Models in Pretraining:

LayoutLMv2 changes shown in red

- LayoutLM: BERT Word & 2D Position Embeddings
- LayoutLMv2: BERT Word, 2D Position & CNN Visual Encodings used as Image Embedding

Analysis

- ***LayoutLM*** clearly outperforms baseline BoW or TF-IDF models
- ***Masking out pads*** and ***appending the document length*** as a feature plays an important role
- ***Fine-tuning on a related task*** does not bring discernable benefits, and could even hurt by overfitting on the simpler supervised task it was trained on.
- ***Fine-tuning on an unrelated task*** performs the same as the vanilla model as it does not overfit.
- ***Vanilla LayoutLMv2 is inferior to LayoutLM***, especially for high number of clusters