RVL-CDIP : IndoML 2022

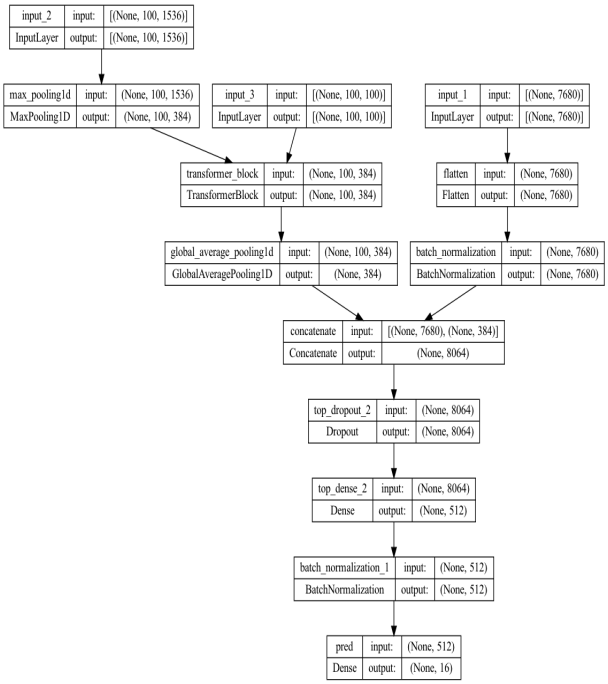
Document Image Classification: Novel RoI Vision Transformer Model

Rishi Dey Chowdhury[1] **Mrinmoy Banik**[2]

1. Affiliation rishi8001100192@gmail.com
2. Affiliation rinmoybanik12@gmail.com

Abstract

Image classification has been one of the leading areas of research and here we have explored both traditional ways using pre-trained CNN models to derive feature vectors from and also some less traditional ones like vision transformer network on a larger set of ROI(Region of Interest) snippets extracted from those document images.But as it turns out, the former methods yields slightly superior results than the latter ones, probably due to constraints on dataset size and computation, we were obliged to follow as this being a kaggle competition.



RoI Vision Transformer + Multi-Part CNN Network based on InceptionResNetV2's 1536 dimensional vectors.

Overview of the Multi-Part CNN Network

We partitioned each document image into many non-overlapping and mutually exhaustive sections, and passed them through a pre-trained CNN model like ResNet152, ResNetV2, InceptionResNetV2, VGG16 and EfficientNetV2L to extract the features. Then we further concatenated these feature vectors together and trained a custom MLP layer to predict the final class. We also experimented image augmentation but they don't seem to make much of an improvement.

Multipart CNN Results

6+1-Part Model: With the whole document as a single piece we were only able to get a mean-f1 score of around 0.6 with EfficientNetV2L. Next with the rescaled whole image(900x600) divided into a 3 by 2 grid or 6 parts(300x300) we were able to increase the score to 0.7 and we got the best results with VGG19 as the feature extractor of each part with a feature dimension of 16 each and the classifier had 2 dense layers with 512 nodes each.

4+1-Part Model: Here we divide the image into 4 pieces and extracted the feature vector from each namely, header (250x600), body left (300x300) , body right (300x300) , footer(250x600) and also concatenated the feature vector from the entire image resized to a 800x600 array. To get the final results we concatenated features from multiple extractors like ResNet152 ,VGG16 & InceptionResNetV2 of each piece. Then to reduce dimension we used PCA for each of these feature vectors on each part to get a final 2500 dimensional feature vector for each part and finally trained 2 dense layers of 512 units and 16 units with softmax to classify the doc type. We got a mean f1-score of 0.785 in this way.

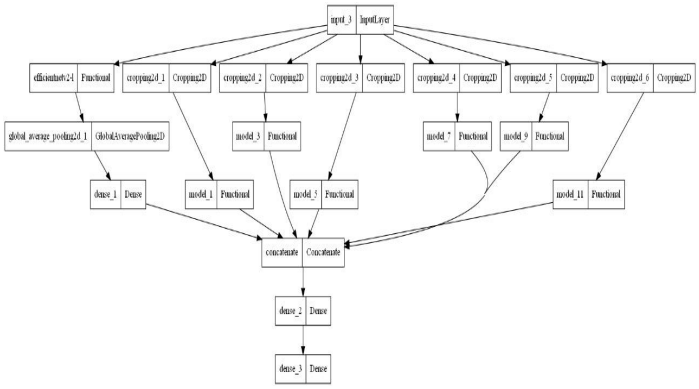
ROI Vision Transformer

We used a pre-trained Layout Parser to identify regions which correspond to text, figures, etc. namely a mask RCNN model. After the regions (we will call them RoI i.e. Region of Interest) were identified, they were extracted and since each RoI are of varying shape each of them were resized and padded to be 400x300 dimension. Now, these RoIs are passed through InceptionResNetV2 to generate 1536 dimensional feature vectors corresponding to each RoI. Now, since the number of RoIs per document is also

varying we padded the number of RoIs to a 100 for each document and maintained a mask for which RoIs are non-padding and which RoIs are padding.

Now, we also have the 4+1-Part Model's feature vectors for InceptionResNetV2 for the 4 parts of the image and the entire image. We arrange the RoIs in top-to-down manner, taking inspiration from how most of the humans read a document i.e. top to bottom with more focus on heading, then body and so on. So, to keep the sequential nature of the RoI's feature vector intact we add sinusoidal positional encoding to the feature vectors. Then, they are passed to the Transformer's Encoder block along with the mask which computes new features after a MultiHeaded Self-Attention and few dense layers. The final output of the transformer is AveragePooled and then concatenated with the 4+1 feature vectors for each part of the image and passed through 2 more Dense Layers to produce softmax output of 16 units for each of the labels.

Figures and Tables



This is one of the model structure of the 6+1-part CNN model which got us mean-f1 scores of around 0.7

Model Used	Mean-F1 Score*	Remark
Whole Part CNN 900x600	~0.6	EfficientNetV2L and 2 new dense layers added
6-Part CNN Classifier (300x300) each	~0.7	VGG19 & a classifier with 2 hidden layers.
4-Part CNN Classifier	~0.785	ResNet152,VGG16,InceptionResNetv2) along with PCA & finally 2 layer MLP
ROI based Vision Transformer	~0.755	InceptionResNetV2 + Transformer Block + 2 MLP Layer classifier.

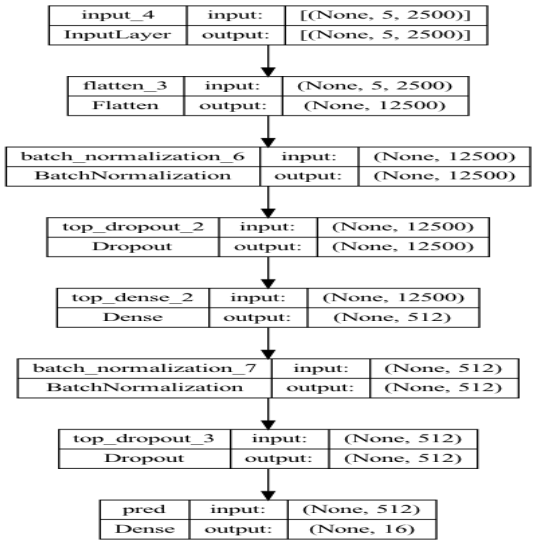
*only approximate maximum obtained score after submission are given

Novelty

Our models are results of weeks long experiments and experience on image classification on the given subset of RVL-CDIP dataset. The Novelty in our method lies in the fact that instead of averaging all the RoI's extracted features which is most commonly observed in various research papers, we used a Vision Transformer to take into account the sequential nature of the appearance of the RoIs. Transformer Models are big and are quite data hungry. Given the small dataset of 16,000 images this is the maximum we got from the Vision Transformer Network. Due to computational and time constraints of the competition we were unable to test various combination of pretrained CNN-Based features apart from InceptioResNetV2 as well as various other interesting form of attention mechanisms i.e. Relation Positional Embedding, etc. But the results are already quite promising with this Basic RoI Vision Transformer Network.

Conclusion

The Vision Transformer turns out to be the most promising if combined with various feature extractors and attention mechanisms.



4+1-part CNN mode using PCA on concatenated multiple feature vector for each part i.e. from ResNetV2, VGG16 and InceptionResNetV2 which got us mean-f1 scores of around 0.785

Github Link for the code is available at https://github.com/RishiDarkDevil/Document-Classification
Kaggle Competition Link: https://www.kaggle.com/competitions/datathonindoml-2022