ABSTRACT:

A region-based Deep Convolutional Neural Network framework is provided in this paper for learning document structure. Effective region-based classifier training and ensembling for document image classification are two key contributions of this work. By considering the RVL-CDIP dataset as a reference and the train_csv file, the 16,000 images are classified into 16 classes (letter, email, invoice, advertisement, etc). Hence the images are categorized based on their class label. Applying common Convolutional Network architectures with DeepDocClassifier. Exporting weights from model architectures that have already been trained on the dataset is used to train a document classifier on complete document images, which is the first level of "inter-domain" transfer learning. Utilizing a region's unique characteristics to quickly train deep learning, transfer learning is utilized. Now shuffle the categorised images and split them into train and test. Now, this test and train are applied to each of the following individual models (document-image-transfer (DIT), VGG-16, LILT-only-base, LayoutLMv2) and store their predicted values in P1, P2, P3, and P4 respectively. These models are integrated to get a new model using multi-model integration and predicted values (P1, P2, P3, and P4) are stacked into a 1D array using stacked generalization with ML. Now the final predicted values are validated with the validation set to find the performance metrics.

LAYOUT:

INTRODUCTION:

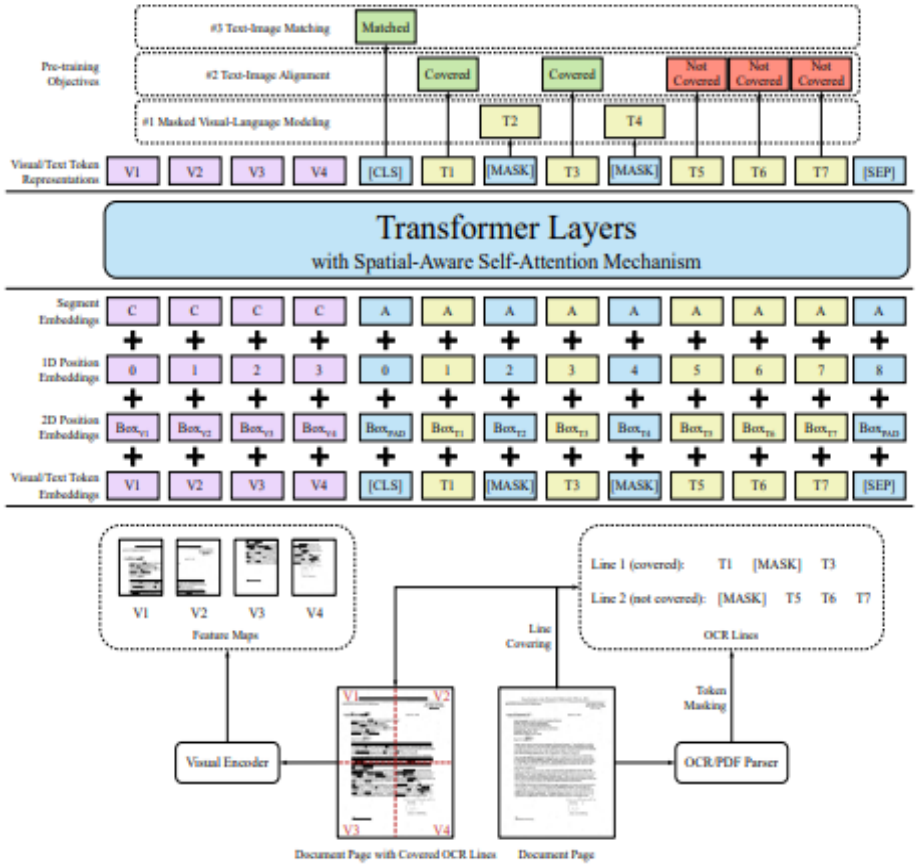
Due to its efficient model design and the benefit of large-scale unlabeled scanned/digital documents, pre-training of text and layout has demonstrated effectiveness in a range of visually complex document interpretation tasks. To simulate the interaction between text, layout, and image in a single multi-modal framework, we suggest the LayoutLMv2 architecture with novel pre-training tasks. In particular, LayoutLMv2 uses a two-stream multi-modal Transformer encoder that not only the new text-image alignment and text-image matching tasks but as well as the current masked visual-language modelling job, improves the pre-training stage's ability to capture the interplay between different modes. The Transformer architecture also has a spatial-aware self-attention mechanism so that the model may completely comprehend the relationship between various text blocks' relative positions. In the document-level classification task RVL-CDIP, we employ the [CLS] output along with a pooled representation of visual tokens as global features.

Fine-tuning for Document Image Classification:

Since this task requires detailed visual information, we deliberately use picture characteristics during the fine-tuning phase. We combine the visual embeddings into a pre-encoder feature that is available globally. LayoutLMv2's visual component outputs representations as a global post-encoder feature. the prior and post-encoder capabilities in addition to the [CLS] Concatenated output features are fed into the last layer of categorization

CONTRIBUTIONS:

1. As new pre-training tactics to enforce the alignment among various modalities, we also include text-image alignment and text-image matching in addition to the masked visual-language model.
2. In the pre-training step, we suggest using a multi-modal Transformer model to integrate the document's text, layout, and visual information. This integrates end-to-end learning of the cross-modal interaction into a single framework. Meanwhile, A self-attention system with spatial awareness in the Transformer architecture has it included.
3. LayoutLMv2 greatly outperforms traditional visually-rich document understanding tasks (VrDU) as well as the visual question answering (VQA) and reaches new outcomes. A job for document photos, demonstrating the enormous potential for the VrDU's pre-training.
   •



An illustration of the model architecture and pre-training strategies for LayoutLMv2

DiT:

Introduction:

DiT is a self-supervised document image transformer model that was trained using massive amounts of unlabelled text pictures. Due to the lack of human-labelled documents, supervised analogs never exist for jobs involving document AI. Analysis of document layouts or optical Character Recognition (OCR) still significantly relies on supervised computer vision backbone models using human-labelled training data. With either supervised pre-training on ImageNet or self-supervised pre-training, Image Transformer has recently shown remarkable success with natural image understanding tasks like classification, detection, and segmentation. The results that pre-trained Transformer models can produce compared to CNN-based pre-trained models with a similar parameter size, and comparable and even superior performance was achieved.

For generic Document AI tasks, we provide DiT, a self-supervised pre-trained Document Image Transformer model that doesn't rely on any human-labelled document images. The DiT model simply uses large-scale unlabeled data to understand the global patch relationship within each document image and does not rely on any human-labelled document images. We assess the previously trained data using four openly accessible Document AI benchmarks,

Dit models comprise the dataset for document layout analysis and the RVL-CDIP dataset for document image classification. According to experiment results, the pre-trained DiT model outperformed the existing supervised and self-supervised pre-trained models, and new state-of-the-art was attained on these tasks.

CONTRIBUTIONS:

- We suggest DiT, a self-supervised pre-trained document image Transformer model that can benefit from massive amounts of unlabeled document photos.
- We use the previously trained DiT models as the foundation for different Document AI tasks, such as document image classification, analysis of document layout, table detection, as well as and text identification for OCR, and create brand-new cutting-edge outcomes.

The model architecture of DiT with MIM pre-training.

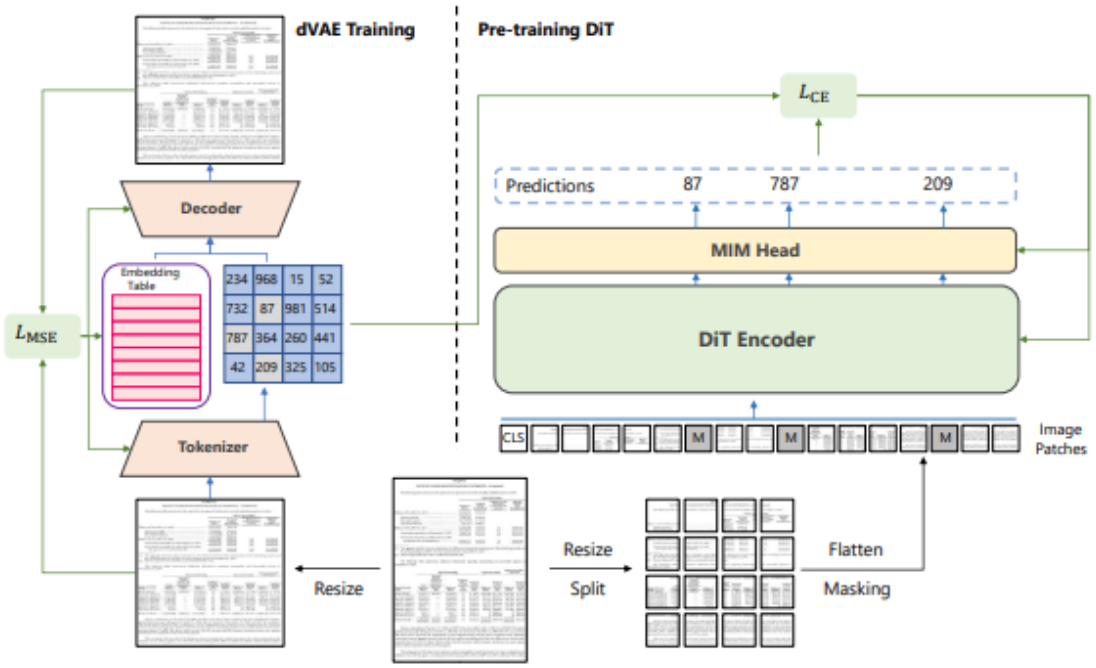
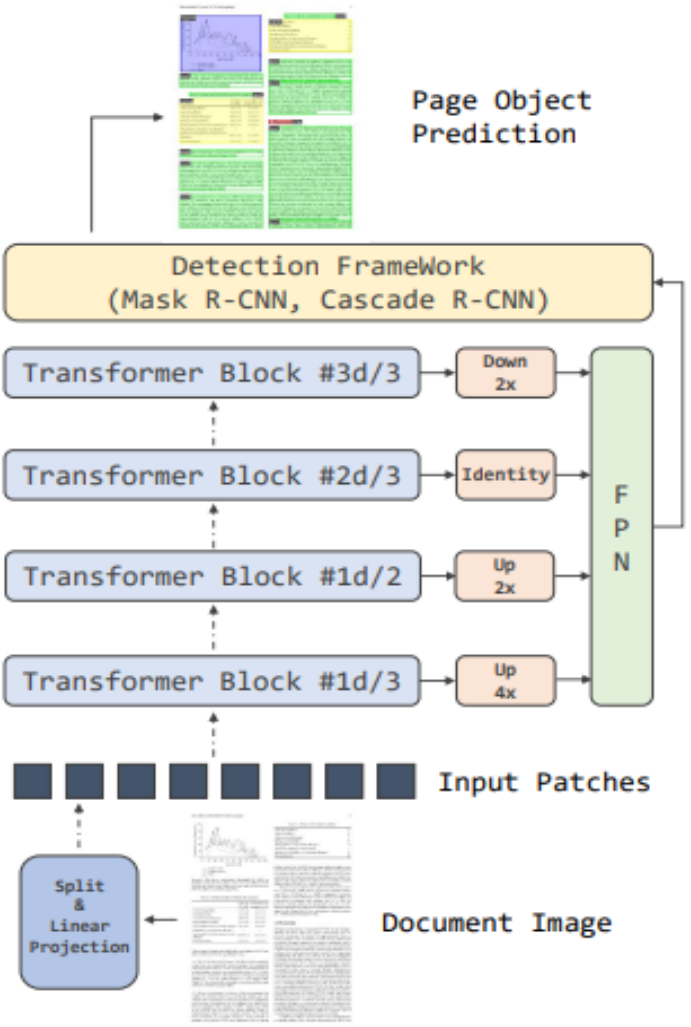


Illustration of applying DiT as the backbone network in different detection frameworks.



Inception-ResNet:

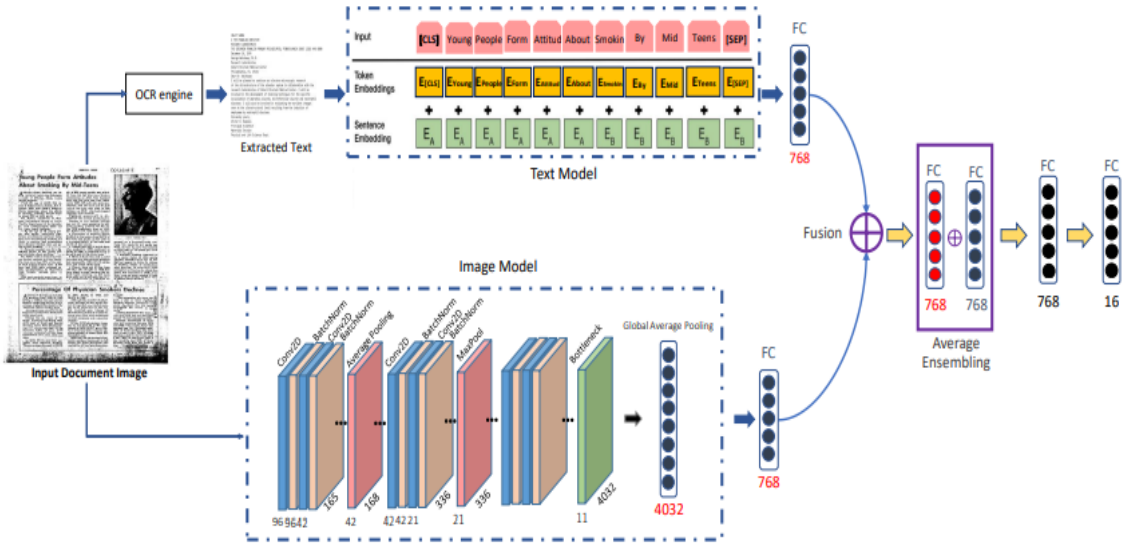
Introduction:

Document images, as opposed to general photographs, can be displayed in a variety of ways because each document is organized differently. However, given their visual structural characteristics and their textual varied content, it is exceedingly difficult to extract correct and structured information from the large diversity of documents. Modern approaches based on visual data of document images tackle the issue as a standard image classification problem because the majority of recent deep learning methods do not involve manually extracting features. The issue of limited inter-class discrimination and substantial intra-class structural variability of highly overlapping document images may arise when classifying documents only based on visual information. We suggest a cross-modal network to learn concurrently from the textual information in document images and the visual structural characteristics. The cross-modal properties that were learned include merged as the final representation of our suggested network to improve the ability of document image categorization. We examine the effects of both heavyweight (i.e., with a lot of parameters) and lightweight (i.e., with a lot fewer parameters) on the picture classification problem. Deep neural network architectures for discovering deep structural characteristics from document images (of parameters). The Inception-ResNet-v2, a heavyweight model with significant size parameters, can reach precision in classification that is modern. Convolutional neural network Inception-ResNet-v2 produced cutting-edge performance on the ILSVRC image classification benchmark. By incorporating the bypass link as in Inception-ResNet-v2, the older Inception V3 model is modified.

CONTRIBUTIONS:

- To categorize the textual content of document photographs, we assess the effectiveness of static and dynamic word embeddings.
- Here, we examine how deep neural networks trained with different weights perform while learning pertinent information and structural data from photos of documents.
- To categorize document pictures, we suggest a cross-modal deep network that makes use of both visual and textual data. We demonstrate that, in comparison to single-modal networks, the joint learning methodology increases overall accuracy.

- **The proposed cross-modal deep networt**



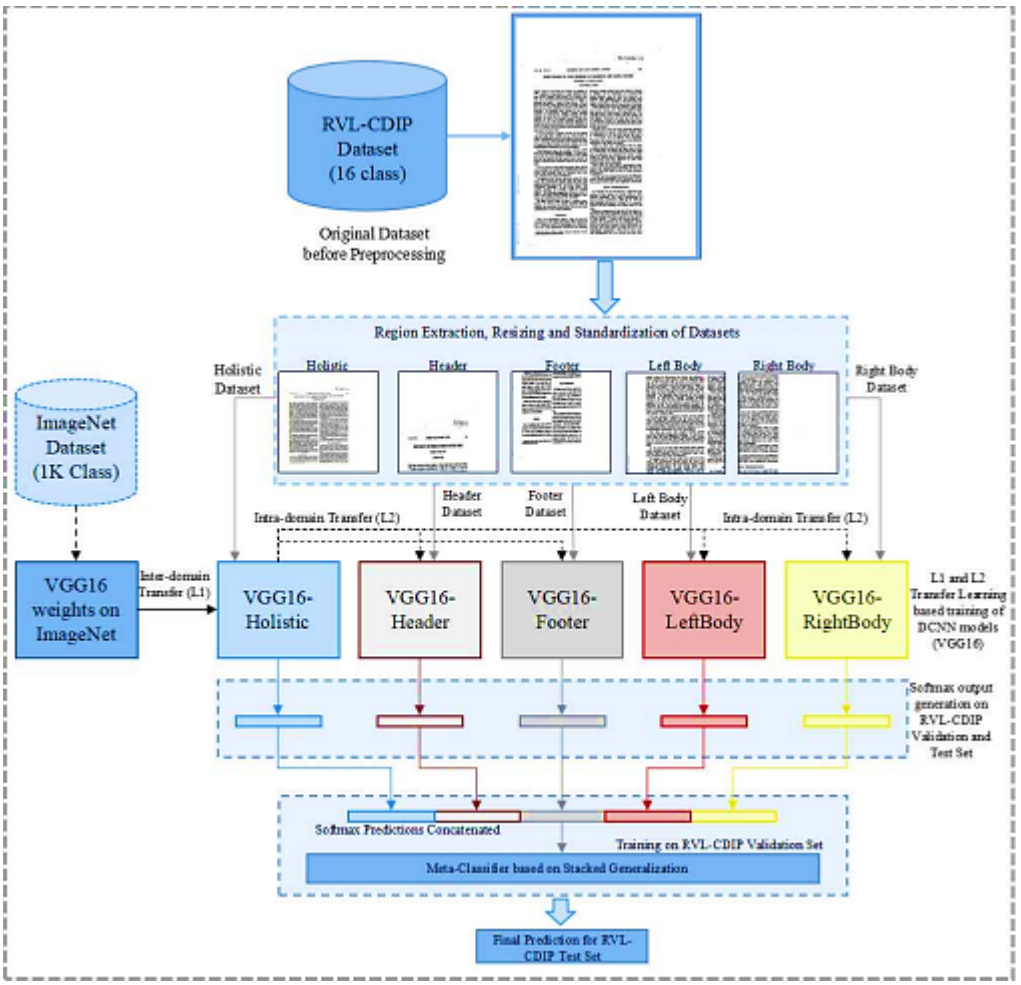
VGG19:

INTRODUCTION:

Since sufficient labelled data must be available for deep learning models to train on. A machine learning-based model for a system is therefore almost impossible to develop. Only a little amount of tagged data in the target domain learning under supervision. In these situations, transfer learning improves learning outcomes considerably. An efficient first step in many Document Image Processing (DIP) activities, including document retrieval, information extraction, and text recognition, is the automatic classification of document pictures. Using optical character recognition (OCR) techniques to extract the words from the document image makes document classifiers based on text contents prone to OCR errors. On the other hand, some OCR systems use a structural analysis technique, first classifying the document picture and then selecting the suitable OCR module. We make use of this data to choose the VGG19 model as our basic classifier model for this challenge. Transfer learning is the process through which a machine learning model transfers its knowledge from one related area to another. All pictures from the RVL-CDIP dataset were used for training. starting weights from the VGG19 model trained on the dataset was transferred. Although it would seem like classifying documents and classifying objects are two different fields, architectures trained on the dataset have shown to be effective generalized feature extractors. Compared to the simple approaches, stacked generalization performs substantially better. In order to learn the final set of predictions, stacked generalization trains a meta-classifier on the output of the base classifiers on a hold-out dataset. The region-based DCNN models in the current work were taken into consideration as the basic models for layered generalization.

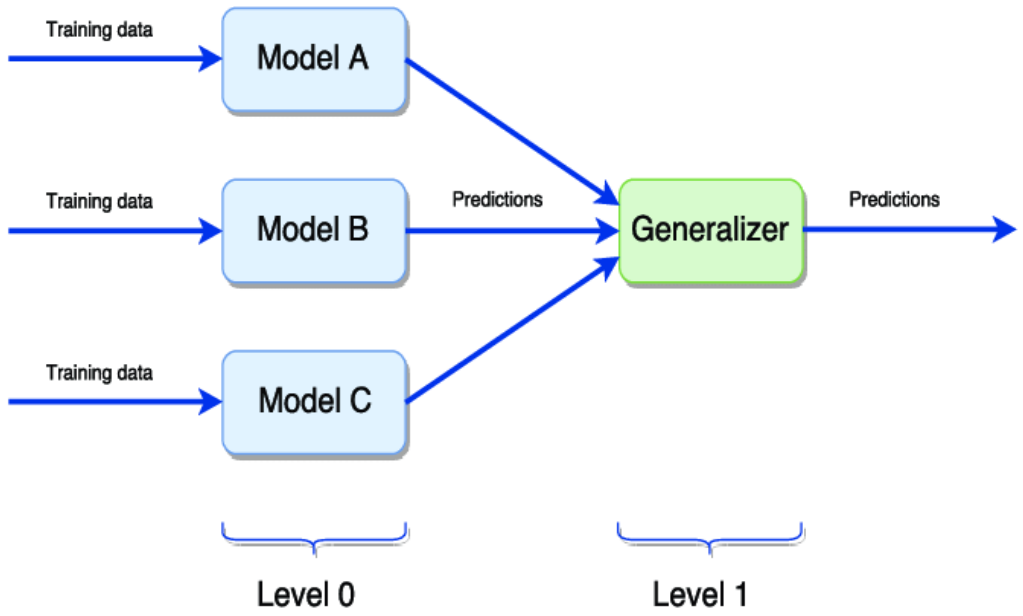
CONTRIBUTIONS:

- Deep neural networks have produced unmatched outcomes in a variety of fields, including the classification of document pictures. Deep convolutional neural networks (DCNN) are utilized in the current study to automatically know a document's structure will help you to the classification's goal.
- Making a case for both inter-domain and intra-domain transfer learning for this issue makes use of the distinctive features of the region-based method to document classification.
- By doing a detailed investigation into meta-classification utilising layered generalisation, the predictions from region-based classifiers are integrated.



Flowchart of Proposed Model for Document Image Classification with L1 and L2 Transfer Learning

Multi-model stacked generalization



Here we stack all the predicted values of above each model by using stacked generalization

Google colab link :-
https://colab.research.google.com/drive/1CFIzkCc_DSsLhT5rLZn0ofaaX2UknUm4?usp=sharing