Transfer Learning Approach for Scanned Document Classification

Project Report

Team:

Sribalaji Thirunavukkarasu

[sribalaji.thirunavukkarasu@ucdenver.edu](mailto:sribalaji.thirunavukkarasu@ucdenver.edu)

Problem Statement and Background:

In current times, banking industry handles and validates a huge number of documents. For example, banks facilitate trade payments for imports and exports, and this involves a high volume of scanned documents. Identifying the document type and extracting relevant information from these scanned documents, is the pain point when it comes to processing and approving the payments promptly. These documents, such as Shipment information, Airway/Seaway bills, invoices, emails from vendors, identity of individual or seller information will be scanned and shared to bank through email or post. A solution to minimise the human error and time consumed is to automatically tag documents and extract the information according to their classification.

Data Source:

RVL-CDIP dataset has been identified as a source of documents data for this purpose.

URL - <https://www.cs.cmu.edu/~aharley/rvl-cdip/>

* + Total Images - 400,000 grayscale Images
  + Classes – 16 and 25,000 images per class.
  + Training Images – 320,000
  + Validation Images – 40,000
  + Test Images – 40,000
  + Available Classes – letter, form, email, handwritten, advertisement, scientific report, scientific publication, specification, file folder, news article, budget, invoice, presentation, questionnaire, resume, memo.

Tobacco3482 dataset from Kaggle web site.

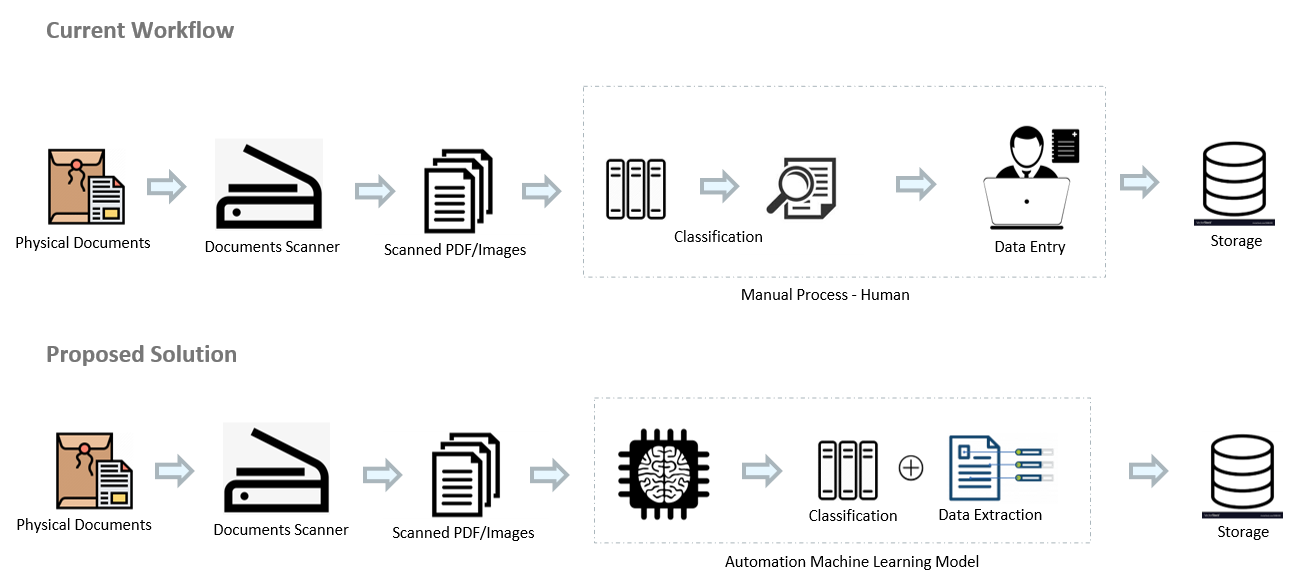
* Dataset from paper - Jayant Kumar, Peng Ye and David Doermann. "Learning Document Structure for Retrieval and Classification." International Conference on Pattern Recognition (ICPR 2012), 2012.
  + URL - <https://www.kaggle.com/patrickaudriaz/tobacco3482jpg>
  + Total Images - 400,000 grayscale Images
  + Classes – 10
  + Available Classes – ADVE, email, form, letter, memo, news, note, report, resume, scientific papers

Based on the hardware resource availability, few classes from this dataset will be handpicked for implementing this project.

Data Source from this paper - A. W. Harley, A. Ufkes, K. G. Derpanis, "Evaluation of Deep Convolutional Nets for Document Image Classification and Retrieval," in ICDAR, 2015

Objectives:

To create a product which classifies a given document to help people in validating the relevant document and to reduce the time. To find the best suited model for this document classification, success metrics such as Accuracy, precision and efficiency will be noted and monitored for various iterations. After analysing the data, other metrics will be decided. The long-term goal of this project would be to extend this idea from classifying a document to extracting the relevant information, specific to the classified document type.



Success Measures:

1. Accuracy & Loss value while training and validation.
2. Training time – To measure time taken to train the model.
3. Prediction time – Time taken to predict the given document image.

Related Works: (Literature Review):

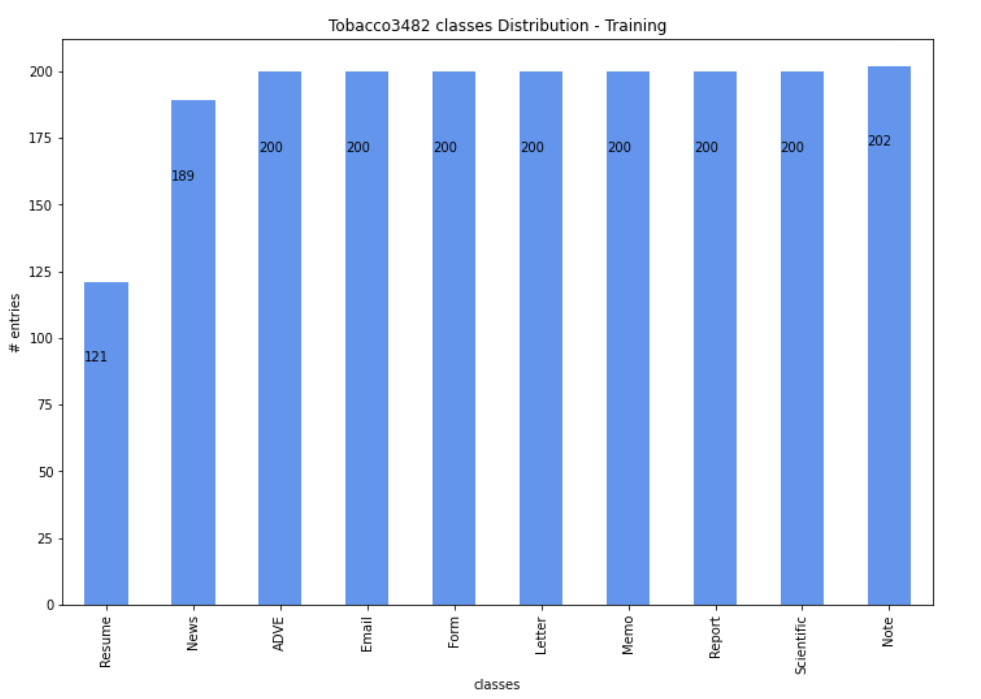
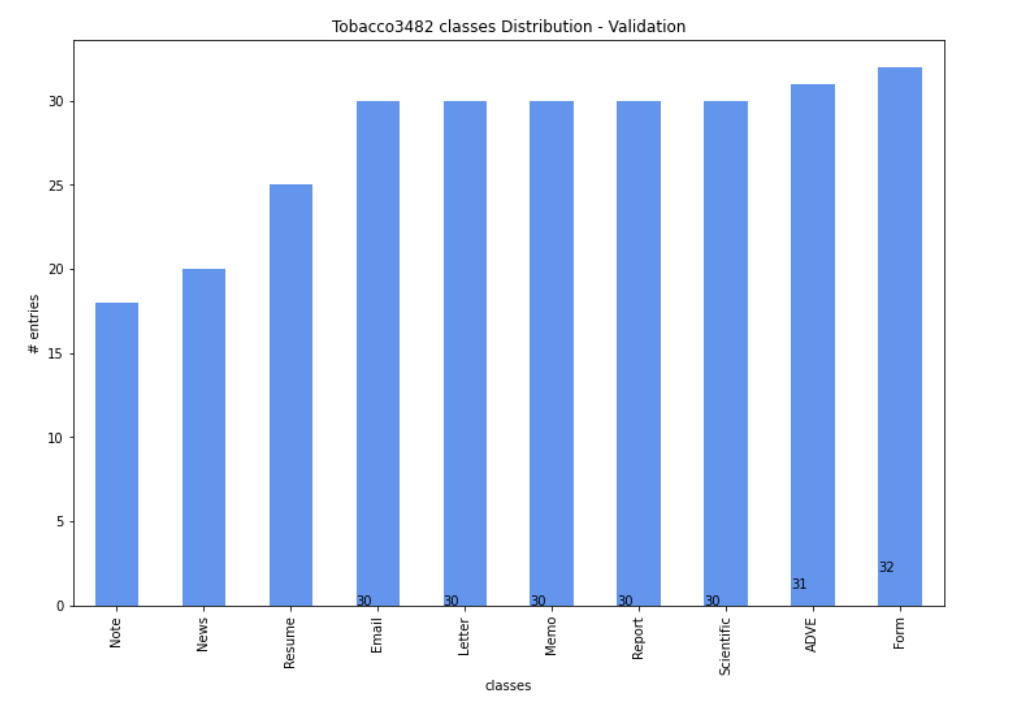
* Evaluation of Deep Convolutional Nets for Document Image Classification and Retrieval - 2015 13th International Conference on Document Analysis and Recognition (ICDAR) – Experimented with three approached of deep convolution neural network (CNNs) as follows:
  + Holistic convolutional neural networks – process the input through stack of CNN layers and then classify the output with two or three fully connected layers.
  + Region-based guidance - approach is to train one CNN for specific region, therefore force multiple CNNs to learn rich region dependent representations, from which features can be extracted and combined for classification.
  + Transfer learning – To train the CNN to another task which has more training data like ImageNet and transferring them to the new problem to facilitate learning on the problem with insufficient training data. Dataset Used – RVL-CDIP.
* Document Image Classification Using SqueezeNet Convolutional Neural Network - 5th Conference on Signal Processing and Intelligent Systems, 18-19 December 2019, Shahrood University of Technology – proposed SqueezeNet for Document Classification
  + To evaluate the performance of SqueezeNet on document image classification, model is trained on the Tobacco-3482 dataset which contains 3482 high resolution document images over 10 classes. In each class, 80 images were used for training, 20 for validation and the rest for testing model performance.
  + To improve the performance, they had three main strategies as follows.
    - Majority of filters used as 1x1 instead of 3x3.
    - Decreases the number of input channels to 3x3 filters.
    - Down sampling is performed later in the network on larger activation maps.
  + Training was performed over 150 epochs where each training epoch took approximately 6 seconds on a Tesla K80 GPU. Due to the small dataset size, training was performed five times with five different models and the accuracies achieved by these models were averaged to get the final accuracy.
* Two Stream Deep Network for Document Image Classification - 2019 International Conference on Document Analysis and Recognition (ICDAR) – Proposed two stream methodology as combination Visual stream and textual stream,
  + Visual Stream – Transfer learning Using InceptionV3, feature map of model is passed to global average pooling layer and passed to fully connected layer of 1024 units.
  + Textual Stream – Image document is passed to OCR module(Tesseract OCR), embedding metrics of top k feature for vocabulary will be passed to CNN and then CNN feature map will be passed to global average pooling layer and then concatenated and normalized.
  + Output from both visual and textual stream will be concatenated and passed to fully connected layer of 128 units with SoftMax layer for classification.
  + Dataset Used – RVL-CDIP and Tobacco3482.
* Document Class Recognition Using A Support Vector Machine Approach - 2nd International Conference on Advanced Technologies for Signal and Image Processing - ATSIP'2016 March 21-24, 2016, Monastir, Tunisia– Proposed System of documents classification that discriminates between photographs, textual and mixed documents. This system has three stage as feature extraction stage, classification, and archiving stage.
  + Features Extraction Stage – to extract low level feature like mean, standard deviation, skewness, document size and entropy extracted from the input document.
  + Classification - SVM classifier with RBF Kernel was used in this paper. Approach is to train the one CNN for specific region, therefore force multiple CNNs to learn rich region dependent representations, from which features can be extracted and then combined for classification.
  + Custom dataset has been used in this experiment which has 750 documents consist of photos, textual and mixed documents divided into 500 documents for train and 250 for test.
* Generalized Stacking of Layer wise-trained Deep Convolutional Neural Networks for Document Image Classification - 2016 23rd International Conference on Pattern Recognition (ICPR) Cancún Center, Cancún, México, December 4-8, 2016 – proposed system consists of six deep convolution neural network model and SVM combining the outputs of DCCNs.
  + Each CNN is trained to specific part of document like header, footer, left body and right body, full document image and flipped document image. A linear Support Vector Machine (SVM) was trained as a meta-classifier to combine the predictions of the individual models.
  + Dataset – Tobacco3482 was used with image resized to 150x150, Training sets of size 80 per class and validation sets of size 20 per class were selected from the dataset at random. The remaining 2482 images were used as the test set.

Initial Dataset for Training:

* For training from Tobacco3482 dataset took ~200 samples for training and ~30 samples for validation from each class.
* Data rescaled to 1./255 and image size 224x224.
* For fine tuning model, some more data from RVL-CDIP dataset will be added.
* For initial training Hardware
  + Intel I7
  + 16gb Ram
  + GTX 1060 – 6gb GPU
* Next step, to increase dataset size and train it in Google Colab.

Classes: Resume, News, Advisement, Email, Forms, Letter, Memo, Report, Scientific Documents and Notes.

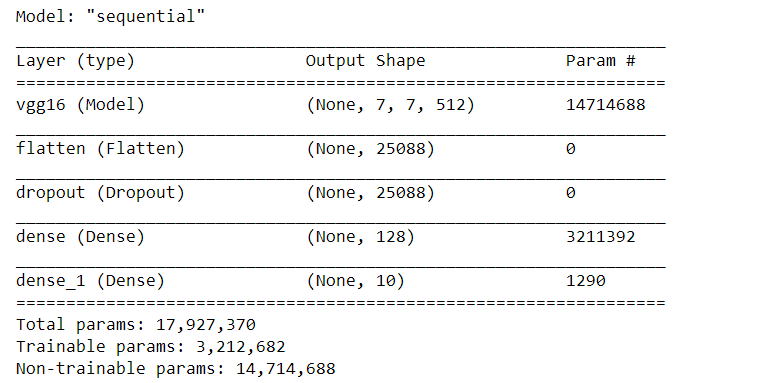
Class distribution for Training and validation

Method Explored (Initial):

The base model was chosen from different pre-trained base models like VGG16, VGG19, ResNet50, Inception\_resnet\_v2, Xception and then output of model to layer like flatten, dropout, fully connected layer with 128 units & ReLu activation function and final fully connected layer with 10 units & SoftMax activation. Below are results from the base model identification process.

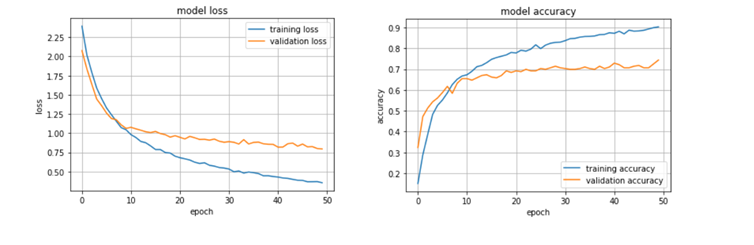
1. VGG16:



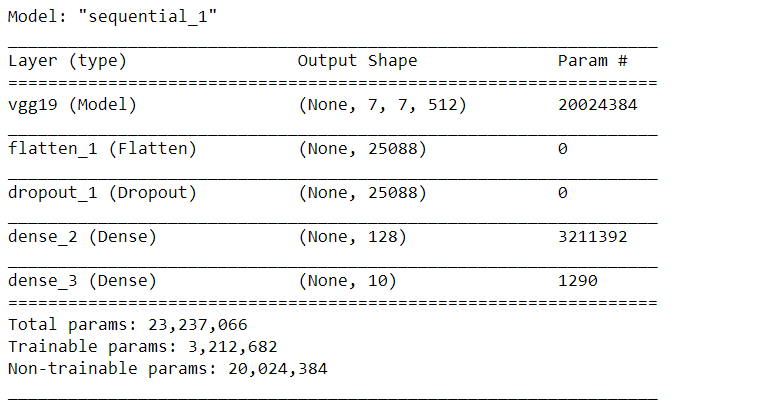
Results

* Image Size – 224 X 224
* Epochs – 50
* Training Loss – 0.2971
* Training Accuracy – 0.9193
* Validation Loss – 0.7682
* Validation Accuracy – 0.7286

Initial Model - Accuracy and Loss vs epoch



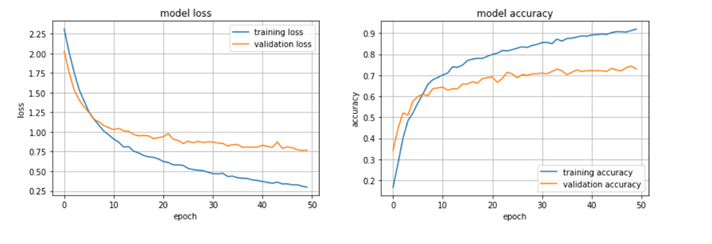
1. VGG19:



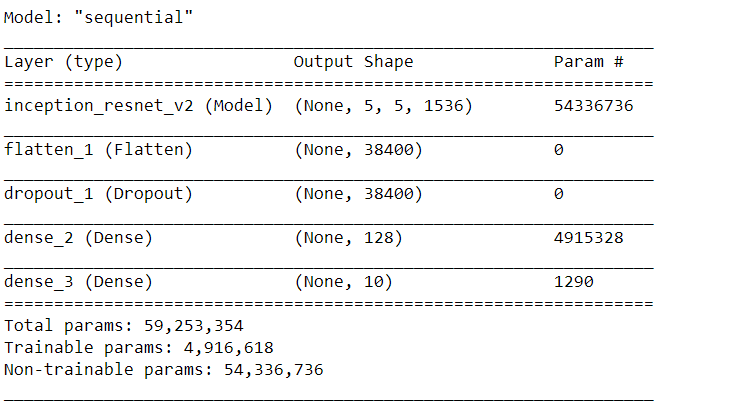
Results

* Image Size – 224 X 224
* Epochs – 50
* Training Loss – 0.2971
* Training Accuracy – 0.9193
* Validation Loss – 0.7682
* Validation Accuracy – 0.7286

Initial Model - Accuracy and Loss vs epoch



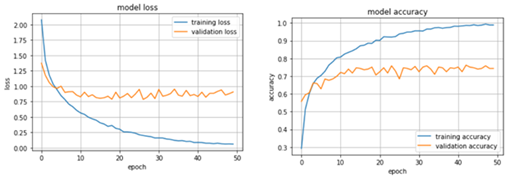
1. Inception:



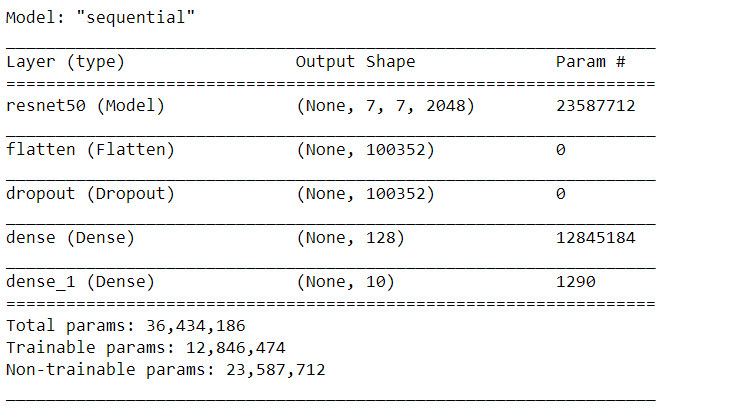
Results

* Image Size – 224 X 224
* Epochs – 50
* Training Loss – 0.3532
* Training Accuracy – 0.9026
* Validation Loss – 0.7935
* Validation Accuracy – 0.7435

Initial Model - Accuracy and Loss vs epoch



1. ResNet50:



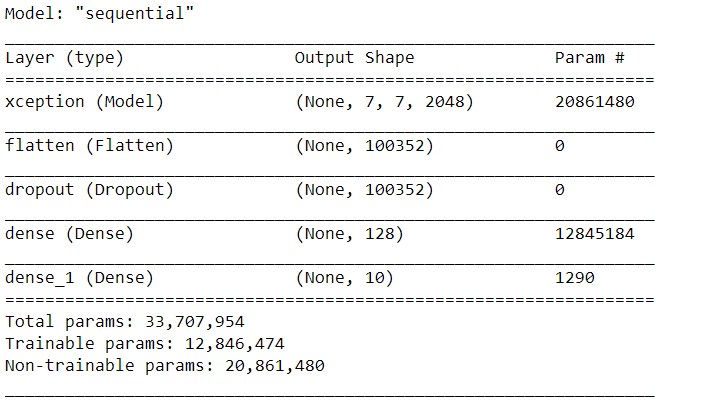
Results

* Image Size – 224 X 224
* Epochs – 50
* Training Loss – 0.8037
* Training Accuracy – 0.7271
* Validation Loss – 1.0127
* Validation Accuracy – 0.6580

Initial Model - Accuracy and Loss vs epoch



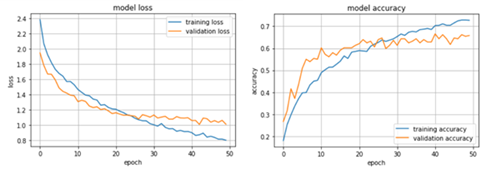
1. Xception:



Results

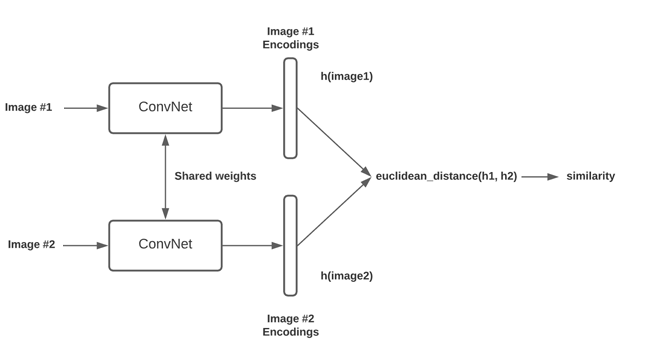
* Image Size – 224 X 224
* Epochs – 50
* Training Loss – 0.0427
* Training Accuracy – 0.9969
* Validation Loss – 0.9180
* Validation Accuracy – 0.7063

Initial Model - Accuracy and Loss vs epoch

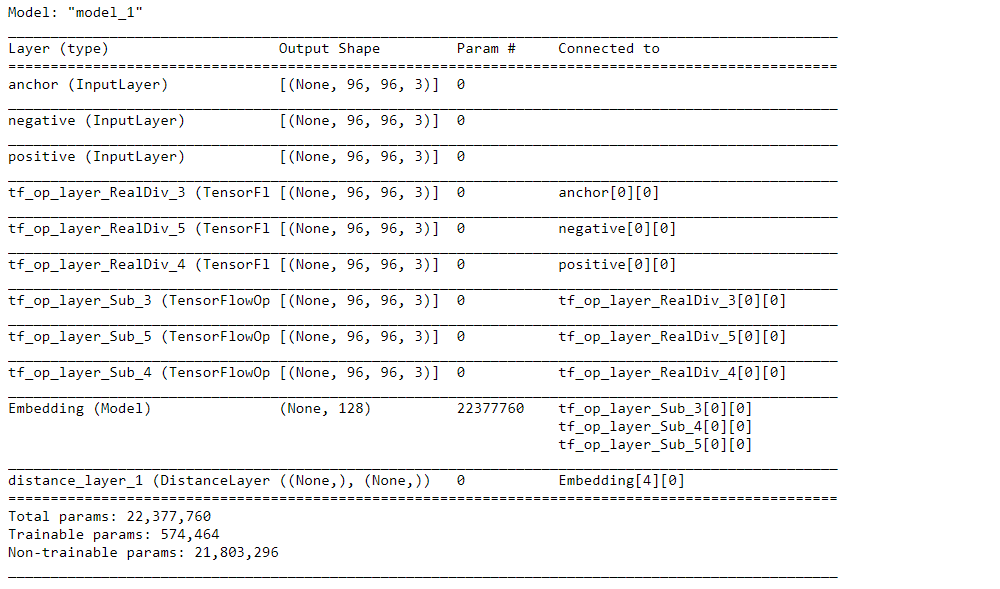


Solution: Siamese Network – Document Image Classification:

Siamese Neural Network – contains two twin network which share the same parameters and weights and their outputs are jointly trained on top with a function to learn the relationship between pairs of input data samples. Once model is fined tuned, by generalising the model we can predict not only the new data but with entirely with new classes from unknown distribution.



Siamese Network – Triplet InceptionV3:



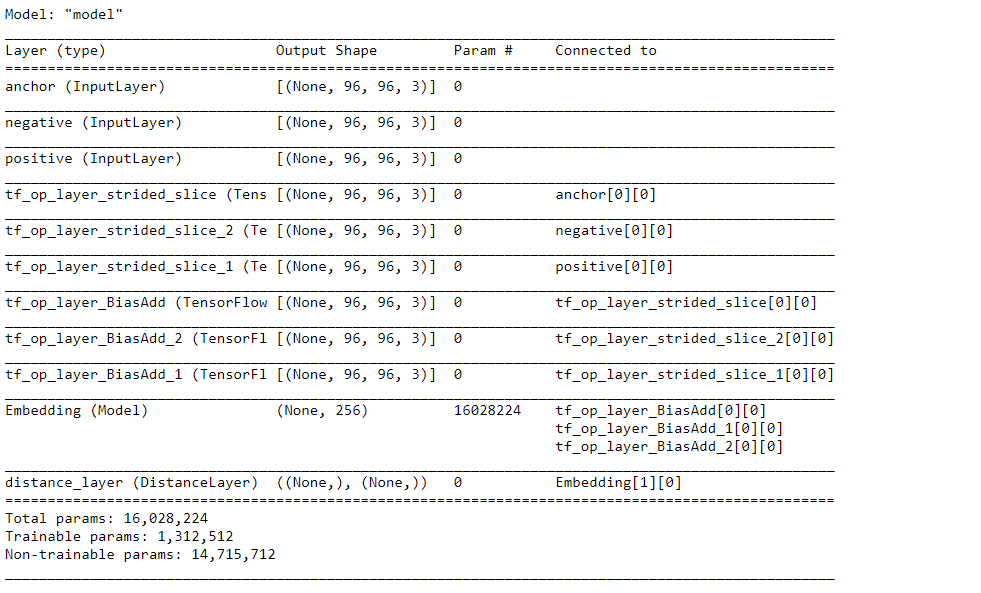
Loss:

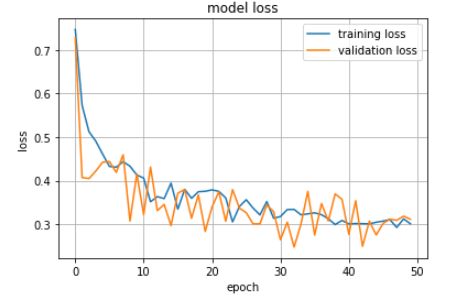


Results:

* Image Size – 224 X 224
* Dataset – 640 images
* Epochs – 50
* Training Loss – 0.1147
* Validation Loss – 0.1611

Siamese Network – Triplet VGG16:

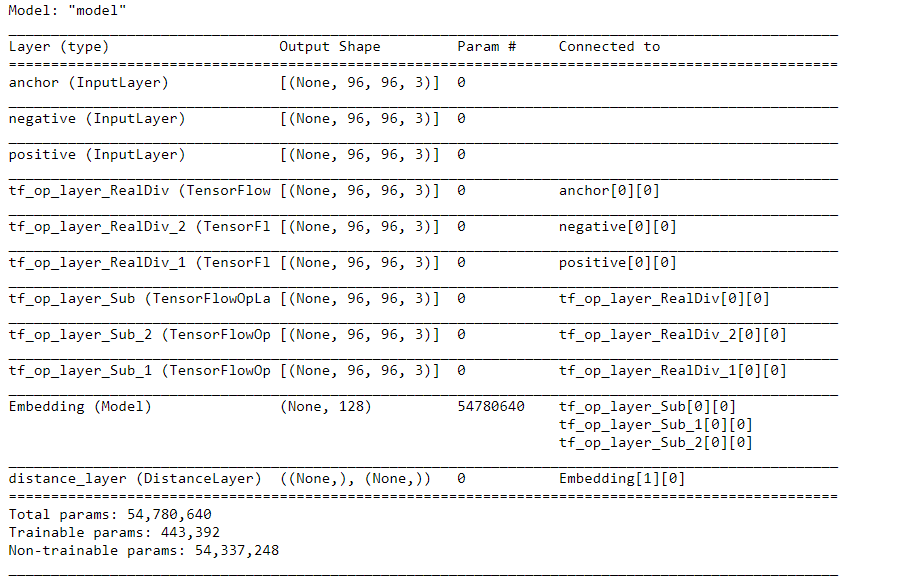




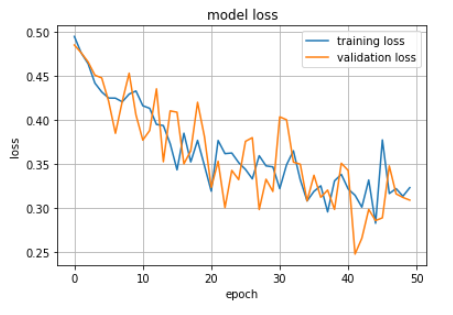
Results:

* Image Size – 224 X 224
* Dataset – 640 images
* Epochs – 50
* Training Loss – 0.3012
* Validation Loss – 0.3115

Siamese Network – Triplet Inception Resnet V2



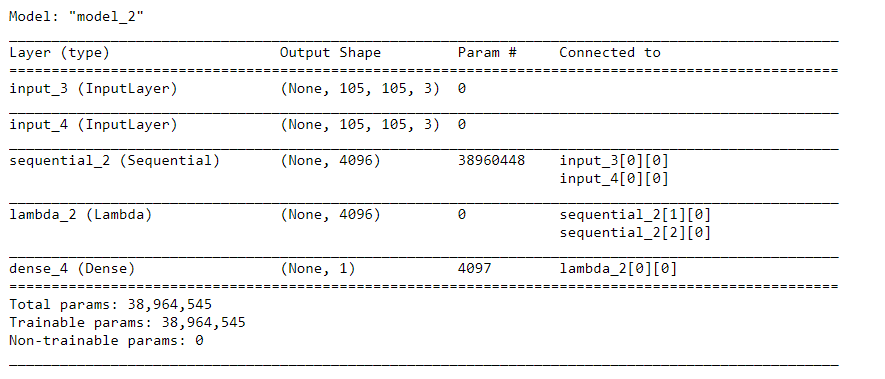
Loss



Results:

* Image Size – 224 X 224
* Dataset – 640 images
* Epochs – 50
* Training Loss – 0.3232
* Validation Loss – 0.3091

Siamese Network – One Shot Learning:



Reference:

https://www.cs.cmu.edu/~rsalakhu/papers/oneshot1.pdf

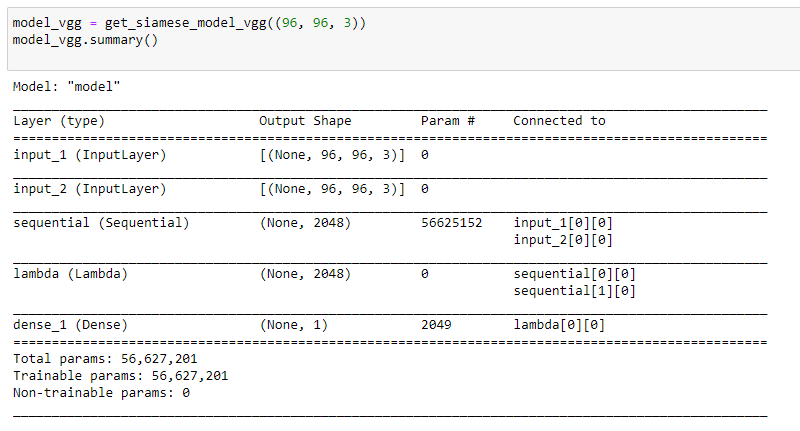
https://sorenbouma.github.io/blog/oneshot/

<https://github.com/tensorfreitas/Siamese-Networks-for-One-Shot-Learning>

Results:

* Image Size – 105 X 105
* Dataset – Tobacco3482
* Epochs – 20000
* Training Loss – 0.0830

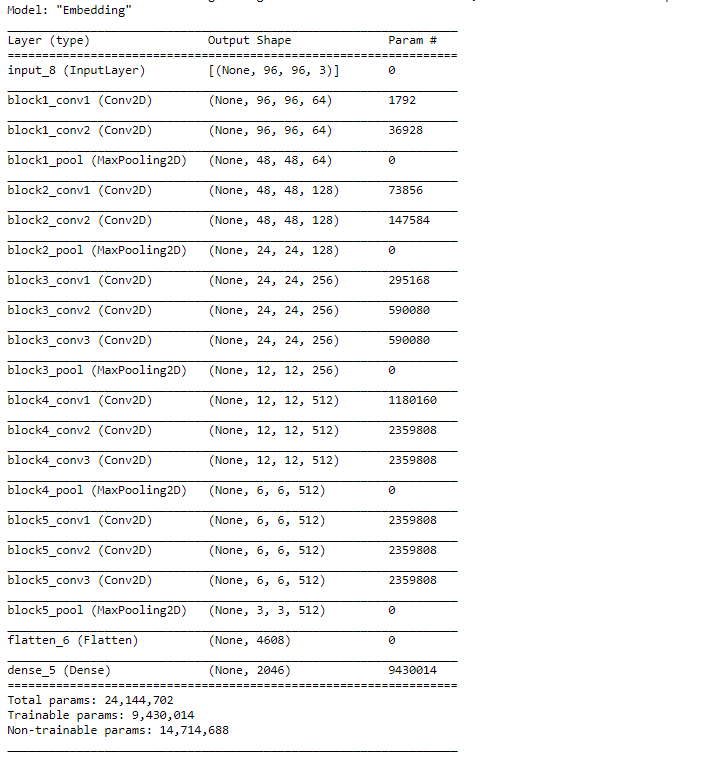
Siamese Network – One Shot Learning (VGG16):



Result:

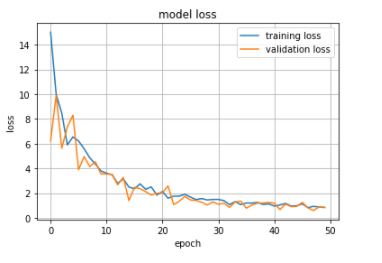
* Image Size – 96x96
* Base model – VGG16
* Dataset – Tobacco3482
* Epochs – 20000
* Training Loss – 0.2012

Final Model:



Result:

* + - * + Image Size – 96 X 96
        + Dataset – 640 images
        + Epochs – 100
        + Training Loss – 0.3012
        + Validation Loss – 0.3115



From the result comparison, it can be concluded that the Siamese Neural Network - Triplet VGG16 is a better performing and suitable model.

**Result Comparison:**

**VGG 16 – Transfer Learning** – performs well with 75% accuracy but requires large data and infrastructure to train the best model. It will not work with new class or unknown distribution without retraining the model. It takes longer training time.

**Siamese Network – One Shot Learning** – even though the model has very less training loss of 0.0830 and gave 95% accuracy of training validation accuracy, it didn’t perform well test data. Model looks like overfitting model. I have noticed the similarity score is very less when I tried trained model on test data and unseen data. Higher execution time while predicting the classes.

**Siamese Network – Triplet Loss** – Model performs well in both train and validation dataset with training loss of 0.3012 and validation loss of 0.3115. When faced with Unseen dataset classes like passport, visa and driving license, the model performed well. Executes faster when compared to one-shot learning model.

Description of Tools:

TensorFlow 2.x and Python 3.7 will be used for training the model and best fit iteration will be exported and stored for demonstration through web application based on Flask, HTML & JavaScript.

Lessons Learnt:

Though, Siamese network will perform better for structured documents, in case of a wide range of inputs, traditional CNN will perform better than Siamese network, with only drawback of increased training time.

Areas of Improvement:

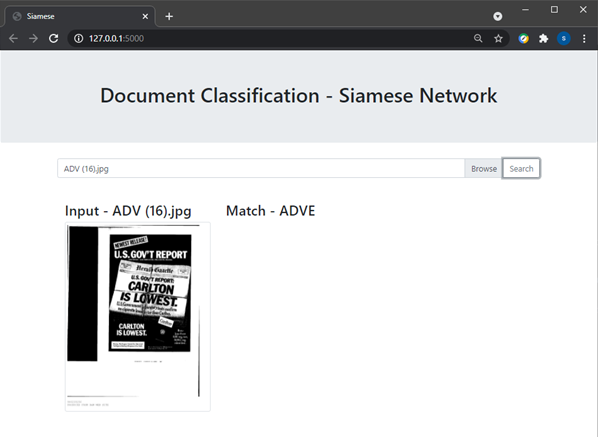
1. There could have been more accurate if more hyperparameter tuning was performed or experimented.
2. Better sampling of dataset - selecting apt Anchor, Positive and Negatives images for the Siamese network is key to increase accuracy.

Reference:

* A. W. Harley, A. Ufkes, K. G. Derpanis, "Evaluation of Deep Convolutional Nets for Document Image Classification and Retrieval," in ICDAR, 2015
* K. Gregory, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition." ICML Deep Learning Workshop. Vol. 2. 2015.
* Antonacopoulos and R. T. Ritchings, "Segmentation and classification of document images," IEE Colloquium on Document Image Processing and Multimedia Environments, 1995, pp. 16/1-16/7, doi: 10.1049/ic:19951197.
* A. M. Awal, N. Ghanmi, R. Sicre and T. Furon, "Complex Document Classification and Localization Application on Identity Document Images," 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), 2017, pp. 426-431, doi: 10.1109/ICDAR.2017.77.
* B. Su, S. Lu and C. L. Tan, "Combination of Document Image Binarization Techniques," 2011 International Conference on Document Analysis and Recognition, 2011, pp. 22-26, doi: 10.1109/ICDAR.2011.14.

Demo Application Screenshot:

* Classifying an Advertisement image which was predicted correctly. Advertisement is part of Tobacco3482 dataset.



* Predicting a Colorado state Driver’s Licence, which was not part of Tobacco3482 dataset. These few images were converted to feature map embedding and stored as database. For given image, feature map embedding will be predicted and using the Euclidean distance, class will be predicted.

