Progress Report – Document image classification for digital library collections

08/20/2019

Mike Pack

Objective

In this report, we aim to report on the following two experimental results: (1) classification performances of *VGG-16*, pre-trained on ImageNet, trained and tested with *RVL_CDIP* dataset and (2) classification performances of *VGG-16*, pre-trained on *RVL_CDIP* in (1), trained and tested with *suffrage_1002* dataset collected from the By the People corpus. The remainder of this report is organized as follows: in Experiment 1, a configuration of a dataset (i.e., *RVL_CDIP*) and training process is described, followed by training and testing results. In Experiment 2, similar to Experiment 1, a configuration of a dataset (i.e., *suffrage_1002*) and training process is described, followed by training and testing results.

Experiment 1:
Training and Testing *VGG-16* pre-trained on *ImageNet* with *RVL_CDIP*

The objective of this experiment is to reproduce the result reported in the work of Harley et al. (2015) which is training a model, *VGG-16*, with a large-scale document image dataset (i.e., *RVL_CDIP*) using transfer learning [1]. The advantage of this experiment is that once we have a model trained on this large-scale document image dataset, we can reuse the rich features that this model has learned for many document analysis tasks, say, one of our main tasks, a document type classification.

Dataset: *RVL_CDIP*

The *RVL_CDIP* dataset, which is publicly available, consists of 400,000 document images that are divided into 16 evenly distributed classes. The dataset is provided in three different sets: training, validation, and test set. The training set contains 320,000 images of 16 different evenly distributed classes (i.e., about 20,000 images per class). Both validation and test sets together contain 40,000 images of 16 different evenly distributed classes (i.e., 2,500 images per class).

Network Architecture: *VGG-16*

We use the original *VGG-16* architecture, but the output tensor is adjusted to have a shape of 16, which is the number of classes found in the *RVL_CDIP* dataset.

Training

As a preprocessing step, in order to make the shape of our input to match with that of *VGG-16*, we convert grayscale images to three-channel images by simply copying the pixel values of the single-channel to three channels. Also, each image is resized to 224 by 224, and normalized to range from 0 to 1 by dividing each pixel's intensity value by 255. In accordance with the size of the training set and under a limited memory constraint, we use a batch size of 126. As an optimizer, we use adaptive momentum estimation, or so-called Adam, which is the state-of-the-art optimizer and also known as the rule-of-thumb [2]. The initial learning rate is set to a small value, 10^{-5}. This is because the model has been already pre-trained on *ImageNet*, so we want to prevent overshooting local minima of the loss function. The training is scheduled to run 80 epochs total, but we use early-stopping to terminate the training process if the validation loss is not improved than that of the previous iteration.

Results

Interestingly, the entire training process took only three epochs to converge with promising classification results. This indicates that features obtained from natural scene images (i.e., ImageNet) are general enough to be applied to documents. The resultant classification performance metrics—precision, recall, and f1-score—are shown in Table 1. On average, each metric shows around 87%, which aligns well with the result reported by Harley et al. (2015). In Figure 1, more detailed classification performance on the test set is visualized as a heatmap. *A series of high support values in diagonal elements indicates that the trained model is capable of producing many correct predictions.*

Table 1. Precision, recall, and f1-score of *VGG-16* trained on *RVL_CDIP* dataset. The alphabetic labels are corresponding to the following labels: *letter*, *form*, *email*, ***handwritten***, *advertisement*, *scientific report*, *scientific publication*, *specification*, *file folder*, *news article*, *budget*, *invoice*, *presentation*, *questionnaire*, *resume*, and *memo*. Our class of interest, ***handwritten***, is bolded.

(unit: %)	A	B	C	**D**	E	F	G	H	I	J	K	L	M	N	O	P	Avg
Precision	86	74	98	**89**	89	73	90	88	89	92	87	91	78	91	92	88	87
Recall	94	79	97	**96**	91	73	93	91	97	86	83	86	79	73	94	91	87
F1	86	77	97	**92**	90	73	91	90	93	89	85	88	79	81	93	90	87

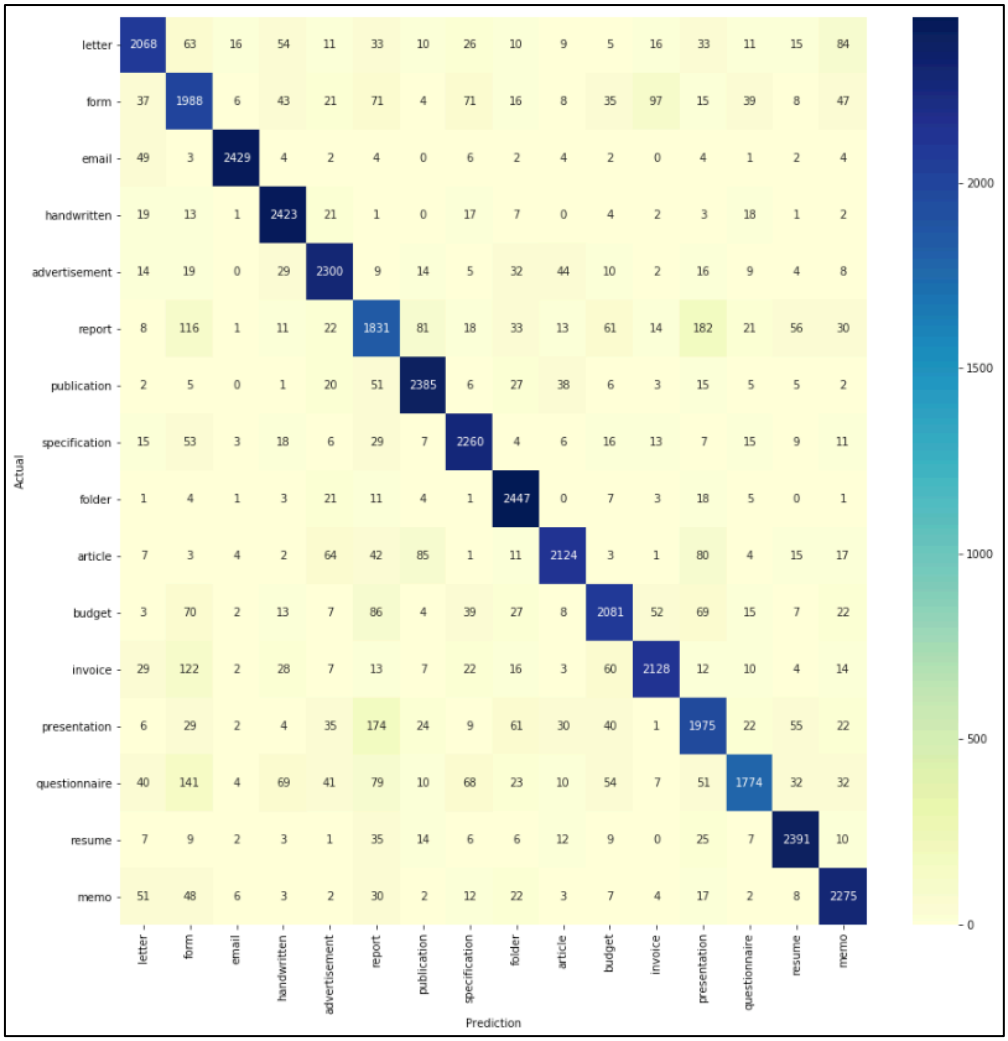

Figure 1. Heatmap of confusion matrices for classification performance of *VGG-16* trained on *RVL_CDIP*. Note that the diagonal elements represent the numbers of occurrences for which the predicted label is equal to the true label, while off-diagonal elements are those that are misclassified by the classifier. The higher the diagonal values of the confusion matrix the better, indicating many correct predictions.

Experiment 2:
Training and Testing *VGG-16* pre-trained on *RVL_CDIP* with *suffrage_1002*

The objective of this experiment is to generate our own model for this specific task—three-class document type classification; *handwritten*, *typed*, *mixed*—by retraining the model obtained from the previous Experiment 1 with our own *suffrage_1002* dataset.

Dataset: *suffrage_1002*

Thanks to Dr. Lorang and Ashlyn Stewart, we have collected a total of 1,002 images from a suffrage collection in By the People corpus[1]. This dataset is a fully balanced set (334 *handwritten*; 334 *typed*; 334 *mixed*) that has been compiled manually. The entire dataset is split into three sets—training, validation, and test—with the ratio of 8:1:1. Note here that in order to keep the class balanced during this split, it is inevitable to drop some datapoints (i.e., three datapoints). Our final dataset configuration is elaborated in Table 2.

Table 2. Configuration of *suffrage_1002* dataset.

	handwritten	**typed**	**mixed**	Total
train	267	267	267	801
validation	33	33	33	99
test	33	33	33	99
Total	333	333	333	999

Network Architecture: *VGG-16*

We use the same *VGG-16* architecture as in Experiment 1, but the output tensor is adjusted to have a shape of 3, which is the number of classes specified in *suffrage_1002* dataset.

Training

All the training configuration is the same as the previous Experiment 1, except for an initial learning rate and batch size. We use an initial learning rate of 10^{-6}, which is smaller than the one used in Experiment 1, since the model is pre-trained on *RVL_CDIP* on top of *ImageNet*. We also use a smaller batch size of 32 in accordance with the size of *suffrage_1002* dataset.

Results

Generally, one can diagnose whether a model is overfitted or underfitted to its training dataset based on a model's training and validation loss. For example, if a validation loss increases while training loss decreases, the learned model is speculated to have overfitted. Taking this into account, as shown in Figure 2, based on the overall decreasing trends of both training and validation loss, during the training, there is no symptom of overfitting or underfitting.
Overall, our model's classification performance on the testing set shows about 90% of precision, recall, and f1-score, as shown in Table 3. Compared to the other two classes, a *mixed* type shows relatively poor recall performance (i.e., 79%). We believe that this is due to challenging characteristics of *mixed* type document images; for example, too small amounts of handwriting

[1] https://crowd.loc.gov/topics/suffrage-women-fight-for-the-vote/

in a *typed* document, or vice versa, as shown in Figure 3. In Figure 4, more detailed classification performance on *suffrage_1002* test set is visualized as a heatmap.

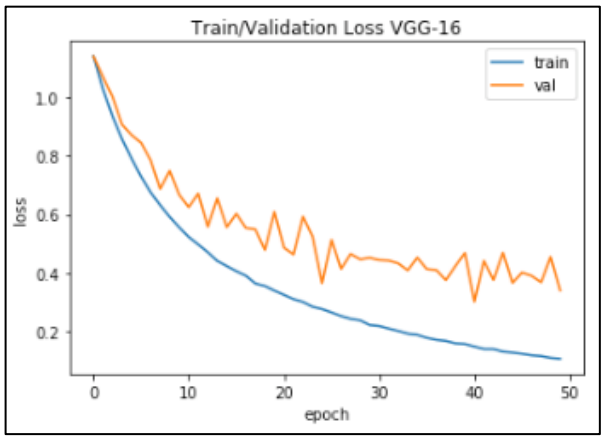

Figure 2. Training and validation loss of *VGG-16* with *suffrage_1002* training and validation set. In spite of some fluctuations, the overall trend of validation loss goes down.

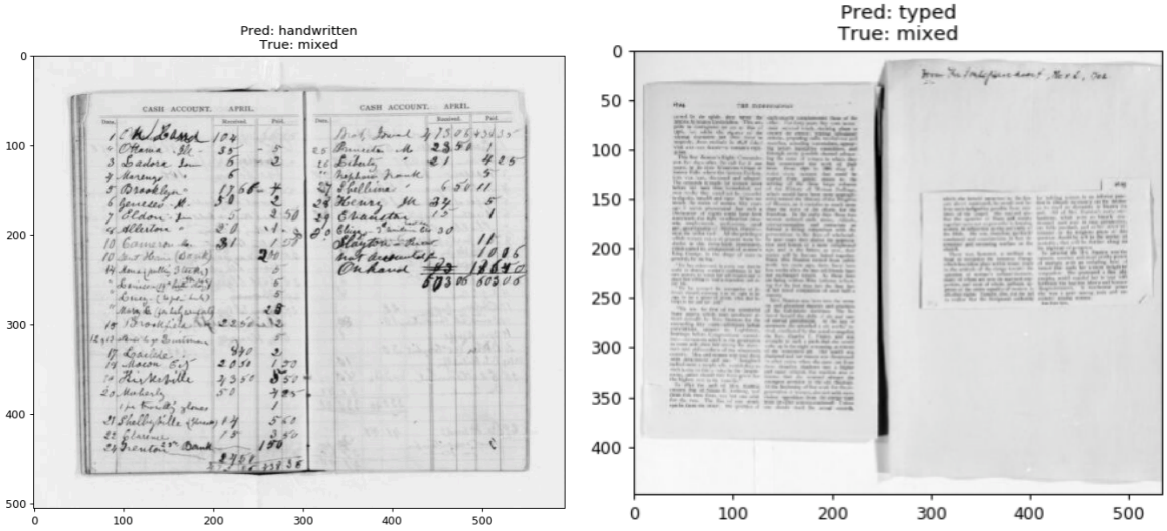

Figure 3. Failure prediction cases. On the left example, a typed region is relatively smaller than that of handwriting. On the right example, a handwriting region is relatively smaller than that of typing.

Table 3. Precision, recall, and f1-score of *VGG-16* on *suffrage_1002* testing set.

(unit: %)	**handwritten**	**typed**	**mixed**	**Avg**
Precision	89	91	90	90
Recall	97	94	79	90
F1	93	93	84	90

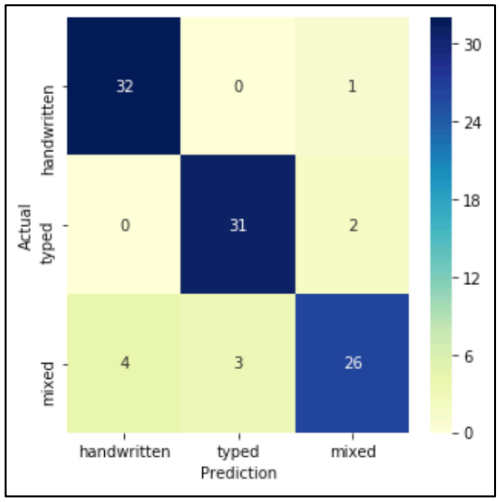

Figure 4. Heatmap of confusion matrices for classification performance of *VGG-16* trained on *suffrage_1002*. Note that diagonal elements contain most of datapoints, which indicates that most of our model's predictions are correct over all three classes.

Reference

[1] Harley, A.W., Ufkes, A. and Derpanis, K.G., 2015, August. Evaluation of deep convolutional nets for document image classification and retrieval. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)* (pp. 991-995). IEEE.
[2] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.