Evaluating Out-of-Distribution Performance in RVL-CDIP Document Classification

Gordon Lim[1], Yutong Ai[1], Stefan Larson[2]
gbtc@umich.edu, [1]University of Michigan, [2]SkySync, Ann Arbor, MI

1. Abstract

The Ryerson Vision Lab Complex Document Information Processing (RVL-CDIP) corpus [1] is the de facto standard benchmark for document classification, yet to our knowledge all studies that use this corpus do not include evaluation on out-of-distribution (OOD) documents. We report on our evaluation of document classifiers trained on RVL-CDIP and tested on a new set of over 3000 OOD documents. Based on our experiments, we find that standard image-based classifiers are not adequate at discriminating between in-distribution and out-of-distribution inputs using uncalibrated confidence scores.

2. Background

The RVL-CDIP corpus consists of grayscale images of scanned documents from the IIT-CDIP collection, a large repository of publicly-available documents that were released as part of litigation against several tobacco-related companies. As such, all documents in the RVL-CDIP corpus are tobacco-related. The corpus consists of 16 categories (see Fig. 1). There are 320, 000 train images (20,000 from each category). There are 40,000 validation and 40,000 test images. All documents from this dataset are from the year 2006 or earlier, with 2006 being the year that the IIT-CDIP collection was released.

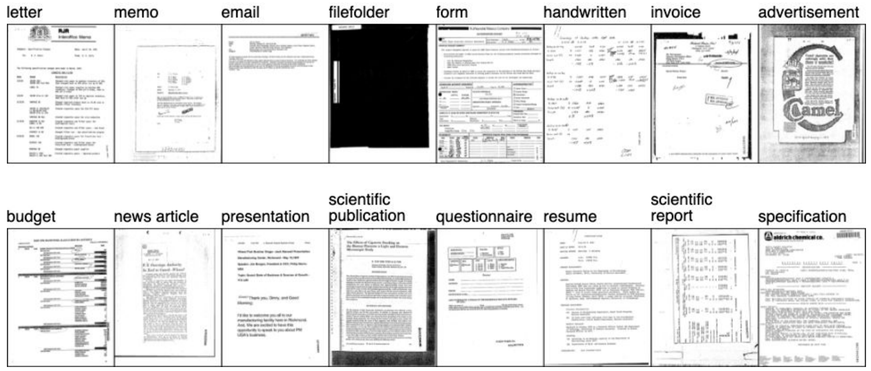

Fig. 1 Example of each of the 16 categories of RVL-CDIP [1]

We consider two definitions to out-of-distribution [2]:

Concept shift: The data is of a category that is not part of the training dataset
Covariate shift: The data is part of the target label set but it is different in style, age etc.

In the context of RVL-CDIP, examples of concept shift OOD documents include music sheets (see Fig. 2) while that of covariate shift OOD documents include letters from more recent years (see Fig. 3). Ideally, models will be able to detect concept shift data (produce low confidence estimates) and also classify covariate shift data correctly.

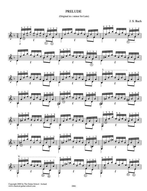

Fig. 2 Example of concept shift OOD document, music sheet

Fig. 3 Example of covariate shift OOD document, letter from 2010

3. Objectives

Prior work has shown that even if classifiers perform well on in-distribution inputs, they may struggle on the task of out-of-distribution prediction (e.g., Larson et al. [3] for short-text classifiers). Moreover, few studies have investigated out-of-distribution performance for document classifiers. We hypothesize that state-of-the-art models trained on RVL-CDIP would perform badly on out-of-distribution prediction.

4. Dataset

Our new out-of-distribution dataset consists of both concept shift and covariate shift OOD documents. Our OOD documents were collected from the Internet. To collect concept shift OOD documents, we used keywords that would not overlap with RVL-CDIP's 16 categories (e.g music sheet, traffic advisory). To collect covariate shift OOD documents, we simply used keywords from RVL-CDIP (e.g handwritten). Every document was then manually reviewed to ensure that they were in fact of the keyword. Documents that are "born-digital" (i.e., they are not scanned versions of physical document) have been processed by the Augraphy tool [4] to add scanner-like noise to our out-of-distribution set to mimic documents from RVL-CDIP (see Fig. 4).

Fig. 4 Example out-of-distribution document images unprocessed (above) and with Augraphy's scanner-like noise (below)

5. Experiments

We trained several image-based classifiers on the full RVL-CDIP training set. These models are VGG-16, ResNet-50, GoogLeNet, AlexNet, and LayoutLMv2. The accuracy scores that we achieved on the RVL-CDIP test set are shown in the second column of Table 1.

Given confidence scores for the in-distribution and concept shift OOD data, we use Area Under the Curve (AUC) to measure the separability between the two distributions. An AUC of 1.0 would mean perfect separation, and that classifiers are able to completely distinguish between in- and out-of-distribution inputs. An AUC of 0.5 (indicating the two distributions are roughly overlap) would mean that classifiers are unable to distinguish between the two types of inputs. For our covariate shift OOD, we measure the accuracy of our models

6. Results

Table 1 charts in-distribution accuracy of each image classifier on the RVL-CDIP test set. Most models come close to reported results in prior work. On our concept shift OOD data, the AUC scores are relatively high. When we use Augraphy to add scanner-like noise to our OOD data, the AUC score drops for all models. The confidence density plots in Fig. 5 provide a nice graphic representation of the overlap between distributions.

Model	ID Acc. (reported)	ID Acc. (achieved)	AUC	AUC (Augraphy)
VGG-16	0.910	0.908	0.888	0.872
ResNet-50	0.911	0.891	0.862	0.842
GoogLeNet	0.884	0.871	0.857	0.842
AlexNet	0.900	0.885	0.886	0.880
LayoutLMv2	0.953	0.887	0.838	0.829

Table 1 In-distribution accuracy compared with AUC for each model

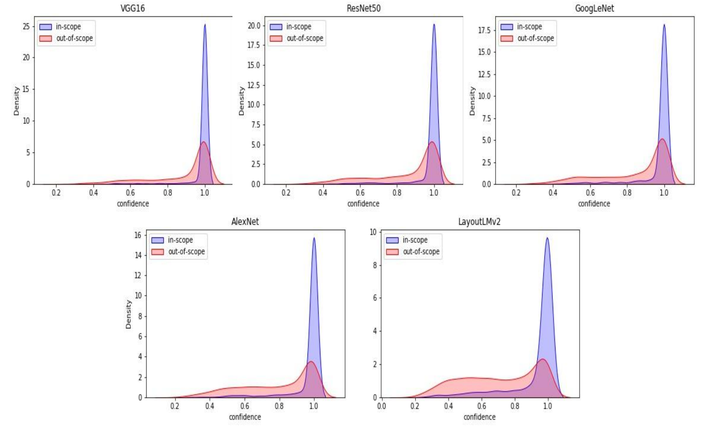

Fig. 5 Confidence density plots of RVL-CDIP test set vs concept shift OOD dataset each titled by the respective model

For our covariate OOD data, there were sharp drops in accuracy from that of the RVL-CDIP dataset (see Table 2). Augraphy preprocessing gave slight improvements to all models except for GoogLeNet.

Model	ID Acc. (achieved)	OOD Acc.	OOD Acc. (Augraphy)
VGG-16	0.908	0.683	0.697
ResNet-50	0.891	0.575	0.616
GoogLeNet	0.871	0.633	0.612
AlexNet	0.885	0.607	0.612
LayoutLMv2	0.887	0.533	0.557

Table 2 Accuracy of each model on covariate shift OOD

7. Conclusion

The AUC scores indicate that the models are able to distinguish between in- and concept shift OOD documents reasonably well. Adding scanner-like noise to the out-of-distribution test set pushes the AUC scores down for all of the supervised models, which seems to indicate that adding the noise to the OOD documents makes them more similar to the in-distribution RVL-CDIP documents.

While the AUC scores are reasonable, we inspected the confidence scores returned by the models and found that many of the content shift OOD test documents are also near 1.0, but typically slightly lower (e.g., 0.985). This tells us that the models still predict the concept shift OOD with high in-distribution confidence.

The poor accuracy on covariate shift OOD data suggests that the RVL-CDIP dataset may not provide enough diversity to enable generalizability to real world data.

In conclusion, the out-of-distribution detection problem is an important yet overlooked problem in the document classification field. Future work can look into alleviating these problems we have identified while usingour new OOD dataset as a benchmark

1. A. W. Harley, A. Ufkes, K. G. Derpanis, "Evaluation of Deep Convolutional Nets for Document Image Classification and Retrieval," in ICDAR, 2015
2. Tian, Junjiao & Hsu, Yen-Change & Shen, Yilin & Jin, Hongxia & Kira, Zsolt. (2021). Exploring Covariate and Concept Shift for Detection and Calibration of Out-of-Distribution Data
3. Larson, S., Mahendran, A., Peper, J., Clarke, C., Lee, A., Hill, P., Kummerfeld, J. K., Leach, K., Laurenzano, M. A., Tang, L., Mars, J.: An Evaluation Dataset for Intent Classification and Out-of-Scope Prediction. In: Empirical Methods in Natural Language Processing ({EMNLP}) (2019) .
4. The Augraphy Project. Augraphy: an augmentation pipeline for rendering synthetic paper printing, faxing, scanning and copy machine processes.