Keywords: Document Image Classification · Data Augmentation · Computer Vision · Document Analysis Systems · Convolutional Neural Networks.

1 Introduction

Document image classification is a challenging fine-grain classification problem. ***[Ronny: might be helpful to explain what fine grain classifcaition means here]*** The variation ***[Ronny: of the images in the same class]*** is often high, distribution shift and the addition of new classes often occurs after deployment, and there is little to no color variation in document images as many are grayscale, which reduces the richness of available features. Nevertheless, high accuracy in these classification models is of paramount importance for downstream task dependencies on classification output labels for information retrieval, language modeling and translation, and many other downstream tasks. However, deployment of these models, such as with OCR applications on handheld or edge devices, require parameter efficiency to be feasible for computing on smaller systems and hardware.

We assess on RVL-CDIP [7] pre-trained weight initializations on select EfficientNet capacities and training configurations as well as on InceptionResNetv2 and vanilla ResNets of varying capacity. Additionally, we examine the effectiveness of various augmentation strategies, including flipping, cropping, cutout [6] and mixup [13] augmentation.

A large and impactful amount of work has already been done in using neural methods for document image classification, and it is impossible to reference all recent work in this area.

Csurka et al [3] assessed various modalities for treatment of document images, including visual, structural and textual features across a variety of document image datasets. Tensmeyer et al[10] perform a robust assessment of CNN architectures and associated training strategies on RVL-CDIP and ANDOC, showing that input dimension size and shearing are critical factors regardless of CNN architecture.

Afzal et al [1] examined the impact of transfer learning for document image classification from ImageNet weights vs models trained from scratch on RVL-CDIP, showing significant improvements in transfer learning capability on Tobacco-3482 and setting a new SOTA on RVL-CDIP at the time of publication using VGG-16. Notably, the highest model performance both on RVL-CDIP and on transfer learning classification on Tobacco-3482 came from VGG-16, showing superior performance to ResNet5, GoogleNet, and Inception Models, despite its significantly simpler architecture.

We also examined the evolution of multi-model ensembles put forward by these authors and others working with CNNs for document image classification. We find the work put forward by A Das et al in [5] to be the SOTA on RVL-CDIP, with 92.47 percent test accuracy. Their approach built a system of five in-

terconnected VGG-16 models, trained on different cropped sections of document images, in combination with a "holistic" VGG16 initialized with ImageNet1K weights. An elegant L1 and L2 transfer learning approach is introduced, and the terminal network output probabilities are provided by an MLP metalearner built on top of the six VGG-16 CNN column softmax probabilities collectively.

2 Methodology

As our approach simply modifies existing CNN architecture for fine tuning, the methodology is relatively simple. For all experiments we use convolutional architectures with MLP classifier atop candidate CNN backbone architectures previously discussed. We connect with a 2D global max pooling layer to the terminal softmax activation, with a single 2D spatial dropout layer in between to limit overfitting.

As document images are often affected by distortions generated by humans such as lines, smudges, and spurious strokes and other artifacts in production systems [8]. Additionally document images suffer deterioration from scan noise, warping, reduction in DPI during format transfer, and many other factors that severely distort the image, from machines. As such, our approach employs a Python batch generator to create desired augmentations during training. Noise is introduced using cutout augmentation [6]. Cutout augmentation works by masking random regions of the document throughout the training cycle. Our approach progressively increases the degree of "cutout" through the training cycle.

3 Experiments

To test our models, we used RVL-CDIP [11]. We initialized with ImageNet-1K weights. We use SGD with momentum, with an initial learning rate of 0.1 and momentum of 0.9 based on analysis of plotting learning rate vs loss per [9]. We trained for a maximum of 500,000 total steps using early stopping and penalized learning rate schedule. For penalized schedule, We penalized learning rate based on performance on validation set after each epoch. If no improvement after N epochs, we scaled the learning rate by a factor of 90 percent. We found this to be a nice blend between stepwise decay and a less responsive penalized schedule. We used Hyperopt [2] library to do all Bayesian search.

We assessed various input resolutions. In general, if we increase image size, we see performance improvement from the higher resolution. The trade-off being much longer training times for much larger images. Most images are 1000x 750 in RVL-CDIP. We provide a small ablation study on test set performance by input image dimension in Table 1.

We use categorical cross entropy to score accuracy of the classification models. We held out the original test set from the RVL-CDIP dataset and used test set results in all of our reporting.

Table 1. Model Performance Details

Model name	Batch Size	Test Accuracy	Total Steps(K)	Image Size	Optimizer	LR	Cutout
EfficientNetB4	64	**0.9281**	500	380	SGD	0.01	Y
InceptionResNetV2	16	**0.9263**	250	512	SGD	0.1	N
EfficientNetB2	64	0.9157	500	260	SGD	0.01	Y
EfficientNetB0	64	0.9053	500	224	SGD	0.01	Y
EfficientNetB0	32	0.9036	247.5	224	SGD	0.01	Y
EfficientNetB0	32	0.8983	145	224	SGD	0.01	Y
EfficientNetB0	64	0.8951	100	224	SGD	0.01	Y
EfficientNetB0	32	0.8921	192.5	224	Adadelta	1.00	Y

Training was done on Tesla V100 GPUs. We scaled the learning rate to accommodate per [12].

Future work

Firstly, the usage of reinforcement learning reinforcement learning-based augmentation strategies, such as AutoAugment [4] likely would be highly complementary to our approach. We would plan to extend our approach to other datasets as well, including pretraining on IIT and other large document image datasets.

References

1. Afzal, M.Z., Kölsch, A., Ahmed, S., Liwicki, M.: Cutting the error by half: Investigation of very deep cnn and advanced training strategies for document image classification. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). vol. 1, pp. 883–888. IEEE (2017)
2. Bergstra, J., Yamins, D., Cox, D.D.: Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures (2013)
3. Csurka, G., Larlus, D., Gordo, A., Almazan, J.: What is the right way to represent document images? arXiv preprint arXiv:1603.01076 (2016)
4. Cubuk, E.D., Zoph, B., Mane, D., Vasudevan, V., Le, Q.V.: Autoaugment: Learning augmentation policies from data. arXiv preprint arXiv:1805.09501 (2018)
5. Das, A., Roy, S., Bhattacharya, U., Parui, S.K.: Document image classification with intra-domain transfer learning and stacked generalization of deep convolutional neural networks. In: 2018 24th International Conference on Pattern Recognition (ICPR). pp. 3180–3185. IEEE (2018)
6. DeVries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552 (2017)
7. Harley, A.W., Ufkes, A., Derpanis, K.G.: 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 (2015)
8. Huang, W.R., Qi, Y., Li, Q., Degange, J.L.: Deeperase: Unsupervised ink artifact removal in document text images
9. Smith, L.N.: A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay. arXiv preprint arXiv:1803.09820 (2018)
10. Tensmeyer, C., Martinez, T.: Document image binarization with fully convolutional neural networks. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). vol. 01, pp. 99–104 (Nov 2017). https://doi.org/10.1109/ICDAR.2017.25
11. Tensmeyer, C., Martinez, T.: Analysis of convolutional neural networks for document image classification. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). vol. 1, pp. 388–393. IEEE (2017)
12. You, Y., Zhang, Z., Hsieh, C.J., Demmel, J., Keutzer, K.: Imagenet training in minutes. In: Proceedings of the 47th International Conference on Parallel Processing. p. 1. ACM (2018)
13. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017)

4 Appendix

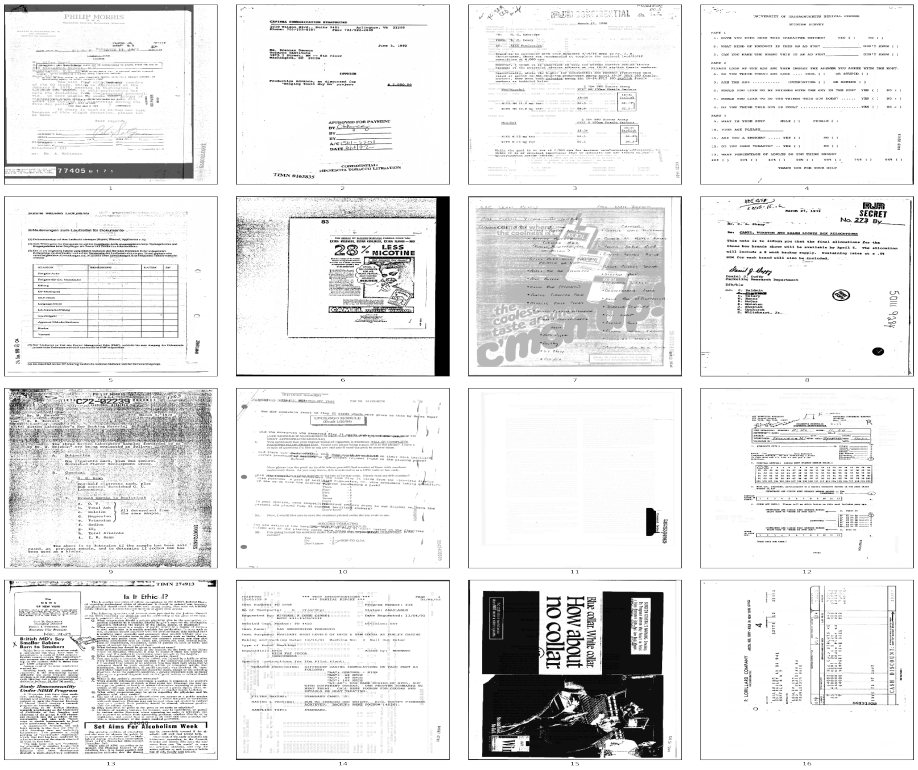

Fig. 1. Example noised images generated from augmentation strategy.