A deep neural network for multi-label fine-grained classification using the Metropolitan Museum of Art digitized collection.

Ksenia Sokolova[1]
[1] Princeton University, Department of Computer Science, Princeton, NJ

PRINCETON UNIVERSITY

ABSTRACT

Digitizing large museum collections is a labor intensive task, that would benefit from automatic or semi-automatic labeling of the artworks. This project uses the digitized and annotated collection of art by the Metropolitan Museum of Art [1] to train a deep convolutional neural network to predict 1103 labels. The best model achieves the mean F2-score of 0.533 and AUC of 0.94 on the validation set. Additionally, the first 8 layers of trained model can be used as a content aware crop tool for object photographs and allow to gain an insight into the activations of the network.

BACKGROUND AND METHODS

Goal

- Multi-label prediction: each object may belong to multiple categories.
- 1103 categories total: 398 'culture' labels and 705 'tag' labels. Culture describes the region of origin, while tag describes the objects shown.

INPUT

Capitol of VA in Richmond, from the General Government and State Capitol Building series for Allen & Ginter Cigarette Brands, 1889

OUTPUT

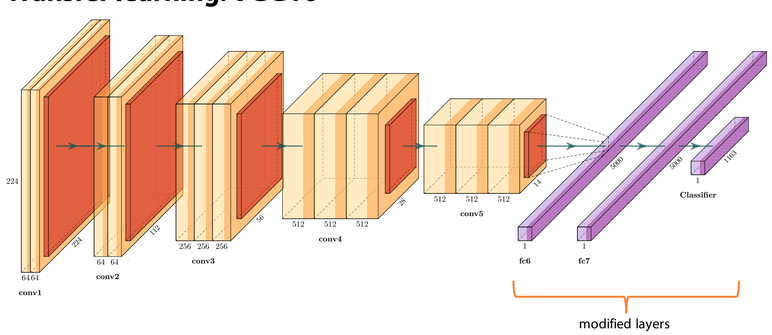
Culture: American
Tag: buildings
Tag: columns
Tag: pediments*

(Pediment: triangular upper part of the front of the building in classical style)

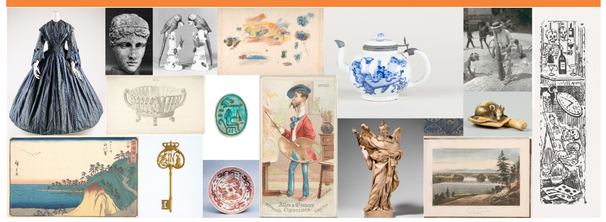
Convolutional Neural Networks

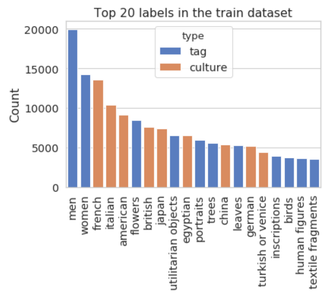
- Class of deep feed-forward neural networks commonly used for analysis of visual data
- When there is a small number of training images available, transfer learning is used. It involves using network weights pre-trained on other dataset. This project uses VGG16 [2] network trained on ImageNet

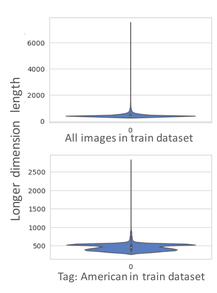
Transfer learning: VGG16



DATASET





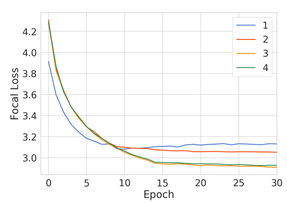


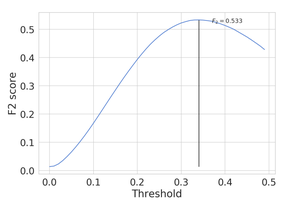
TRAINING AND SELECTING MODELS

- Special evaluation metrics: F2-score and Focal Loss [3].

$$F_2 = 5\frac{precision * recall}{4 * precision + recall} \qquad FL(p_t) = -(1-p_t)^{\gamma}\log(p_t), p_t \text{ is probability of error of the prediction}$$

- Best F2 score for all models was found by search over different thresholds.
- In total, compared 4 different model settings. The best model was VGG16 with batch normalization and intermediate layer of 5000 in the fully connected stack, learning rate of 0.01 with a step learning rate decrease of 0.1 every 15 epochs and batch size of 128.



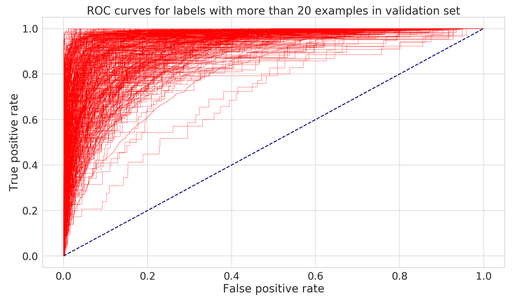


Left: Validation error vs epoch; **Right**: F2 score vs the threshold for accepting label for model 1

		Avg. AUC	F2	Loss	Batch size	Lr rate	StepLR
1	VGG16	0.932	**0.533**	3.082	10	0.001	0.1/10
2	VGG16_bn	0.928	0.51	3.052	128	0.01	0.1/10
3	VGG16_bn	**0.94**	**0.533**	**2.91**	128	0.01	0.1/15
4	VGG16_bn*	**0.94**	0.53	2.92	128	0.01	0.1/15

* 6000 neurons in fc layer

MODEL PERFORMANCE AND CONTENT AWARE CROP



Label	# in Val	AUC
tag::men	3991	0.911401
tag::women	2831	0.916577
culture::french	2705	0.875143
culture::italian	2138	0.923297
culture::american	1783	0.969055
tag::flowers	1684	0.93337
culture::british	1543	0.904896
culture::japan	1472	0.980762
culture::egyptian	1323	0.980658

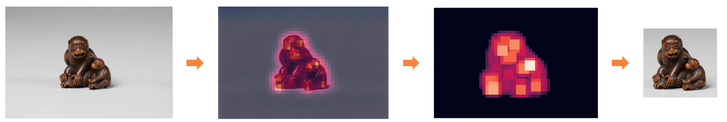
Truth: American, pants
Prediction: American, pants, *clothing and accessories*

Truth: Indonesia, rings
Prediction: Indonesia, rings, *jewelry*

Truth: French, inscriptions, men, profiles, *doctors*
Prediction: French, inscriptions, men, profiles

Truth: Japan, *textiles*
Prediction: Japan, *flowers, leaves, textile fragments*

Examples of images, predictions (by model 1) and ground truth labels from the validation set. Observe different types of discrepancies in labeling.



Input image — Normalized activations after 8 conv layers — Peak localization for automatic cropping — Resulting crop

Example of using convolutional layers of the trained model 1 for cropping photographs of various objects. The feature maps are normalized by layer, and layers with low standard deviation are not used. 'Nearest' interpolation is done to scale feature maps up.

DISCUSSION

- Models achieve good performance overall. Training parameters upon the start play an important role in quality of predictions.
- There are classes that have only 1 or 2 images associated with it. That complicates the analysis and training procedure, and may result in overtraining the model for that class.
- The ground truth labels are noisy, as each image is labeled by one person without checks. This results in the higher loss, when model correctly predicts the label upon manual examination, but the ground truth is missing the label.

FUTURE WORK

- Cleaning the dataset labels in a supervised fashion. Ideally, labelling the dataset at least one more time would be beneficial to estimate the level of human error.
- Using other loss measure: as can be seen, focal loss and F2 score are not directly related.
- Using other pre-trained neural network and/or more modifications in the last layers.

REFERENCES

[1] "iMet Collection 2019: dataset", *Kaggle, https://www.kaggle.com/c/imet-2019-fgvc6/data*
[2] "Very Deep Convolutional Networks for Large-Scale Image Recognition", *K.Simonyan & A.Zisserman,* 2014
[3] "Focal Loss for Dense Object Detection", *Lin, Goyal, Girshick, He & Dollar,* 2017

sokolova@cs.princeton.edu