Detecting Faces in Animes Using Supervised Domain Adaptation

Cem Koc Can Koc Brian Su

Department of Electrical Engineering and Computer Sciences, UC Berkeley

Introduction

Motivation: Domain adaptation is necessary because often times, models trained on a certain large dataset don't perform well on a new target domain. Here we explore adapting pre-trained models on real world images to images from Animes.

Problem Statement: Given an image from an anime, determine whether it contains a face.

Definition of a Face: We consider an image to contain a face if it satisfies the following conditions:

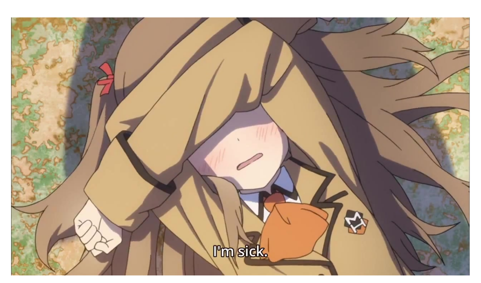
- Face belongs to a human character
- 50% or more of the face is showing
- Facial features can be easily discerned such as eye(s), nose, mouth etc.

Face (Label: 1)

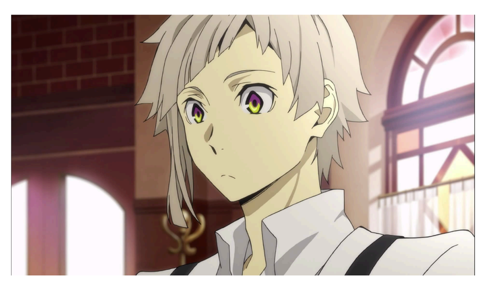
Face (Label: 1)

No Face (Label: 0)

Face (Label: 1)



No Face (Label: 0)



Face (Label: 1)

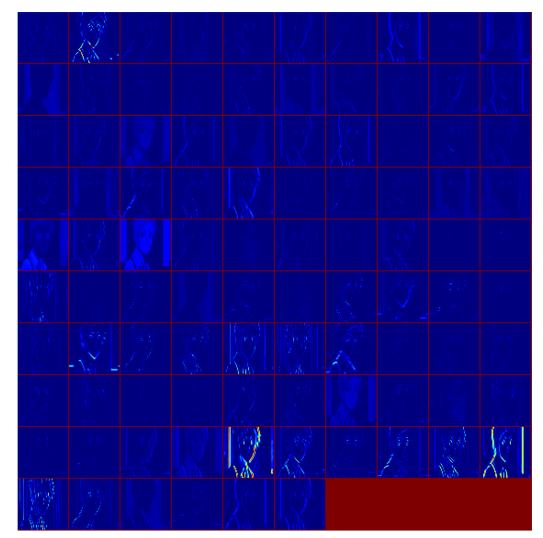
Methods

- **Dataset:** We created our own Anime Dataset that contains randomly sampled frames from 11 different animes and hand labeled them whether they contain face or not.
- We used 5 different CNN architectures, pre-trained on large real world datasets such as ILSVRC 14 [1]. As it was described by Donahue et al. [2], earlier layers of CNNs learn more generic features. Hence we explored extracting features from different layers using Caffe and Keras frameworks to find out which generalized better to our anime domain.
- We use these new anime image features to train an SVM and classify our test images.

Rectified Responses

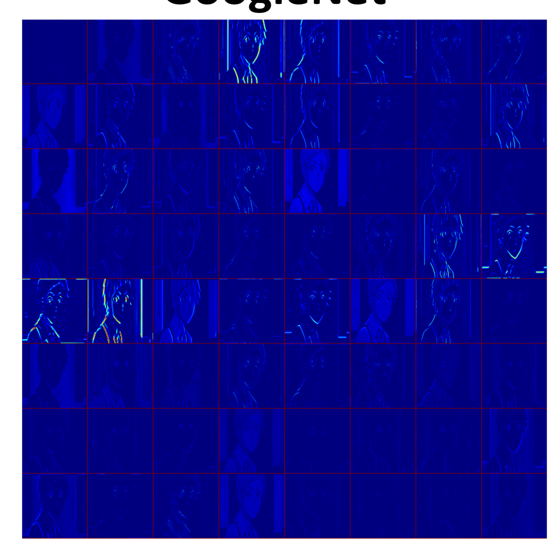
Image

AlexNet



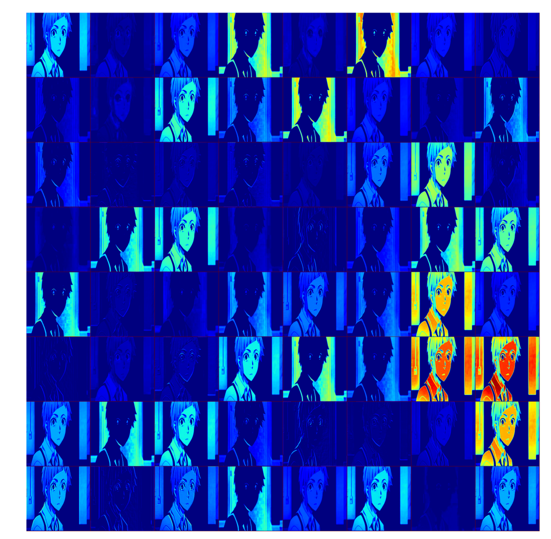
Conv1 Layer

GoogleNet

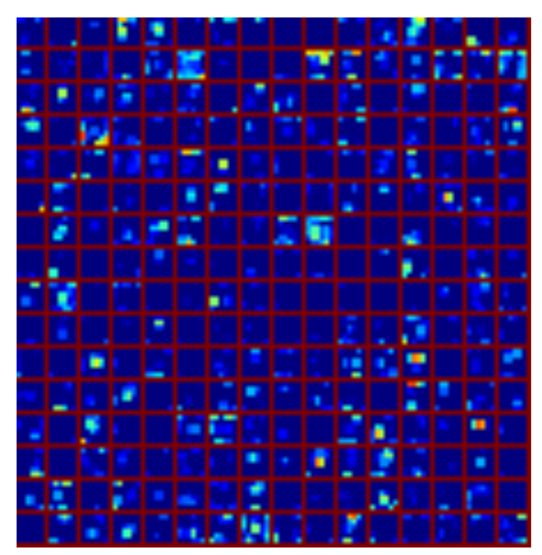


conv1/7x7_s2 Layer

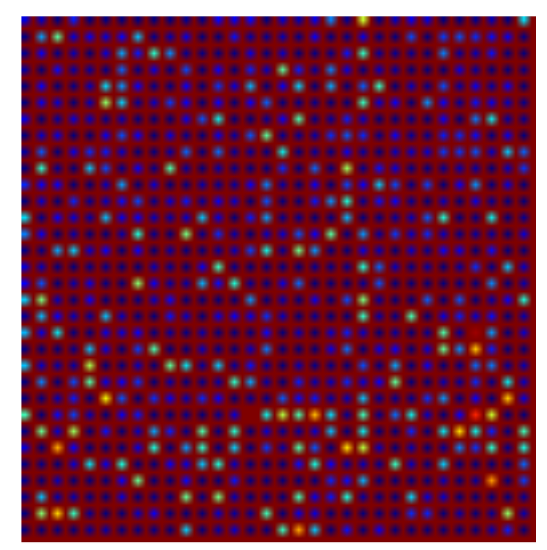
VGG FACE



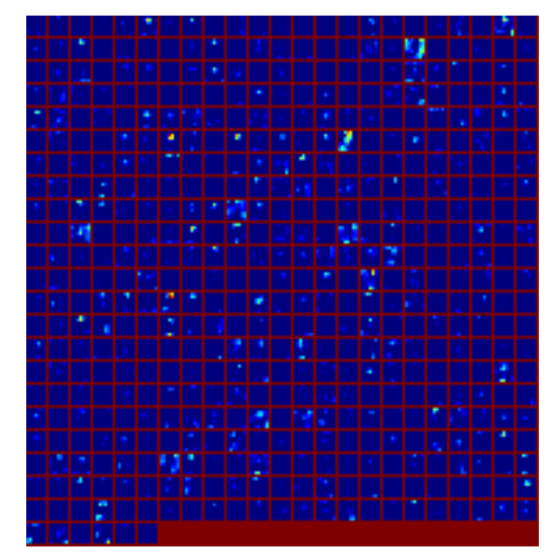
Conv1 _1 Layer



Pool5 Layer



pool5/7x7_s1 Layer



Pool5 Layer

Training/Validation/Test Accuracies

	AlexNet	GoogleNet	VGG 16	VGG 19	VGG-Face
Linear SVC	.941, .922, .778	1, .929, .931	.997, .891, .904	.999, .901, .907	.988, .885, .646
RBF SVC	1, .67, .455	1, .595, .607	1, .615, .583	1, .615, .583	1, .67, .455
Poly deg 3	1, .97, .787	1, .956, .96	1, .938, .953	1, .945, .95	1, .958, .688
Poly deg 4	1, .969, .785	1, .953, .960	1, .930, .94	1, .935, .946	1, .961, .698
Poly deg 5	1, .966, .778	1., .952, .961	1, .923, .933	1, .927, .94	1, .958, .726
Poly deg 6	1, .961, .770	1, .953, .955	1, .917, .916	1, .92, .924	1, .956, .738
Poly deg 10	1, .93, .753	1, .923, .936	1, .868, .853	1, .867, .873	1, .912, .742
Poly deg 20	.385, .331, .545	1.832, .833	.428, .404, .439	.428, .404, .439	.386, .331, .545

Conclusion

In this paper, we followed a feature extraction approach to adapt pre-trained models on large image datasets to perform face detection in anime images. We created a dataset of several thousands of face/no-face images from several animes for this task. Because of our limited dataset, we took the activations of each pre-trained model before its fully connected layer as our input to our support vector models - RBF, linear, and polynomial models of various degrees. Our results were promising and our top SVM models were able to achieve at least 70% test accuracy for each pre-trained model, with our best accuracies of 95-96% obtained from Googlenet, VGG16, and VGG19, using polynomial SVMs of degree 3 and 4. Furthermore, on our distinct test set where images sampled from animes that were not present in the training set, our best model GoogleNet achieves 86% accuracy using SVM using polynomial of degree 3.

Future Work

The similarities and the differences between the real domain and the anime domain provide an interesting domain adaptation problem for the real world trained models. In this paper, we explored the problem of face detection. A natural extension is to perform face recognition - i.e. can we identify characters across various scenes. Not only this

There are much more domain adaptation problems that we'd like the explore beyond this paper. One is simultaneous localization and recognition of objects in the anime domain by using a RCNN. Second is to be able to recognize emotions in anime faces. Although depicted slightly differently than human emotions, a good start for this task may come from Figure 3 which provides a mapping of emotions to faces.

References

- [1] Olga Russakovsky*, Jia Deng*, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg and Li Fei-Fei. (* = equal contribution) ImageNet Large Scale Visual Recognition Challenge. IJCV, 2015.
- [2] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In ICML, 2014.