Analyzing Self Supervised and Supervised Learning Methods on STL-10 Dataset

Vikas Patidar
New York University
vp1274@nyu.edu

Hemant Singh
New York University
hs3644@nyu.edu

Abstract

In this paper, we described various methods to perform classification task on STL-10 dataset. The target of this work[1] was two-fold: first, trying out various self supervised approaches, analyzing and comparing the feature representation for the task in hand and second trying out a novel approach named "Harmonic Networks" with our enhancements to come up with a highly efficient approach for classifying STL-10 data which doesn't use the unlabeled data for training. We performed popular pretext tasks including Rotation Prediction, JigSaw Puzzle Solving, Inpainting under same experimental conditions and downstream architecture and analyzed the essence of features representation learnt during pretext phase for downstream classification task. We found that among all these techniques Jigsaw Puzzle as pretext task provides the most suitable feature learning for classification. Also we found that the pretext models with very high training accuracy perform worse than the models with relatively low training accuracy, which goes on to show that they overfit the data and the features learned are far less meaningful. Finally, we trained Harmonic network and got a classification accuracy of 90.56%.

1 Introduction

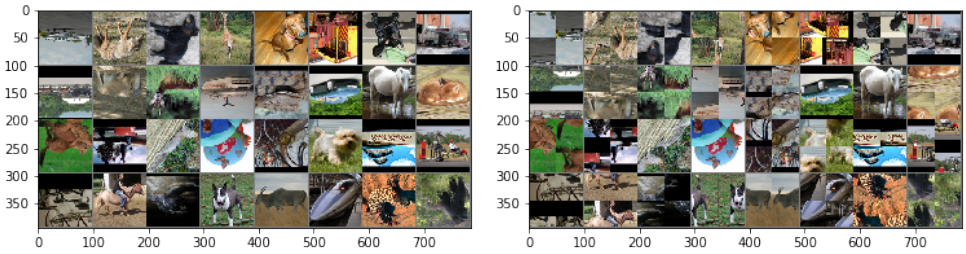
Currently, the deep neural networks have evolved to a level where given a task and enough labels, supervised learning can solve it really well. But collecting labels manually is expensive and hard to scale up. Considering the size of unlabeled data available compared to human annotated data, it can be efficiently used to learn some image representations, which can be coupled with the limited supervised data to improve performance on popular vision tasks like classification, detection, segmentation etc.

[1] https://github.com/hemant-git10/CV-Project

To exploit the vast amount of unlabeled data available we perform self supervised learning. Self supervised learning approach is comprised of two phases: first phase is solving pretext task where design some tasks for network to solve and visual representations are learned by learning objective functions of pretext tasks. Second phase is Downstream task, which is one of the most popular computer vision applications that are used to evaluate the quality of features learned by self-supervised learning. These applications can greatly benefit from the pretrained models when training set is scarce and this is where self supervised learning approaches help. We take trained models from pretext phase and use them for downstream task. Some of the popular pretext tasks include Rotation prediction, Jigsaw Puzzle Solving, Inpainting, Colorization, Relative Position Prediction. In this paper, we use Rotation Prediction, Jigsaw Puzzle solving and Inpainting as pretext tasks and compare the qualities of features learnt for classification task especially on STL-10 dataset. Generally, Convolutional neural networks generally tend to overfit when trained with limited data. Here we observed this behavior when we trained the network just on the training data(5K images). To overcome it and to improve the classification accuracy we used harmonic neural network, which replaces the conventional neural network architecture with harmonic blocks. With this we got an improved classification test accuracy of more than 90%.

2 Related Work

Deep ConvNets have been successful in a lot of object detection and classification tasks as they are very good at learning higher level visual representations. However, to learn these representations, they require a massive amount of training data and



(a) (b)

Figure 1: (a) Original Images (b) 2x2 Jigsaw puzzle

manual labels which is impractical to scale for various real-world tasks. Learning these representations in an unsupervised way has got a lot of interest in the past few years. Self-supervised learning is an interesting way to do that. In this type of learning, a pretext task is trained first with the help of a surrogate supervision signal that is obtained from the visual information present in the image. For example, in (Gidaris et al., 2018), visual representations are learnt by training ConvNets to recognize the geometric transformation that is applied to the image that it gets as input. Images are transformed by rotating them in four ways (i.e. 0, 90, 180, 270 degrees). Then they are trained as a four way classification task and it is expected to learn the object representations by recognising the geometric transformation. In (Noroozi and Favaro, 2016), a CFN (context-free network) is trained to solve jigsaw puzzles, from which it learns both feature mapping of object parts as well as their correct spatial arrangement. Solving jigsaw puzzles is the pretext task and classification and detection are the target tasks. We train the model by permuting girds in the puzzle to learn the ground truth (i.e actual setting of the patches). The cardinal idea here is to teach the model that an object is made up of parts and what they are by solving jigsaw puzzles and make use of permutation invariance property of CNNs. These learned visual representations are used for other object detection and classification tasks. In (Pathak et al., 2016), a convolutional neural network is trained to find the missing pixels in the given image. They trained a context encoder to capture the context of an image into a compact latent feature representation to produce the missing image content. On the encoder side, they showed that encoding just the context of an image patch and using the resulting feature to retrieve nearest neighbor contexts from a dataset produces patches which are semantically similar to the original patch. The encoder here learns higher level representations in the last layers which are further augmented by fully connected layers to solve other classification and detection tasks.

The STL-10 dataset (Coates et al., 2011) is an image recognition dataset for developing unsupervised feature learning, deep learning, self-taught learning algorithms. Each class has fewer labeled training examples than in CIFAR-10, but a very large set of unlabeled examples is provided to learn image models prior to supervised training. The primary challenge is to make use of the unlabeled data to build a useful prior. In (Wang et al., 2019), Ensemble of Auto-Encoding Transformations (EnAET) is trained to learn from both labeled and unlabeled data based on the embedded representations by decoding both spatial and non-spatial transformations. They obtained an error rate of 1.99% on CIFAR-10 and 4.52% on STL10. In (Berthelot et al., 2019), they unified the current dominant approaches for semi-supervised learning to produce a new algorithm, MixMatch. MixMatch obtains state-of-the-art results by a large margin across many datasets and labeled data amounts. On CIFAR-10, they reduce error rate from 38% to 11% and by a factor of 2 on STL-10. In (Ji et al., 2019), a novel clustering objective that learns a neural network classifier from scratch from unlabelled data samples. The model discovers clusters that accurately match semantic classes. It obtained an accuracy of 88.8% on STL10 classification.

3 Methodology

3.1 Rotation Pretext Task

Unsupervised feature learning is of crucial importance in order to successfully harvest the vast

(a) (b) (c)

Figure 2: (a) Original Image (b) Pixelwise Dropout (c) Concentrated Dropout

amount of visual data that is available today. We learned image features here by training CNN to recognize the 2d rotation that is applied to the image that it gets as input. This apparently simple task actually provides a very powerful supervisory signal for semantic feature learning. We used FAIR Self-Supervised Learning Integrated Multimodal Environment (SSLIME) environment to learn the image representation using different architectures including custom Alexnet, VGG and Resnet.

Figure 3: Rotation Prediction

3.2 Jigsaw Puzzle Solving as Pretext Task

A $n \times n$ grid is randomly selected from a jigsaw puzzle. A permutation from this grid is extracted by rearranging the patches and the model is trained to output n^2 probabilities for each index. Multiple puzzles (i.e. $n \times n$ grids) are extracted from the input image to make sure that the features would carry semantic meaning. The experiment was tried with 2×2 and 3×3 patches with 1000 permutations. The network processes each patch independently with shared weights and outputs a probability vector per patch index out of a predefined set of permutations. To control the difficulty of jigsaw puzzles, patches are shuffled according to a predefined permutation set and the model is configured to predict a probability vector over all the indices in the set.

3.3 Inpainting as Pretext Task

Inpainting is one of the most important tasks for meaningful representation learning in self supervised setting. We used custom designed autoencoders for pretext tasks. For all the below mentioned tasks, we train an autoencoder to reconstruct the image from given input. Once the pretext phase is done, we take the encoder module of the network and add fully connected layers to it. Now we freeze the encoder module and train the fully connected layers in the network with the help of supervised images and use this fine-tuned model for classification on the test set.

3.3.1 Plain Image reconstruction

In this task we train an autoencoder to reconstruct the image and use it in our pretext phase. This reconstruction pretext task gives training accuracy of 99.49 and test accuracy of 98.98. But using this trained encoder for downstream classification task doesn't perform well with test accuracy of around 32%, which shows that although the auto encoder learnt the features well for image reconstruction, those are not representative features for image classification.

3.3.2 Image Reconstruction with Random Dropout

In this task we randomly apply pixelwise dropout in the image and try to reconstruct the image using the same network and architecture. We kept the same architecture for these experiments as the purpose of conducting these experiments is to know the quality of features learned for representation learning and how suitable these features

are for classification task. We tried bunch of values for dropout ranging from 30-75. Once we get the models from reconstruction task based on the dropout values, we compared their performance on the downstream task.

3.3.3 Image Reconstruction with Concentrated Dropout

We take fixed size patches in the image and apply dropout to those patches randomly. The main thought behind this idea is that as a random patch is selected and dropout is applied at each training step, the features learnt exhibit spatial locality, which are very useful for classification, detection and segmentation tasks. Again we reconstruct the images using the same architecture as in previous experiments for consistent comparison. We try different dropout values and patch sizes to train the network and compare models' performance on classification task.

3.4 Harmonic Networks

Convolutional neural networks extracts correlation of input image with the applied filters. The pixels in the local neighbourhood of images are observed to be correlated and the applied convolutions will produce correlated signals. This will tend to overfit the model in a limited amount of training data. So we apply transformation methods to decorrelate signals forming an image. The use of predefined filters reduces the impact of overfitting and decrease computational complexity. Harmonic Networks consist of harmonic blocks which rely on using windowed cosine transform at various frequencies in place of learned filters. These harmonic blocks learn weights to combine several frequency outputs and dropping out the uninformative frequencies to improve the computational complexity of the model without comprising the performance. The harmonic blocks work in 2 parts - firstly, the input features undergo harmonic decomposition by a Discrete Cosine Transform(DCT). In the second stage, the transformed signals are combined by learnable weights. Control over the filters allows selecting subsets of filters to approximate the signal. We here replaced the Resnet50 architecture with the harmonic blocks. The difference from standard convolutional network is that the optimization algorithm is not searching for filters that extract spatial correlation, instead learns the relative importance of preset feature extractors(DCT filters) at multiple layers.

Algorithm 1: Harmonic block

Input: h^{l-1}

for $n \in \{0, \cdots, N-1\}$ **do**

$z^l_{n,u,v} \leftarrow \sum_{u=0}^{K-1} \sum_{v=0}^{K-1} \psi_{u,v} * * h^{l-1}_n$

if *normalize* **then**

$\mu^l_{n,u,v}, \sigma^l_{n,u,v} \leftarrow$ estimate mean and standard deviation of $z^l_{n,u,v}$ over the batch dimension

$z^l_{n,u,v} \leftarrow \frac{(z^l_{n,u,v} - \mu^l_{n,u,v})}{\sigma^l_{n,u,v}}$

end

end

for $m \in \{0 \dots M-1\}$ **do**

$h^l_m \leftarrow \sum_{n=0}^{N-1} \sum_{v=0}^{K-1} \sum_{u=0}^{K-1} w_{m,n,u,v} \; z_{n,u,v}$

end

Output: h^l

Figure 4: Harmonic Block

Discrete Cosine Transform(DCT): DCT is a transformation method that decomposes an image to its spatial frequency spectrum. It expresses the decomposition in terms of sum of cosine functions oscillating at different frequencies. The contribution of each cosine function is determined by its coefficient during the transformation. DCT is also used for image and video compression.

In all above experiments during the training of downstream task and in harmonic networks, data augmentation techniques were applied to increase the labelled dataset.

4 Dataset

All the experiments were performed on STL-10 dataset, which consists of 10 classes, 500 training images per class, 800 testing images per class and 100K unlabeled images. Each image is of dimension 96×96.

5 Experiments & Results

5.1 Rotation Prediction

For Rotation Prediction pretext task, we tried multiple architectures including Alexnet, VGG-16, Resnet-50. Out of these Resnet-50 gives the best accuracy of over 73 %. The results are shown in Table 1.

5.2 Jigsaw Puzzle

We used 2×2 and 3×3 patches for jigsaw puzzle solving task. The results of these experiments are shown in Table 2. The results show that although the training accuracy is higher for 2×2 puzzle,

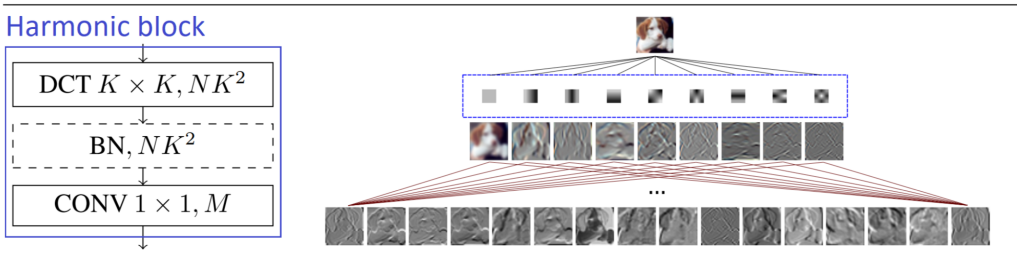


Figure 5: Harmonic Network

Pretext Model	Classification Accuracy
Alexnet	60.78
VGG	63.7
Resnet	**73.2**

Table 1: Rotation Prediction Results

Patch Size	Pretext Train Accuracy	Classification Test Accuracy
2x2	96.4	45.3
3x3	93.1	67.8

Table 2: Jigsaw Puzzle Results

accuracy on STL-10 test data is way lower compared to the 3×3 puzzle setting, which suggests that the features learnt during the 2×2 setting are not informative for classification task.

5.3 Inpainting

5.3.1 Plain Image Reconstruction

The plain reconstruction task gives very high train accuracy of over 98 in few epochs, but it gives poor test accuracy of 32%, which shows that plain image reconstruction task is not a suitable task for classification.

5.3.2 Pixelwise Dropout

We try different values of dropout ranging from 0.45 to 0.8 on the whole image. Low dropout value of 0.45 gives high pretext train accuracy, but performs badly on downstream classification task. High dropout value of 0.8 gives relatively lower training accuracy values in the pretext phase and low test accuracy in classification phase too. So to find the optimal dropout value we repeat the experiment with multiple dropout values and find p = 0.6 as the optimal value for dropout, which provides highest classification accuracy. After this we show that even for the pretext tasks, which learn meaningful feature representation for downstream task, better training accuracy doesn't mean higher end task test accuracy. We train the model with optimal dropout value of 0.6 and take three screenshot of the trained model for STL-10 classification task after 50, 100 and 300 epochs respectively. Table V shows that 50 and 150 epoch models features underfit and overfit the data respectively. The results of this experiments are shown in Fig 2 and Fig 3.

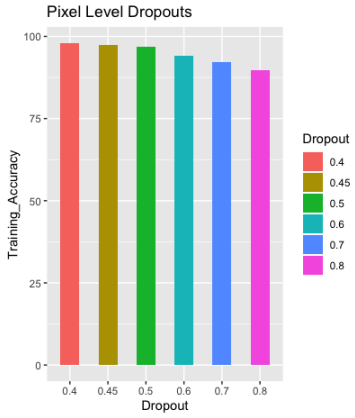


Figure 6: Pixel level dropout with different dropouts

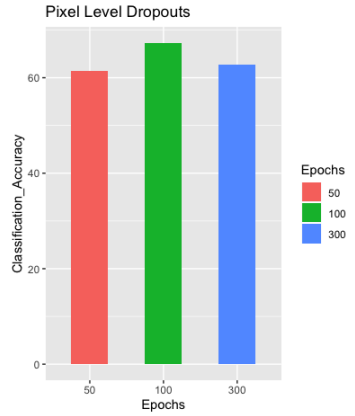


Figure 7: Effect of Pretext training on classification

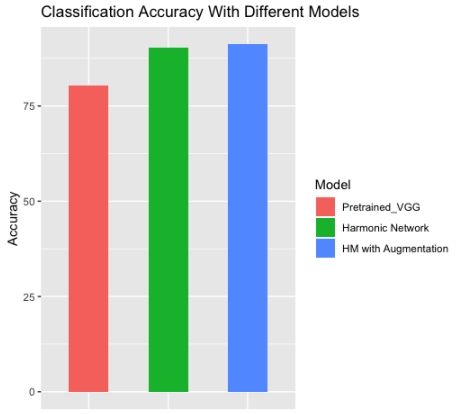


Figure 8: Harmonic Network Accuracy on STL-10 Test data

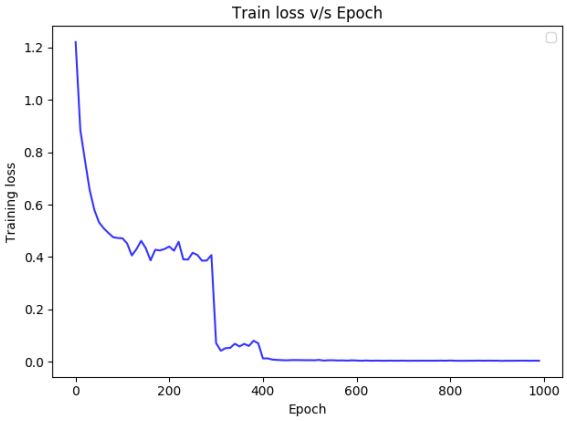


Figure 9: Harmonic Network Train loss

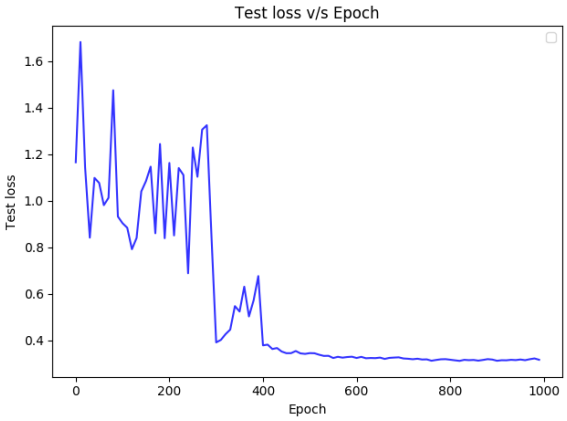


Figure 10: Harmonic Network Test loss

Dropout	Pretext Train Accuracy	Classification Accuracy
0.4	**97.6**	41.6
0.45	96.2	44.1
0.5	94.7	53.5
0.6	91.3	**66.7**
0.7	88.9	54.5
0.8	86.7	42.2

Table 3: Concentrated Dropout Results

5.3.3 Concentrated Dropout

We take a fixed patch size and randomly choose this patch in the image. In the chosen patch we apply random dropout and try to reconstruct the original image using same network architecture, which was used in Pixel wise dropout experiment. Similar to previous experiment we try a range of dropout values and for the optimal dropout value we find classification results with models trained on different epochs. The results are tabulated in tables 3.

5.4 Harmonic Network

The Harmonic networks have a variety of model and training parameters which can be changed to see the effect on the classification accuracy. We performed various experiments by changing model depth, width, dropout, batch size, learning rate etc. We replaced the convolutional blocks in conventional ConvNets with harmonic blocks. We tried with architectures of VGG and Resnet50, but Resnet50 provided the best accuracy. The results are plotted in figures 8, 9 and 10.

6 Conclusions

- We performed the classification task first using self supervised learning with 3 pretext tasks - Rotation prediction, Jigsaw and Inpainting. Out of these 3, Rotation prediction proved out to be the best pretext task for classification.

- Higher training accuracy on Pretext task doesn't show correlation with the higher test accuracy on downstream task. This shows that features learnt with relative less fitting the data in pretext phase, comprise of more meaningful feature representations compared to the features learnt while overfitting the data.

- During Jigsaw task, moving from patch size of 2×2 to 3×3 showed significant improvements in the test results, which shows that jigsaw with larger patch sizes learns more representative features for classification task. Further studies can be done to see whether increasing the patch size even further improves the test results.

- Concentrated Dropout results show that applying dropout in concentrated window helps in learning better spatial features compared to applying dropout on complete image, which are important for object detection, classification and segmentation tasks.

- Harmonic Network is highly efficient and optimized network architecture for the classification task, which only makes use of the supervised training data for learning and gives excellent test accuracy of over 90 percent.

References

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. 2019. Mixmatch: A holistic approach to semi-supervised learning. *arXiv preprint arXiv:1905.02249*.

Adam Coates, Andrew Ng, and Honglak Lee. 2011. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. 2018. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*.

Xu Ji, João F Henriques, and Andrea Vedaldi. 2019. Invariant information clustering for unsupervised image classification and segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9865–9874.

Mehdi Noroozi and Paolo Favaro. 2016. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, pages 69–84. Springer.

Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. 2016. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544.

Xiao Wang, Daisuke Kihara, Jiebo Luo, and Guo-Jun Qi. 2019. Enaet: Self-trained ensemble autoencoding transformations for semi-supervised learning. *arXiv preprint arXiv:1911.09265*.