Self-Supervised Representation Learning by Rotation Feature Decoupling: Implementation

Abraham Jose
ID :5068109, CAP6614
abraham@knights.ucf.edu

Abstract

The paper deals with unsupervised semantic feature learning without requiring manual annotation effort to detect the rotation in the image. These unsupervised learning approaches no longer requires manual annotation efforts to learn crucial features in vast amount of data and to harvest them. They also proves that the simple task of unsupervised detection of rotation using RotNet provides a very powerful supervisory signal for semantic feature learning.They were able to drastically reduce the gap between unsupervised and supervised feature learning benchmarks throught RotNet. Further using the RotNet they were able to mitigate the influence of rotation label noise, as well as discriminate instances without regard to image rotations. By decoupling the rotation discrimination from instance discrimination, the resulting feature has a better generalization ability for more various tasks such as object recognition, object detection, and object segmentation. We will implement RotNet and then will decouple the rotation related noise by decoupling the rotation depended features in the classifier model.

1 Introduction

Deep learning in computer vision has helped us make great strides that we couldn't have done without it. The 2D feature extraction and feature response or activation when cascaded helped us extract higher level feature vectors for many of the computer vision tasks including Segmentation, Detection and Classification. However, learning images in hyper-dimentional feature space makes it really hard to understand the dynamics and nature of the Deep Convolution Neural network model. Rotation is one of many well studied property of the visual representation, which is rarely appreciated and exploited in the context of deep convolution network based self-supervised representation learning methods. By decoupling the rotation specific parts/features from the images, we can extract the rotation unrelated parts/features in the image and make a model out of it which will make the model robust to noise. We can split model by jointly predicting image rotations and discriminating individual instances which will create make it orientation agnostic model.
In classical computer vision, these efforts can be compared to features like SIFT and RIFT which are rotation invariant and thus will give robust result for samples with different orientations. The unsupervised rotation feature decoupling sheds light to make the deep neural model rotation invariant, which makes the model robust to rotation related noises.

2 Model Architecture

RotNet: RotNet is the previous work from the author to learn the orientation of natural image(rotation in particular) in unsupervised manner. This unsupervised model is the state-of-the-art self-supervised model available up-to-date for extracting rotation related features. However if the image sample is rotation agnostic, then it causes severe noise in the training label.These label

noises(rotation related) in unsupervised training are modeled as a positive unlabeled(PU) learning problem, which learns from the weight of instance to mitigate the influence of rotation ambiguous/agnostic images. ie, we are accounting for images that are not rotation invariant. Using PU we can use the relationship between the conditional probability and its estimation to model the mislabeled rate. We can further use them to exclude the image samples with low confidence, labeling image samples with high probability or by reweighting the samples. Here we will go for the third technique of reweighting the noisy image samples/labels.

In a given training set $S = \{X_i\}_{i=1}^{N}$ of N images, the rotation transformation $G = \{g(X; y)\}_{K}^{y=1}$ for each image X. The i^th images Y^th rotation can be denoted as $X_{i,j} = g(X_i; y)$. For rotation transformation, we have ConvNet model $F(.; \theta)$ which will be trained to classify the rotation of image with an objective.

$$min_\theta \frac{1}{NK} \sum_{i=1}^{N} \sum_{y=1}^{K} l(F(X_{i,j}; \theta), y)$$

However, we will use the PU learning to mitigate the effects of noisy labels by updating the weight of the i^{th} image as follows based on PU learning where $\tilde{F}(Xi, y)$ is the probability of an image being positive estimated from the pre-trained model.

$$w_{i,j} = \begin{cases} 1 & y = 1 \\ 1 - \tilde{F}(Xi, y)^\gamma & otherwise \end{cases}$$

Reweighted model objective can be represented as follows which will mitigate the rotation noises in the image.

$$min_\theta \frac{1}{NK} \sum_{i=1}^{N} \sum_{y=1}^{K} w_{i,j} l(F(X_{i,j}; \theta), y)$$

AlexNet: The AlexNet architecture competed on ImageNet Large Scale Visual Recognition Challenge achieving the state-of-the-art results burgeoning era of Deep Learning in Computer Vision. We will use the PU probabilities using RotNet(based on AlexNet) architecture to create a decoupling model extending AlexNet architecture as a decoupling model of rotation related and rotation non-related parts. This will ensure that the sample images has least label error. The network architecture is as follows.

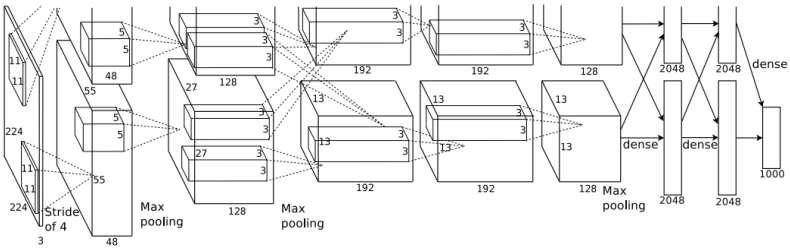

Figure 1: AlexNet Architecture with 60 million parameters and 650,000 neurons

Rotation Irrelevance using AlexNet: We have to decouple the input image features to rotation related and non-related features. ie, for an image X, we have high-level features $f = [f^{(1)^T}, f^{(2)^T}]$ where $f^{(1)}$ be the rotated related feature and $f^{(2)}$ as rotation non-related feature. For rotation features, we will use the loss that gives least error as per equation(2). For rotation irrelevance, we enforce the similarity between features of the same image in multiple rotation angles. So the rotated copies of image, $\{X_y\}_{y=1}^{K}$ has same high level features $\{fy^{(2)}\}_{y=1}^{K}$. Thus the loss will be difference between each images $\{fy^{(2)}\}_{y=1}^{K}$ and the mean feature vector, $\bar{f}$.

Image Instance Classification: From the features extracted from Rotation Irrelavant modeling, Since these features $f^{(2)}$ are already rotation irrelevant, we can use these features to classify the image.

To train the model we will use the negative log-likelihood loss, $L_n = -\sum_{i=1}^{N} logP(i, \hat{f})$ where $\hat{f}$ is the L_2 normalized version of $\bar{f}$. Mean feature will be linearly mapped to a 128-dimensional vector before normalization, NCE and proximal regularization. The loss function for the decoupling model will be to minimize $min_{\theta_f, \theta_c} \lambda_c L_c + \lambda_r L_r + \lambda_n L_n$.

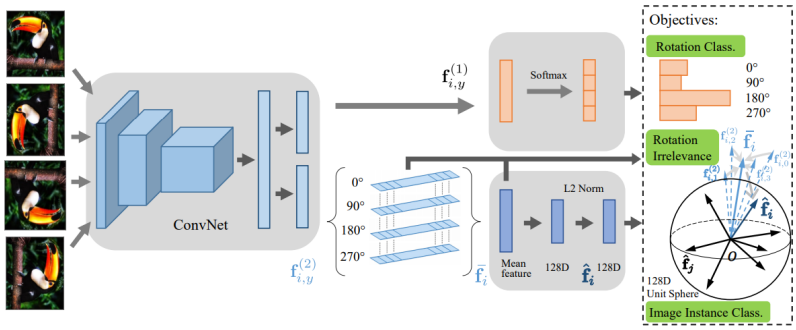

Figure 2: Proposed Architecture

3 Experiments

I have followed the same implementation details from the original paper. Used the ImageNet data-set with 1000 classes for test. Both for the RotNet and the Decoupling model. The original implementation was using batch size of 192 which has been furthered down to carry out the experiments.

RotNet Implementation From Unsupervised Representation Learning by Predicting Image Rotation' paper from Spyros Gidaris et al. AlexNet serves as a primary backbone for the RotNet's feature extractor. Stichastic Gradient Decent optimization is used with batch size 192, momentum 0.9 and weight decay $5e-4$ and l_r of 0.01. Rotated inputs are given at each mini-batch that we use. The model returned top-1 training accuracy in the ImageNet dataset. We are using the PU Learning probability from the github repo available for the implementation which is saved as prob.dat.

Decoupling Model The decoupling model has the feature extracting backbone from AlexNet as we already calculated the prob.dat for the PU probability to mitigate the error due to wrong labeling in the input sampling. Further we will use the discussed training losses to learn and to decouple Rotation responses, Rotation Irrelevance and Image Instance classes which has rotation features mitigated using the Alexnet backbone to extract the features accounting for the wrong labels.

4 Results

Implemented the Image-Net Decoupling model extending the Alexnet architecture as given in Figure 2. We will start with same learning parameters for all feature, classifier and norm(128D mean vector) of network. The initial parameters set for learning are as follows for all of these 3 instances of learning and details about data-set. We are using Nesterov's accelerated gradient decent which will help us to reach optimum by overshooting in the direction of the momentum.

Optimizer	Stochastic Gradient Decent
lr, momentum	0.1, 0.9
weight_decay	0.0005
LUT_lr	[(90, 0.01), (130, 0.001), (190, 0.0001), (210, 1e-05), (230, 0.0001), (245, 1e-05)]
batch size	192
no of epoch	245
no of training & validation images	1281167, 50000
no of classes	1000

However, the model training is still on progress as the training data-set is ImageNet and based on the resources that could be allocated(single node, single GPU of 32G) the model takes almost 8 hours in

an average to complete an epoch on 1281167 images and 50000 validation images. Hence I am not able to provide the final results on ImageNet data-set, which includes layer-wise information retrieval. Statistics on last epoch is as follows:

Training stats: prec_cls: 42.8316, loss: 11.92, loss_cls: 1.2359, loss_mse: 0.0007, loss_nce: 10.6834
Evaluation stats: prec_rot: 56.748, prec_inv: 24.9892, loss_rot: 1.018, loss_inv: 1.4003,

Configurations : Image Instance Classification, Linear Classifier from the filtered rotation non-related features and Non-Linear Classifier on the features in ImageNet were trained *Decoupling-Linearclassifier.slurm, Decoupling-NonLinearclassifier.slurm, Decoupling.slurm* using the given slum files. Results from stdio.out are available in the .out file during training.

5 Conclusion

The presented unsupervised representation method(RotNet) learns semantically meaningful features which are rotation dependent and rotation independent. Thus it is possible to decouple the features and use the rotation agnostic features for training classifier, which will make the model robust to rotation as opposed to feeding rotated images for training the model. The feature decoupling techniques methods opens a whole new world to further incorporate well-analyzed properties like rotation, contrast, etc.. and to extract property agnostic features to learn robust representations for tasks.

References

[1] Zeyu Feng, Chang Xu, Dacheng Tao; "Self-Supervised Representation Learning by Rotation Feature Decoupling" The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10364-10374

[2] Spyros Gidaris, Praveer Singh, Nikos Komodakis "Unsupervised Representation Learning by Predicting Image Rotations", ICLR 2018 Conference.

[3] Krizhevsky, Alex; Sutskever, Ilya; Hinton, Geoffrey E. (2017-05-24). "ImageNet classification with deep convolutional neural networks". Communications of the ACM

[4] https://github.com/philiptheother/FeatureDecoupling

Appendix

All relevant codes used for this experiment to reproduce the result is attached as zip file.