Unsupervised Representation Learning With Deep Convolutional Generative Adversarial Networks

Authors: Alec Radford, Luke Metz & Soumith Chintala
Publisher: Computing Research Repository
Year: 2015

Keywords: DCGAN,

1. Abstract:

"In recent years, supervised learning with convolutional networks (CNNs) has seen huge adoption in computer vision applications. Comparatively, unsupervised learning with CNNs has received less attention. In this work we hope to help bridge the gap between the success of CNNs for supervised learning and unsupervised learning. We introduce a class of CNNs called deep convolutional generative adversarial networks (DCGANs), that have certain architectural constraints, and demonstrate that they are a strong candidate for unsupervised learning. Training on various image datasets, we show convincing evidence that our deep convolutional adversarial pair learns a hierarchy of representations from object parts to scenes in both the generator and discriminator.Additionally, we use the learned features for novel tasks - demonstrating their applicability as general image representations."

2. Introduction:

- Computer vision has adapted many of the CNN applications with supervised learning, whereas unsupervised learning had been neglected.
- So, the authors input the data set into GANs to create the features from generators and discriminators to later use them for supervised tasks.
- For image classification, the trained discriminators were used and the results were compared with remaining unsupervised learning approaches.
- Moreover, as in word vectors, the authors also implemented algebraic additions with the images to output the resultant image with required features.

3. Related Works:

- As said before only some of the works had been previously focused on the unsupervised learning. Out of which clustering the data had been a traditional approach.
- Some of the continued approaches like auto-encoders, ladder structures had resulted in showcasing the better features of the images.

- Generally, the models were made to compare the objects or match the existent images with the database and further applied in super-resolution, texture synthesis and in-painting etc.,
- Usage of GAN's had also been advanced, but the model had been resulting in the noisy images and unstability.

4. Content and Approach:

- The previous works to generate the combination of GANs and CNNs had been failed, which later limited to generating low resolution images.
- To accomplish the task, the authors had implemented three core modifications as follows. The first is that to substitute strided convolutions in place of the spatial polling functions.
- Next, is to remove the fully connected layers which were above the convolutional features generated. This is because as the authors observed one of the example global average pooling, it elevated the stability but affected the speed of convergence.
- Then comes the Batch Normalization(BN), which is robust and stable by normalizing the input to mean and variance to 0 and 1 respectively.
- In contrast BN had created instability when applied to all the layers. So, when the BN is eliminated for the output of generator and input layer of discriminator, it solved the problem.
- Instead of maxout activation function used in the reference paper GAN, the authors observed that the leaky rectified activation to worked better.
- The pre-processing is not implemented to the model. The model was trained with SGD optimization with the mini batch size of 128.
- The weights were normalized with zero mean and 0.02 variance. Normal distribution was used for the weights.
- The momentum of 0.9 caused the oscillation of training, which stabilized later when the momentum decreased to 0.5.
- The model had been trained on three different datasets, one consisting of bedrooms, other with faces and the last one with numbers.
- The LSUN dataset which consists of 3M of bedrooms images had caused overfitting, causing noise in the generated images. No data augmentation was also applied during the training to the images.
- For the face dataset, Open CV face detector was run and stabilized the resolution of the faces with 3.5M face boxes. Even for this data set the augmentation had not been applied.
- The popular method to evaluate the unsupervised training is to retrieve the features and apply them to supervised learning. Then comparison is done by the linear models befitting onto these features.

- So to evaluate the unsupervised learning, they had used the K-means algorithm and extracted the features by training Imagenet dataset.

5. Results:

- This unsupervised learning had resulted in the best of 82.8% accuracy in matching the features with supervised learning.
- That accuracy is in fact achieved with the minimum feature units of 512 compared to all other K-means algorithms.
- The other surprising result, the authors encountered is that algebraic expressions, (addition, subtraction) of features resulting in the images with resultant feature.
- The Street View House Numbers dataset had given only the 22.8% of error rate which is far less than any other architectures.
- The trial of removing some features from LSUN such as windows, the resultant images were also created without images, from which we can state that the architecture had been considering the objects as features.

6. Conclusion and Future Work:

- Despite of little noise and instability of the results, the architecture had been more robust and stable for many of the applications in computer vision.
- Moreover, the application of GAN to unsupervised learning and extracting features followed by supervised learning had resulted in generative model. This model as referred in section 4, can be applied for various operations.
- Future works can be focused on decreasing the instability of the resultant vectors and oscillation through the process. Further extension can be done to various fields of audio and video inputs.
- Moreover the works on latent space would also be exciting to develop. Some of the works during the process were also left to future work, such as fine tuning the discriminator descriptions.