Summaries for "ImageNet Classification with Deep Convolutional Neural Networks"

Weituo Hao 109241801
Stony Brook University
Computer Science Department
Advanced Computer Vision

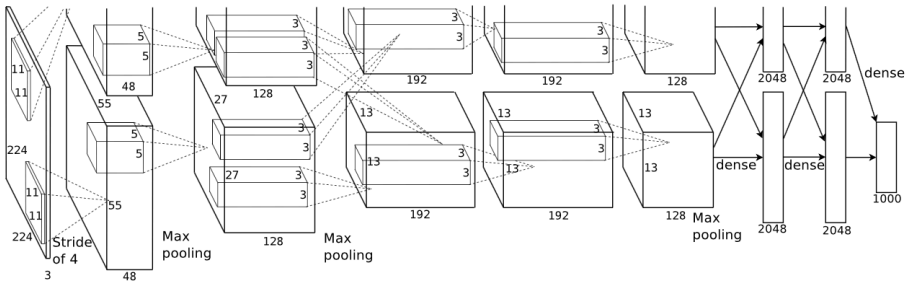
Contribution

The author mainly implemented an efficient GPU computing mode and used it to train a large, deep convolutional neural network to classify the 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into 1000 different classes and achieve top-1 and top-5 error rates of 37.5% and 17.0%

Main idea

The method is convolutional neural networks. But the author makes following changes to the usual CNNs architecture.
First, the author adopts f(x)=max(0,x) instead of f(x)=tanh(x) or f(x)=$(1+e^{-x})^{-1}$, and claims that CNNs with this ReLUs train several times faster than their equivalents with tanh units.

Second, the whole method is implemented on Multiple GPUs. Two GPUs just run parallel as following picture shows and the author states that this kind of architecture reduces the top-1 and top-5 error rates by 1.7% and 1.2%, respectively.



Third, the writer does not use any regularization because of the ReLUs. For a better generalization, following equation called Local Response Normalization is used:

$$b_{x,y}^{i} = a_{x,y}^{i} / (k + \alpha \sum_{j=\max(0,i-n/2)}^{\min(N-1,i+n/2)} (a_{x,y}^{j})^{2})^{\beta}$$

where constants in the equation are k=2, n=5(adjacent kernel maps), $\alpha = 10^{-4}$ and $\beta = 0.75$. The author claims that this local response normalization reduces top-1 and top-5 error rates by 1.4% and 1.2%, respectively.

In addition, two kinds of method to reduce overfitting are proposed. The first is data augmentation. The second one is dropout. The former one comes either in form of generating image translation and horizontal reflections or performing PCA on the

set of RGB pixel values throughout the ImageNet training set. And the author claims this scheme reduces top-1 error rate by over 1%.

Experimental results

The author's results on ILSVRC-2010 achieve top-1 and top-5 test error rates of 37.5% and 17.0%, which is better than the published result at that time. Also this method is applied in the ILSVRC-2012 competition and achieves a top-5 error rate of 18.2%. For the ImageNet with 10,184 categories and 8.9 million images, the method achieves top-1 and top-5 error rates 67.4% and 40.9%, respectively.

Conclusion

The author adopts convolutional neural networks without regularization or any unsupervised pre-training by two GPUs implementation, and achieves record-breaking results on large dataset.

Summaries for "DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition"

Weituo Hao 109241801
Stony Brook University
Computer Science Department
Advanced Computer Vision

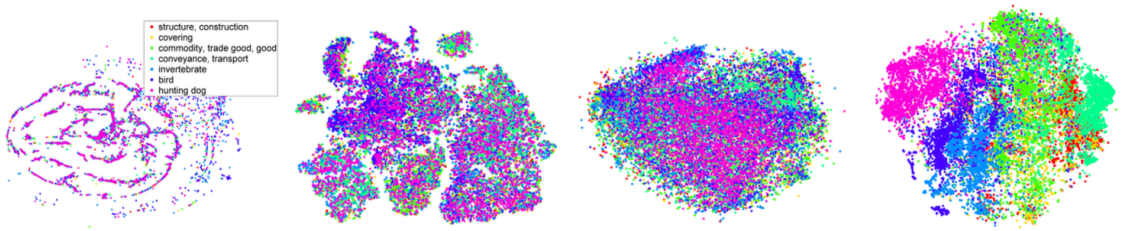
Contribution

The author claims that features extracted from the activation of a deep convolutional network trained in a fully supervised fashion can be used for multiple generic tasks including scene recognition, domain adaptation, and fine-grained recognition challenges, and the results outperform the state-of-the-art on several vision challenge tasks.

Main idea

The author implements the algorithm in the last summaries with following adaptation. First, they ignore the image's original aspect ratio and warp it to 256*256, rather than resizing and cropping to preserve the proportions. Second, no data augmentation is performed.

Then the writer compares the feature extracted by the convolutional neural networks to GIST features and LLC features by t-SNE algorithm. By visualizing the features the author concludes that the features extracted by CNNs display a clear semantic clustering in the latter but not in the former, which is exactly the common learning knowledge that the first layers learn low-level features whereas the latter layers learn high-level feature. However, other features like GIST and LLC fail to capture such characteristics. The comparison figure is shown as follows:



(a) LLC (b) GIST (c) $DeCAF_1$ (d) $DeCAF_6$

On the other hand, the author shows us a break-down of the computation time of CNNs and states that in large networks the last few fully-connected layers require the most computation time as they involve large transform matrices.

Experimental results

The most important part of this paper is the application of features extracted by CNNs to different tasks and comparison with other approaches.
First the author tests the feature on object recognition task on the Caltech-101 dataset. Using DeCAF6 and DeCAF7 as the features, SVM and logistic regression

classifiers can performs equally well. And the author claims that their performance by SVM outperforms the state-of-the-art work by 2.6%. Their on-shot learning results suggest that with sufficiently strong representations like DeCAF useful models of visual categories can often be learned from just a single positive example.

Second experiment is done about the domain adaptation on the office dataset. The author concludes that DeCAF not only provides better within category clustering, but also clusters same category instances across domains. Also by DeCAF6 and DeCAF7 features, the results outperforms dramatically the baseline SURF feature available with the office dataset as well as the deep adaptive method.

Third experiment is about the subcategory recognition on the Caltech-UCSD birds. The author uses two approaches. First one is to use DeCAF6 combining with multi-class logistic regression model. The second one is to use DeCAF6 combining with pre-trained DPM model, obtaining accuracy of 58.75% and 64.96%, respectively. Both results outperform the best record result.

Lastly, the author applies the feature in scene recognition on SUN-397 large-scale scene recognition database. Using DeCAF 7 combining SVM classifier resulting in 40.94% with an improvement of 2.9% to state-of-the-art work.

Conclusion

By the work presented in the this paper, the author concludes that features extracted from deep convolutional neural networks can also be used to different kinds of tasks and outperforms several previous state-of-the-art work, demonstrating the activation feature is suitable for many vision recognition tasks.