B-CNN Models For Fine-Grained Image Classification

Tran Anh Tuan
Hanoi University of Science and Technology
trananhtuan230102000@gmail.com

Abstract

Bilinear Convolutional Neural Networks (B-CNNs) is a pooled outer of features derived from two "ordinary" CNNs and capture localized feature interactions in a translationally invariant manner. This network is a simple and effective architecture for the Fine-Grained Image Classification Tasks. Currently, most "deep learning" models employ the softmax activation function for prediction and minimizing cross-entropy loss, a different approach that is replacing the softmax layer with a linear support vector machine. In this paper, we present a small alternative but equivalent effect with a linear SVM. Our research focuses on classifying images on the FGVC-Aircraft dataset, the experiment's results were illustrated by the accuracy and the loss graphs. Our code is available in https://github.com/tuantran23012000/BCNN-SVMs.git.

1. Introduction

Recognition with fine-grained precision Tasks like recognizing the species of an airplane model is difficult because the visual distinctions between the categories are minor and can be readily overpowered by factors like perspective or placement of the item in the image.

A frequent method for ensuring resilience against these bothersome aspects is to first locate various sections of the item and then model the appearance based on their discovered positions. One disadvantage of these techniques is that annotating components is far more difficult than gathering image labels. Furthermore, manually created pieces may be ineffective for the ultimate recognition job.

Another approach is to use a bilinear CNN [2] architecture including the what pathway is involved with the object identifier associated with the recognition, and where the pathway is involved with processing the object's spatial location relative to the viewer, then, using a k-way softmax layer for prediction. A novel approach based on a one-vs-all linear SVMs layer leads to must optimize an unconstrained convex optimization problem. With a convex function, any local minimum of it is also a global minimum, the gradients can be backpropagated to learn lower-level features of this problem.

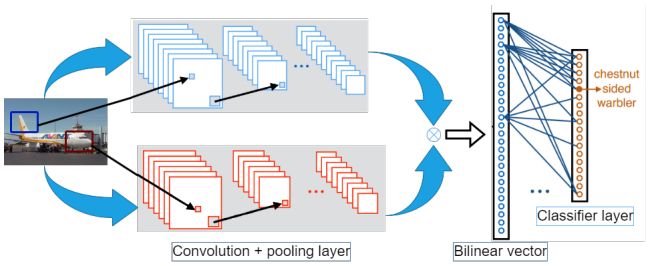

Figure 1. **A bilinear CNN model for image classification.**

2. Related work

2.1. Bilinear models for image classification

A bilinear model $\mathcal{B}$ for image classification consists of a quadruple $\mathcal{B} = (f_A, f_B, \mathcal{P}, \mathcal{C})$. Here f_A and f_B are feature functions, $\mathcal{P}$ is a pooling function and $\mathcal{C}$ is a classification function. A feature function is a mapping $f : \mathcal{L} \times \mathcal{I} \rightarrow R^{c \times D}$ that takes an image $\mathcal{I}$ and a location $\mathcal{L}$ and outputs a feature of size $c \times D$. We refer to locations generally which can include position and scale. The feature outputs are combined at each location using the matrix outer product, i.e., the bilinear feature combination of f_A and f_B at a location l is given by bilinear $(l, \mathcal{I}, f_A, f_B) = f_A(l, \mathcal{I})^T f_B(l, \mathcal{I})$. Then, this bilinear vector is passed through a classification layer such as softmax layer with cross entropy loss or linear layer with multi-class hinge loss.

We initialized with two D-nets [3] denoted by B-CNN [D, D]. Identical to the setting in FV-CNN, the input images are first resized to 448×448 and features are extracted using the two networks before bilinear combination, sum-pooling, and normalization.

2.2. Classifier training

In all our experiments, we focus on using a one-vs-all linear SVMs model on the extracted features are trained by

setting between ℓ_2 regularization and Hinge loss function.

Although the SVM model is a binary classifier, researchers work to extend it to solve multi-class classification problems. The earliest attempt is the one versus all (one versus rest) strategy that solves this problem by training more binary SVM models. In this strategy, the training process divides binary SVM problems into $\frac{c(c-1)}{2}$ subproblems in total. Instead of handing multi-class classification by solving multiple subproblems, Weston proposed to use one single objective function [Weston and Watkins, 1998],

$$\min_{W\in\mathbb{R}^{d\times c},\mathbf{b}\in\mathbb{R}^c} \frac{1}{2}\sum_{j=1}^{c}\|\mathbf{w}_j\|_2^2 + C\sum_{i=1}^{n}\sum_{j\neq y_i}\xi_i^j \quad \text{(CSVM1)}$$
$$\text{s.t. } \mathbf{w}_{y_i}^T\mathbf{x}_i + b_{y_i} \geq \mathbf{w}_j^T\mathbf{x}_i + b_j + 2 - \xi_i^j$$
$$\xi_i^j \geq 0, \forall i, j \in \{1,\ldots,c\}\backslash y_i,$$

where, c is number of classes, ξ_n are slack variables which penalizes data points which violate the margin requirements. The corresponding unconstrained optimization problem is the following:

$$\min_{W,\mathbf{b}} \frac{1}{2}\sum_{j=1}^{c}\|\mathbf{w}_j\|_2^2 + C\sum_{i=1}^{n}\sum_{j\neq y_i}\max\Big(0, 2 - \mathbf{w}_{y_i}^T\mathbf{x}_i - b_{y_i} + \quad \text{(CXSVM1)}$$
$$\mathbf{w}_j^T\mathbf{x}_i + b_j\Big).$$

3. Method

Proposition 3.1 *([1]). If $f_i(x)$, where i through any set, are convex functions in D, the set of all points x of D at which $\sup_i f_i(x)$ is finite is convex and $\sup_i f_i(x)$ is a convex function in this set.*

From Proposition 3.1, we propose to build an equivalent problem with (CSVM1) as follows,

$$\min_{W,\mathbf{b}}\{\frac{1}{2}\max_{j\in\overline{1,c}}\left(\|w_j\|_2^2\right) + C\sum_{i=1}^{n}\sum_{j\neq y_i}\xi_i^j\} \quad \text{(CSVM2)}$$
$$\text{s.t. } \mathbf{w}_{y_i}^T\mathbf{x}_i + b_{y_i} \geq \mathbf{w}_j^T\mathbf{x}_i + b_j + 2 - \xi_i^j$$
$$\xi_i^j \geq 0, \forall i, j \in \{1,\ldots,c\}\backslash y_i.$$

The corresponding unconstrained optimization problem is the following:

$$\min_{W,\mathbf{b}} C\sum_{i=1}^{n}\sum_{j\neq y_i}\max\Big(0, 2 - \mathbf{w}_{y_i}^T\mathbf{x}_i - b_{y_i} + \mathbf{w}_j^T\mathbf{x}_i + b_j\Big) + \quad \text{(CXSVM2)}$$
$$\frac{1}{2}\max_{j\in\overline{1,c}}\left(\|\mathbf{w}_j\|_2^2\right).$$

In reality, $\frac{1}{2}\max_{j\in\overline{1,c}}\left(\|\mathbf{w}_j\|_2^2\right)$ and $\frac{1}{2}\sum_{j=1}^{c}\|\mathbf{w}_j\|_2^2$ are regularization terms that help the model avoid overfitting. In circumstances, the weights in a class are becoming huge, regularization deal with encouraging the weights to be small. Hence, we can completely utilize regularization terms such as,

$$\min_{W,\mathbf{b}} \frac{1}{2}\sum_{j=1}^{c}\|\mathbf{w}_j\|_1 + C\sum_{i=1}^{n}\sum_{j\neq y_i}\max\Big(0, 2 - \mathbf{w}_{y_i}^T\mathbf{x}_i - b_{y_i} + \quad \text{(CXSVM3)}$$
$$\mathbf{w}_j^T\mathbf{x}_i + b_j\Big),$$

or,

$$\min_{W,\mathbf{b}} \frac{1}{2}\max_{j\in\overline{1,c}}\left(\|\mathbf{w}_j\|_1\right) + C\sum_{i=1}^{n}\sum_{j\neq y_i}\max\Big(0, 2 - \mathbf{w}_{y_i}^T\mathbf{x}_i - b_{y_i} + \quad \text{(CXSVM4)}$$
$$\mathbf{w}_j^T\mathbf{x}_i + b_j\Big).$$

4. Experiment

Table 1. **Classification results**. Accuracy on the FGVC-Aircraft testing dataset. We trained on a NVIDIA GeForce GTX 1660Ti with batch size 32, image size 448×448.

Method	FGVC-Aircraft
CNN [D]	54.85
B-CNN [D, D] + CXSVM1	67.11
B-CNN [D, D] + CXSVM2	67.11
B-CNN [D, D] + CXSVM3	67.07
B-CNN [D, D] + CXSVM4	67.07

5. Conclusion

In this paper, we follow the idea of replacing equivalent regularization terms of the SVMs classifier model, based on these ideas, we can expand regularization into expanded convex function classes.

References

[1] Werner Fenchel and Donald W Blackett. *Convex cones, sets, and functions*. Princeton University, Department of Mathematics, Logistics Research Project, 1953. 2

[2] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear cnn models for fine-grained visual recognition. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1449–1457, 2015. 1

[3] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv 1409.1556*, 09 2014. 1