BAGNETS REPRODUCIBILITY CHALLENGE

Anonymous authors
Paper under double-blind review

ABSTRACT

This report reimplements the ICLR paper *Approximating CNNS with Bag-of-Local-Features Models Works Surprisingly Well on ImageNet*. The proposed methodology and experiments are reproduced. The conclusion of the paper that the improvements of DNNs were achieved by fine-tuning instead of using high qualitative decision strategies was proved. This report also proves that larger receptive fields do not always have positive effects on classification. And BagNets are relatively insensitive to scramble images.

1 INTRODUCTION

A problem that has always been plaguing people is to understand DNN's complex decision-making mechanism.Brendel & Bethge (2019) tried to explore the decision mechanism of deep neural networks (DNNs) with an easy-to-analyse model (BagNets) and concluded that the improvements of DNNs are mostly achieved by better fine-tuning rather than by qualitatively different decision strategies.

Bag-of-Local-Features models (BagNets) classify an image based on the occurrences of small local image features without taking into account their spatial ordering. It limits the receptive fields of the convolution layers, so each element of the convolved tensor before the final adaptive average pooling layer holds the local feature of a small image patch. Then BagNets combine those local features linearly by feeding this convolved tensor into an adaptive average pooling layer then a fully-connected layer, and produce the logit tensor which is for the final classification. The mechanism of BagNets is almost the same as ResNets', the only difference is the size of receptive fields of BagNets is small.

2 MATERIALS AND METHODS

2.1 DATASET PREPARATION AND FEATURE EXTRACTION

Considering training efficacy and performance, we altered datasets from ImageNet-1K to the scene recognition images utilised in Computer Vision Assignment 3. The 15-scene-class database contains a training and a test sets (Lazebnik et al., 2006). We randomly split 1200/300 as training/validation sets from the original training set, and randomly select 1200 images from the test set for final performance evaluation. Before feeding into CNNs, all images are resized to 224×224 and normalised.

Kernel Size	Stride	Dilation	Padding	Input Size	Output Size	Receptive Field
3	1	1	0	224	222	3
3	2	1	0	222	110	5
3	2	1	0	110	54	9
3	2	1	0	54	26	17
3	1	1	0	26	24	33

Figure 1: A brief Receptive Fields demonstration of BagNets

2.2 BAGNETS REPRODUCIBILITY

The architecture of BagNets is based on ResNets. Its key innovation is the alteration from some 3×3 kernels to 1×1 kernels and the elimination of max-pooling layers. By controlling the number of 3×3 kernels and their strides, the receptive field of each element of the convolved tensor is limited to 33 before the final adaptive average pooling layer in BagNet33 (Figure 1). Figure 1 only lists 3×3 kernels which increase receptive fields. By replacing the last one or two 3×3 kernel(s) to 1×1 kernel(s), the architecture will have 17 or 9 final receptive fields, also known as BagNet17 or BagNet9, respectively.

Contrary to most people's intuition, at training stage, full-size images are fed into BagNets but not the cropped image patches. This means that the mechanism of BagNets is almost the same as other convolutional neural networks', the only difference is the size of receptive fields of BagNets is small. What BagNets try to prove is that the linear combination of all local features of an image is the key of classification, i.e. the spatial relationship and ordering of local features do not significantly affect the final decision. Heatmaps which could provide evidence for decision-making according to the author are also reproduced by us.

When generating heatmaps, a full-size image are cropped into small patches by a sliding window with stride 1. For example, the input image is cropped as a series of 33×33 subimages in BagNet33 and 17×17 in BagNet17. Through the BagNet, each cropped image patch generates a 15-dim logit. Thus, a 224×224 image produces a 50176×15 tensor where the second dimension (*i.e. 15, also called as class layers*) represents classes. The logits of corresponding class (*i.e. a* 50176×1 *tensor*) are extracted and resized to an image-size heatmap. The heatmap can visualise output logits of the input image, it reveals which local features are emphasised through a well-trained model.

3 RESULTS AND DISCUSSIONS

3.1 PERFORMANCE COMPARISON

Table 1 shows their validation and test sets accuracy of their well-trained model. The comparison of BagNet9, BagNet17 and BagNet33 shows that the accuracy is enhanced from 0.840 to 0.925 as the increase of receptive fields, which indicates that the number of features in receptive fields do influence the performance of classification. The improvement of accuracy conforms to the tendency in the original paper (Brendel & Bethge, 2019). Surprisingly, BagNet33 outperforms ResNet50 whose receptive field is 483 pixels covering all image area. Therefore, other variables, such as maxpooling layers, padding/stride methods, and network depth, may have significant negative effects when training neural networks. More variable controlling experiments would be designed and implemented to reveal the effects of reducing the performance of ResNet50. In general, Cbam-resnet50 achieves the best performance (0.934) in the test set. The new CBAM module provides significant improvement to ResNets. Different from naive ResNet50, it introduces attention maps along the channel and spatial dimensions. Besides, its accuracy is higher than that of BagNet33 by around 0.011.

Table 1: The details of CNN performance.

CNN Architectures	Validation Accuracy	Test Accuracy	Test(Scrambling) Accuracy
Resnet50	0.913	0.907	0.50
Densenet121	0.933	0.918	0.57
Densenet169	0.953	0.922	0.58
Cbam-resnet50	0.953	0.934	0.73
BagNets9	0.853	0.840	-
BagNets17	0.913	0.905	-
BagNets33	0.947	0.925	0.90

3.2 THE CONTRIBUTION OF LOCAL FEATURES

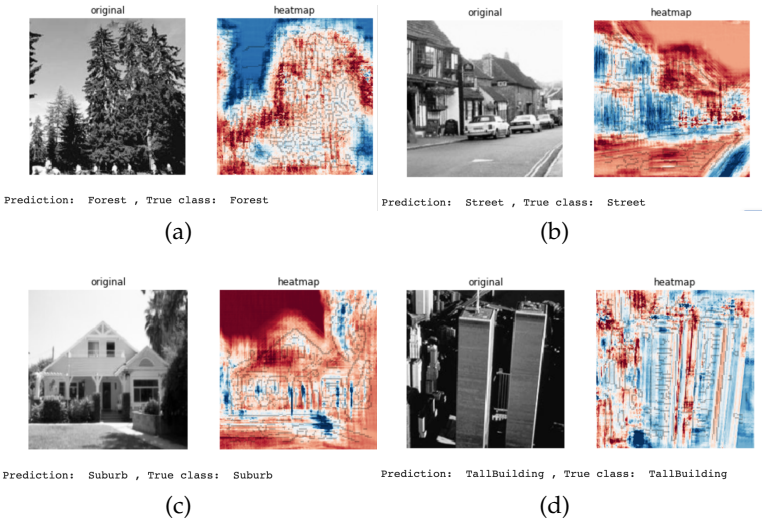

Figure 2: Four example heatmaps generated from BagNet33 of (a) Forest class; (b) Street class; (c) Suburb class; (d) Tallbuilding class

Figure 2-a to d illustrate four example figures (TallBuilding, Forest, Street and Suburb) and their heatmaps through the BagNet33. The examples indicate that the distinct features of different types are learned for image differentiation, and some other common features may also assist in classification. For example, Figure 2-a to call contain the sky, but the model most concentrate on the feature of leaves, branches, and trunks. The sky in Figure 2-b has positive effects on street type classification but the most significant features are roofs and road surface. The Suburb sub-class achieve 1.0 precision and recall values. In its heatmap, the sky parts have the highest logits and the substantial contribution, and the building structure is also considered as classification evidence. To some distinctly different image types, such as TallBuilding (Figure 2-b), most vertical borders of the building are learned.

Apart from some experiments described in the original paper (Brendel & Bethge, 2019), we compared the logit heatmaps of same images generated from different BagNets (Figure 3 -a to f). Figure 3 -a to c and d to f are heatmaps generated from two images by BagNet9, 17, and 33 respectively. The heatmaps illustrate that the size of receptive fields affects the distribution of logits and which local features would contribute more on classification. In same *Suburb* class, the sky area generates lowest logits in BagNet9 (Figure 3-a and d)whereas it was significantly concentrated in the results of BagNet33 (Figure 3-c and f). In BagNet17 heatmaps (Figure 3-b and e), the sky area generates medium-level logit values, and some features of roofs and yards contribute more in classification.

3.3 THE PERFORMANCE UNDER IMAGE SCRAMBLING

To investigate how similar the decision-making of BagNets to high-performance DNNs like VGG-16, ResNet-50, DenseNet-121 and DenseNet-169, the authors of the paper compared the performance of the networks with scramble images. The conclusion they reached through the results is that the high-performance DNNs do not rely on global shape integration for perceptual discrimination but rather on statistical regularities in the histogram of local image features.

In order to verify their conclusions, we divided the image into 16 small pieces and randomly scramble together and test networks according to their approach. The results we got showed similar conclusions that they proposed. For the human, it is hard to recognize what scrambling images are, but BagNet and other networks just got little affected. To be more precisely, BagNet showed 93% accuracy on clean versus 90% on scramble image and

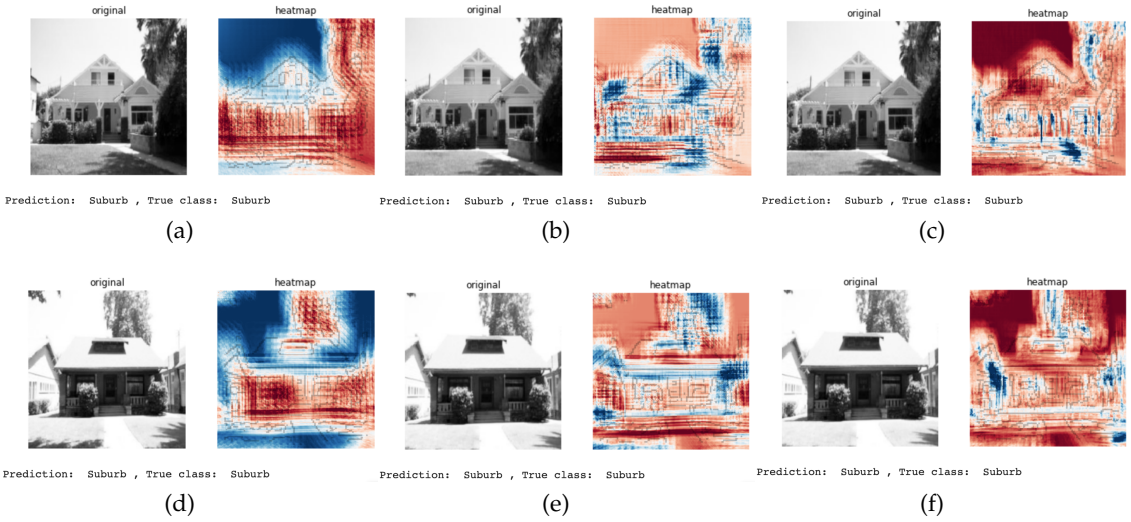

Figure 3: Three example Suburb heatmaps generated by (a) BagNet9; (b) BagNet17; (c) BagNet9; Another three example Suburb heatmaps generated by (d) BagNet9; (e) BagNet17; (f) BagNet33

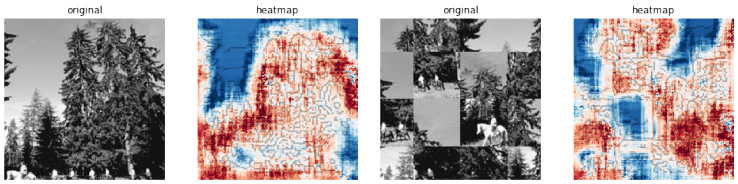

Figure 4: The heatmaps of an example image and a scrambling image : (a)Origin image ; (b)Scrambling image

Cbam-resnet50 showed 93% accuracy on clean versus 73% on scrambling image. What's more, the heatmap of scrambling image also helps us to support the conclusion(See Fig 4). Even if the segmentation of the image destroys some features, the intact local features still can help the network make a correct judgment.

4 CONCLUSIONS AND FUTURE WORKS

Although the maximum receptive fields are restricted in BegNets, the last average pooling layer and fully-connected layer integrate the logits throughout input images in the gradient descent step. The BagNets' concept which is classifying images only by their local features should be suspected, and the training process should also be improved.

REFERENCES

Wieland Brendel and Matthias Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. *arXiv preprint arXiv:1904.00760*, 2019.

Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pp. 2169–2178. IEEE, 2006.