# Visual Attention Mechanism on BAGNETS

Jeet Sen Sarma [1]

23rd August 2022

[1]Dario Zanca, Thomas Robert Altstidl

# Contents

**Abstract**

Deep Neural Networks(DNNs) have excelled in many perceptual tasks but it is notoriously difficult to understand their decision making process. In contrast, the variant Bag-Of-Local features or BagNets model classifies an image based on the presence of small local image features, independent of their spatial ordering. By doing so it achieves a Top5 accuracy of 80.5% on ImageNet Dataset, which is similar to AlexNet. Another important feature in perceptual tasks is attention. In this study we apply various Human attention models to BagNets replacing the Simple Average model used by the author. Experimental results on the ImageNet Dataset, show that the various attention models performed worse than the Simple Average model. This implies that decreasing the number of patches of an Image used for Classification using Attention Mechanism doesn't help in improving accuracy in this case.

**Index terms: DNN, BagNets, Attention, ImageNet**

# I Introduction

## A Bag-Of-Local Features(BagNets)

The BagNets model is based on the bag-of-feature (BoF) models [5] which along with extensions such as VLAD encoding or Fisher Vectors were the most successful approaches to large-scale object detection before the advent of deep learning (up to 75% top-5 on ImageNet) and were able to classify images based on the count of a set of local features, and not taking into account their spatial relationships. Unlike DNNs, which non-linearly integrate information across the whole image, basic BoF models generally apply a spatial average of the patch-wise features.

The model downsamples image to an array of small patches of size 9×9, 17×17 or 33×33 depending on the architecture. Features are extracted from these small image patches which are each fed into a linear classifier yielding one logit heatmap per class. These heatmaps are averaged across space and passed through a linear classifier to get the final class probabilities. The classifier and the spatial aggregation being linear allows interchangeability and helps to pinpoint how evidence from image patches is integrated into one image-level decision.

## B Attention Mechanism

Attention Model has now become an important concept in neural networks and Artificial Intelligence(AI) community and has been researched within diverse application domains including Natural Language Processing(NLP) [7], Speech [2] and Computer Vision(CV) [4]

The intuition underlying attention is best explained in terms of human biological systems. In cognitive science, due to bottlenecks in information processing, only a fraction of all visible information is perceived by humans. Inspired by this, scientists have tried to simulate the Attention mechanism of humans, so as to model the distribution of human attention when observing images as well as videos and expand its applications [3]. It has been shown that attention mechanisms can improve the model performance, and are also compatible with perceptual mechanisms in the human brain and eyes.

In Computer Vision, one typically finds the use of an Image mask for combining deep learning and visual attention based mechanisms. The Image mask is a pixel-wise mask applied to the image with different weights or

values for identifying key features in the Images. By training, deep neural network can learn the areas where attention needs to be applied in each new image, thereby forming an attention mechanism.

# II Our Approach

We describe the main elements of BagNets, before presenting our approach. The BagNets model is similar to the ResNet-50[REf] architecture except for a few changes in strides and replacing the 3×3 convolution with 1× convolution. It has 3 variants which downsamples the input image into patches of size 9×9, 17×17 or 33×33 , which are then combined linearly (i.e. a simple average) and passed through a linear classifier to get the class probabilities of the images. This aggregation of small patches removes their spatial correlation and the model classifies the image based on the number of occurrences of a particular class.

For our approach, we want to find out if the model's attention on a subset of small image patches is enough classify the Image correctly. In this regard, we selected the pre-trained BagNet-17 Model which is the intermediate between the three variants and thus a good sample for our study. Next we prepare the ImageNet Dataset to be of size 3×224×224 which is the required input size for the model. This was followed by the selection of a small subset of Attention mechanisms and generating weight masks corresponding to them, having the shape of the image patches. The simple average function is replaced by the complex weighting schemes generated by the attention mechanisms and passed to the linear classifier to generate the class probabilities. Although our weight masks are not equally distributed but they combine with the image patches linearly, thus allowing insights into the model's decision procedure.

# III Weighting Schemes

Here we discuss the attention mechanisms that we used. In each case, we use an Image mask of different weights corresponding to each of the local patches. The corresponding plots for the weight matrices can be found in Figure 1.

## *A Global Average*

First we try out global average over all the patches. Here the mask has equal weights for all patches. This gives similar result to original Bagnets and thus is used as a baseline.

## *B Gaussian Weight Matrix*

Next we use a Gaussian Weight Matrix. Here we weigh the patches with a weight matrix where the values are taken from a Gaussian distribution. We vary the std deviation of the Gaussian distribution and apply it over the patches before taking a **Global Average**. This results in mimicking the center focus attention bias present in humans.

## *C Unit Square Weight Matrix*

This weight matrix is similar to **Gaussian** one except this has a hard boundary and all weights are either 0 or 1. This acts as a binary map in the sense that it either keeps or rejects a local patch. We vary the length of the square and apply the weights to the patches.

## *D Random Weighting*

Here the weight matrix has 1's and 0's **Randomly** placed. This also acts as a binary map and helps us to check that exact accuracy doesn't depend on any particular patch. The number of 1's placed is varied to generate distinctive random weight matrices.

## *E Saliency Map*

Here the weight matrix is computed using the Saliency Map, which represents **Saliency** at every position of the Image, presented by Itti, Koch and Niebur [1]. They use Multiscale Image features to capture saliency and combine them into one topographical saliency map which is used by a neural network to select positions in order of decreasing saliency. We use the Saliency map as it is generated to weigh our patches accordingly.

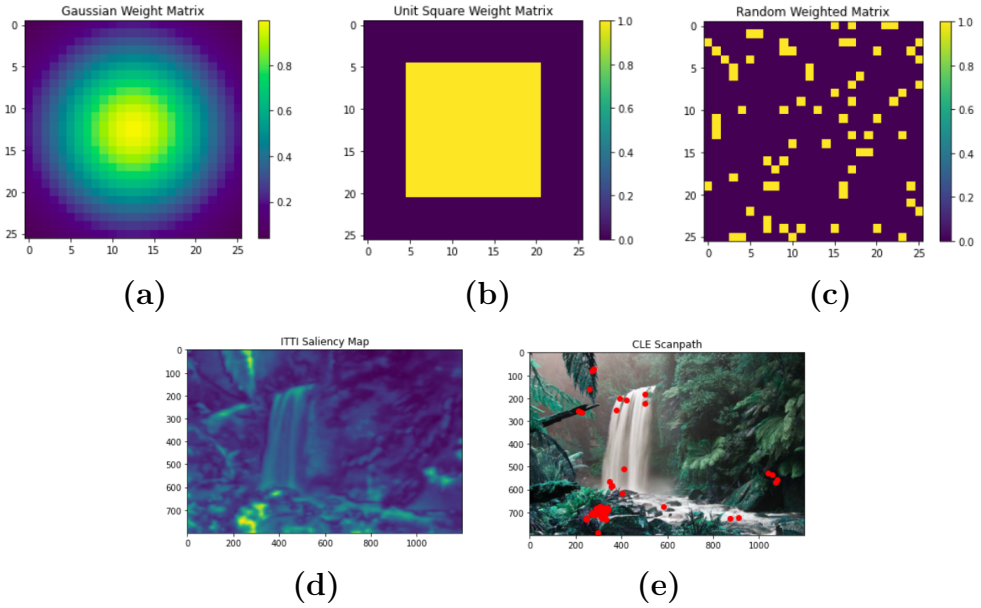


Figure 1: **(a)** Gaussian Weights **(b)** Unit Square Weights **(c)** Random Weighting **(d)** Saliency Map **(e)** CLE Scanpath

### *F CLE Scanpath*

In this case, we generate a scanpath for the image using Constrained Levy Exploration [6], which estimates the probability of the eye movements of a human subject to presence of significant details and motion in Image or Video which are then integrated over time to generate Positions of Interest. We then select patches for Inference which coincide with the points of interest. The number of Points of Interest taken into account was varied to find the optimum number for inference.

## IV Results

The results obtained from our experiments can be seen in Figure 2. We calculated the Top-1 and Top-5 accuracy of the Bagnets model using different Attention mechanisms on the ImageNet Dataset. The Gaussian weights attention model, seen in Fig:2(a), peaks in accuracy at around Standard Deviation 7 but shows no significant improvement over the baseline Top-5

accuracy of 80.5%. The plot for the Center-Focused weights or the Unit Square weights, in Fig:2(b), show that the accuracy keeps increasing till the length of square increases, and reaches the baseline when all patches are considered. Results obtained from using Random weights, in Fig:2(c), show that the accuracy reaches near the Baseline at approximately half the total number of patches but fails to show any considerable improvement in the end. Using the Saliency Map as weight matrix, we obtain a Top-5 accuracy of 78.512% and Top-1 accuracy of 55.314%, thus shows no considerable improvement. Lastly, the plot obtained using the CLE scanpath in Fig:2(d), shows that it performs the worse among all other cases and fails to reach even the Baseline of the original model.

In conclusion, we were unable to achieve any improvement in accuracy over the original model using the stated Attention Mechanisms.

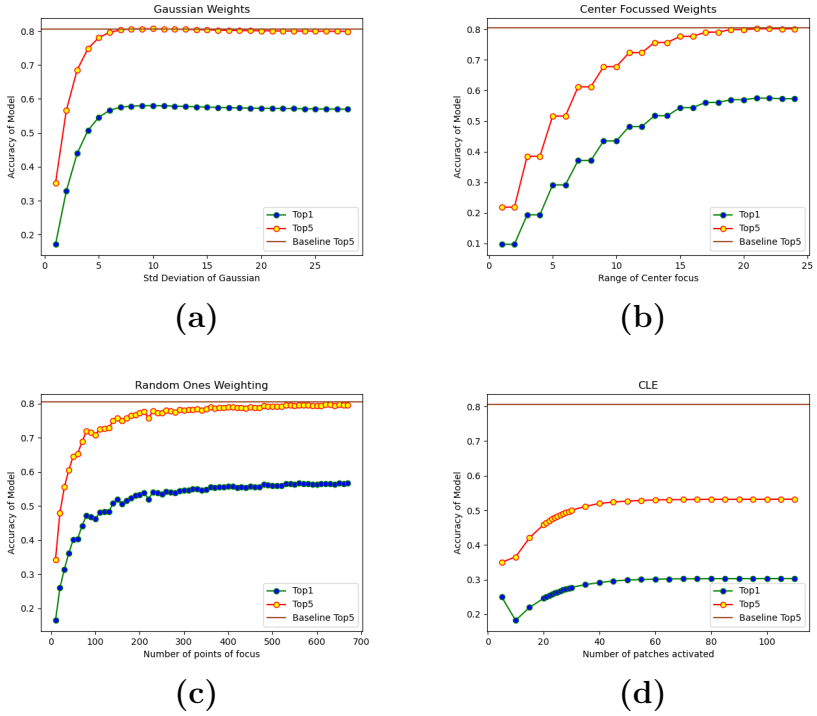


Figure 2: The Top1 and Top5 accuracy are calculated over all ImageNet Images using, **(a)** Gaussian Weights **(b)** Unit Square weights **(c)** Random Weights **(d)** CLE scanpath

# V Discussion & Future Work

From the results it is evident that traditional attention mechanisms don't have any significant improvement on the classification accuracy of BagNets model and that Global Averaging performs the best. This might be attributed to the fact that all the weighting schemes chosen by us, selects a subset of the already small number of patches generated by the BagNets model, before passing them to the linear classifier. This may cause a loss of information which prevents any improvement in performance.

We have only used traditional Attention mechanisms. Alternatively, Deep Neural Network based attention mechanisms can be used. Due to shortage of time we could conduct our experiments on only one of the variants of BagNets, namely BagNets17. The above methods can also be applied on other variants for verifying possible improvements.

# References


[1] L. Itti, C. Koch, and E. Niebur. "A model of saliency-based visual attention for rapid scene analysis". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20.11 (1998), pp. 1254–1259. DOI: 10.1109/34.730558.

[2] Kyunghyun Cho, Aaron Courville, and Yoshua Bengio. "Describing Multimedia Content Using Attention-Based Encoder-Decoder Networks". In: *IEEE Transactions on Multimedia* 17.11 (2015), pp. 1875–1886. DOI: 10.1109/TMM.2015.2477044.

[3] Kelvin Xu et al. *Show, Attend and Tell: Neural Image Caption Generation with Visual Attention*. 2015. DOI: 10.48550/ARXIV.1502.03044. URL: https://arxiv.org/abs/1502.03044.

[4] Feng Wang and David Tax. "Survey on the attention based RNN model and its applications in computer vision". In: (Jan. 2016).

[5] Wieland Brendel and Matthias Bethge. "Approximating cnns with bag-of-local-features models works surprisingly well on imagenet". In: *arXiv preprint arXiv:1904.00760* (2019).

[6] Dario Zanca, Stefano Melacci, and Marco Gori. "Gravitational Laws of Focus of Attention". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42.12 (2020), pp. 2983–2995. DOI: 10.1109/TPAMI.2019.2920636.

[7] Andrea Galassi, Marco Lippi, and Paolo Torroni. "Attention in Natural Language Processing". In: *IEEE Transactions on Neural Networks and Learning Systems* 32.10 (Oct. 2021), pp. 4291–4308. DOI: 10.1109/tnnls.2020.3019893. URL: https://doi.org/10.1109%5C%2Ftnnls.2020.3019893.