Measuring the Spectrum of Deepnet Hessians

Abstract

We apply state-of-the-art tools in modern high-dimensional numerical linear algebra to approximate efficiently the spectrum of the Hessian of modern deepnets, with tens of millions of parameters, trained on real data. We decompose the Hessian into different components and study the dynamics with training and sample size of each term individually.

1 Introduction

We consider a C-class classification problem, whereby we are given n training examples in C classes, $\cup_{c=1}^{C}\{x_{i,c}\}_{i=1}^{n}$, and their corresponding labels and the goal is to classify future data. State-of-the-art methods tackle this problem using deepnets. These are trained using stochastic gradient descent by minimizing the empirical cross-entropy loss,

$$\mathcal{L}(\theta) = \text{Ave}_{i,c}\{\ell(f(x_{i,c};\theta), y_c)\}. \tag{1}$$

Here we denoted by $f(x_{i,c};\theta) \in \mathbb{R}^C$ the output of the classifier and by $\theta \in \mathbb{R}^p$ the concatenation of all the parameters in the network into a single vector. Our goal is to investigate the Hessian of the loss averaged over the training (or testing) data,

$$\text{Hess}(\theta) = \text{Ave}_{i,c}\left\{\frac{\partial^2 \ell(f(x_{i,c};\theta), y_c)}{\partial \theta^2}\right\}. \tag{2}$$

Using the Gauss-Newton decomposition, the Hessian can be written as a summation of two components:

$$\text{Hess}(\theta) = \underbrace{\text{Ave}_{i,c}\left\{\sum_{c'=1}^{C} \left.\frac{\partial \ell(z, y_c)}{\partial z_{c'}}\right|_{f(x_{i,c};\theta)} \frac{\partial^2 f_{c'}(x_{i,c};\theta)}{\partial \theta^2}\right\}}_{H} \tag{3}$$

$$+ \underbrace{\text{Ave}_{i,c}\left\{\frac{\partial f(x_{i,c};\theta)}{\partial \theta}^T \left.\frac{\partial^2 \ell(z, y_c)}{\partial z^2}\right|_{f(x_{i,c};\theta)} \frac{\partial f(x_{i,c};\theta)}{\partial \theta}\right\}}_{G}. \tag{4}$$

Moreover, G can be further decomposed, as explained in the following section.

2 Three-level hierarchical decomposition

Define the p-dimensional vectors

$$g_{i,c,c'}{}^T = (y_{c'} - p(x_{i,c};\theta))^T \frac{\partial f(x_{i,c};\theta)}{\partial \theta}, \tag{5}$$

where $p(x_{i,c};\theta) \in \mathbb{R}^C$ are the softmax probabilities of $x_{i,c}$. Note that for $c = c'$,

$$g_{i,c,c}{}^T = (y_c - p(x_{i,c};\theta))^T \frac{\partial f(x_{i,c};\theta)}{\partial \theta} = \frac{\partial \ell(z, y_c)}{\partial z}^T \bigg|_{z_{i,c}} \frac{\partial f(x_{i,c};\theta)}{\partial \theta} = \frac{\partial \ell(f(x_{i,c};\theta), y_c)}{\partial \theta} \tag{6}$$

and so $g_{i,c,c}$ is simply the gradient of the i'th training example in the c'th class. Following the same logic, $g_{i,c,c'}$ is the gradient of the i'th training example in the c'th class, if it belonged to class c' instead. Denote $p_{i,c,c'}$ to be the c'-th entry of $p(x_{i,c};\theta)$ and define the following quantities

$$\begin{aligned}
g_{c,c'} &= \frac{1}{p_{c,c'}} \sum_i p_{i,c,c'} g_{i,c,c'} & g_c &= \frac{1}{p_c} \sum_{c' \neq c} p_{c,c'} g_{c,c'} \\
\Sigma_{c,c'} &= \frac{1}{p_{c,c'}} \sum_i p_{i,c,c'} (g_{i,c,c'} - g_{c,c'})(g_{i,c,c'} - g_{c,c'})^T & \Sigma_c &= \frac{1}{p_c} \sum_{c' \neq c} p_{c,c'} (g_{c,c'} - g_{c'})(g_{c,c'} - g_{c'})^T \\
p_{c,c'} &= \sum_i p_{i,c,c'} & p_c &= \sum_{c' \neq c} p_{c,c'}
\end{aligned} \tag{7}$$

The left equations cluster gradients for a fixed pair of c, c', whereas the right equations cluster gradients with a fixed c. Leveraging this definitions, we prove that G can be decomposed as follows:

$$G = \underbrace{\sum_c \frac{p_c}{nC} g_c g_c^T}_{A_1} + \underbrace{\sum_c \frac{p_{c,c}}{nC} g_{c,c} g_{c,c}^T}_{A_2} + \underbrace{\sum_c \frac{p_c}{nC} \Sigma_c}_{B_1} + \underbrace{\sum_c \underbrace{\sum_{c'} \frac{p_{c,c'}}{nC} \Sigma_{c,c'}}_{B_{2,c}}}_{B_2} .$$

3 Deliverables

Software for analyzing the spectra of the deepnet Hessians. We release software implementing state-of-the-art tools in numerical linear algebra, allowing one to approximate efficiently the spectrum (or log spectrum) of the Hessian and its constituent components of modern deepnets such as VGG and ResNet.

Confirmation of bulk-and-outliers (Figure 1). We confirm previous reports of bulk-and-outliers structure in the Hessian of toy models and synthetic data [1, 2], this time at the full scale of modern state-of-the-art nets and on real natural images.

Attribution of outliers to G and bulk to H. We observe that the spectrum of H does not contain outliers; thereby we can solidly attribute the outliers of the Hessian to G. Moreover, we observe that most of the energy in the bulk of the spectrum can be attributed to H.

Distribution of H. We document the spectrum of $\log(H)$ and find that it **does not** follow a semicircle distribution, as previously suggested, but rather a power law trend.

Attribution of the upper tail of the bulk. We document the dynamics of the train Hessian, G and H as function of training and sample size. We find that a transition occurs, whereby for low sample size and low epochs the upper tail of the Hessian bulk can be attributed to G, while for higher sample size and higher epochs it can be attributed to H.

Attribution of outliers in G to A_1. We show the outliers of the Hessian, which we attribute to G, are in fact due to G being a second moment matrix. Moreover, the three-level hierarchical structure unveils that the unsubtracted means correspond to the matrix A_1.

Corroboration of three-level hierarchical structure in G (Figure 2). We provide evidence for the three-level hierarchical structure in the spectrum of G in the form of t-SNE plots of g_c and $g_{c,c'}$. Moreover, as the decomposition predicts, there exist two bulks in the spectrum corresponding to B_1 and B_2.

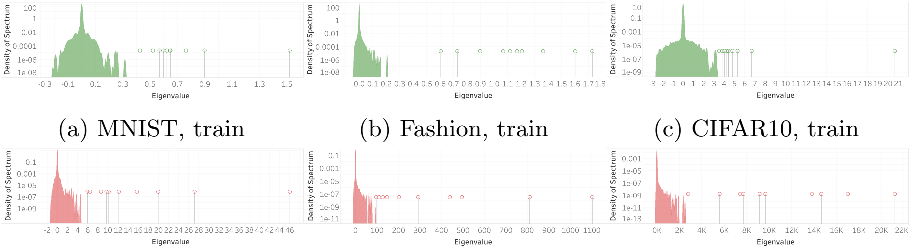
Dynamics of structure in G with sample size and training. We document the dynamics of the hierarchical structure in G, across different sample sizes and epochs. We find that fixing sample size and increasing the epoch causes the bulks B_1 and B_2 to separate. While fixing the epoch and increasing sample size causes the two bulks to draw closer. We also observe that fixing sample size and increasing the epoch or fixing the epoch and increasing sample size causes the outliers due to A_1 to separate from the bulk due to B_2.

Massive experiments. We confirm our results across different datasets (MNIST, FashionMNIST, CIFAR10, CIFAR100, ImageNet), networks (VGG11, ResNet18, DenseNet40), sample sizes, epochs and other hyperparameters.

References

[1] L. Sagun, L. Bottou, and Y. LeCun, “Eigenvalues of the hessian in deep learning: Singularity and beyond,” *arXiv preprint arXiv:1611.07476*, 2016.

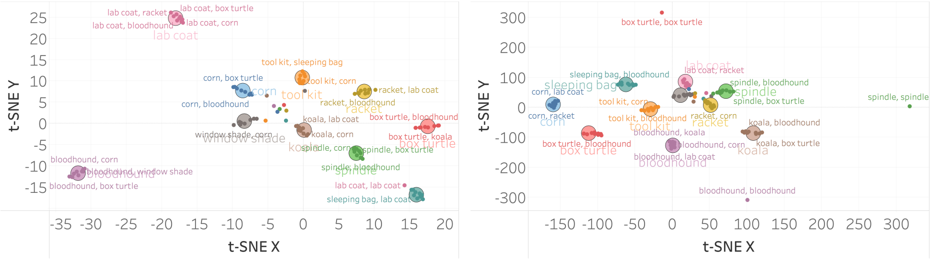
[2] L. Sagun, U. Evci, V. U. Guney, Y. Dauphin, and L. Bottou, “Empirical analysis of the hessian of over-parametrized neural networks,” *arXiv preprint arXiv:1706.04454*, 2017.



(a) MNIST, train (b) Fashion, train (c) CIFAR10, train

(d) MNIST, test (e) Fashion, test (f) CIFAR10, test

Figure 1: *Spectrum of the Hessian for VGG11 trained on various datasets.* Each column of panels documents a famous dataset in deep learning. The panels in the top row correspond to the train Hessian, while those in the bottom row to the test Hessian. We observe a clear bulk-and-outliers structure. Arguably the number of outliers is equal to the number of classes.



(a) VGG16. (b) ResNet50.

Figure 2: *t-SNE visualization of the hierarchical structure in ImageNet.* Each panel depicts the two-dimensional t-SNE embedding of g_c and $g_{c,c'}$ for a different architecture. All circles are colored according to the class c. The g_c are marked with large circles and have a label written in large font attached to them. The $g_{c,c'}$ are marked with small circles and a subset of them also have a label, written in smaller font, attached to them. The label is a concatenation of the two class names corresponding to c and c'. This plot asserts the three level hierarchy. At level one we have the cluster centers $\{g_c\}_c$. At level two, next to each cluster center g_c, we find cluster members $\{g_{c,c'}\}_{c' \neq c}$. For visualization purposes, we subset randomly ten classes.