A GRAPH-BASED APPROACH TO ESTIMATE MUTUAL INFORMATION

Meenakshi Khosla
mk2299@cornell.edu

Evan M. Yu
emy24@cornell.edu

ABSTRACT

Recently, there is an increasing need to understanding the internal representation of deep neural networks (DNNs) in order to shed light into their learning dynamics. One way to accomplish this task is to measure the mutual information (MI) between the input data and the latent space. It has been shown that both MI in stochastic DNNs and binned MI in deterministic DNNs tracks the amount of clustering in the latent space. As a result we propose to use density-based spatial clustering of applications with noise (DB-SCAN) to directly study the geometry of the latent space. By creating a graphical representation of the space, we remove the dimensional dependency when estimating the MI. The computational complexity of DB-SCAN is at most $\mathcal{O}(n^2)$. We demonstrated our method in Shwartz-Ziv & Tishby (2017) and MNIST dataset. Our estimates of MI closely follow the trajectory of binned MI while taking significantly less computational time. We hope to facilitate research techniques in estimating MI on DNNs with real-world datasets and probe open questions about the relationship between compression/clustering and generalization.

1 INTRODUCTION

Recently, there has been an increasing interest in measuring the mutual information $I(X;T_l)$ between the input data X and the hidden representation T_l at a layer l of a deep neural networks (DNNs). The sparked interest was partly due to the introduction of information bottleneck framework to understand DNNs (Shwartz-Ziv & Tishby (2017)). The authors of this work proposed that the goal of DNNs is to optimize the information bottleneck (IB) or in other words, an optimal trade-off between compression and prediction of label Y, as measured by $I(X;T_l)$ and $I(X;Y)$ respectively. Although most claims by Shwartz-Ziv and Tishby have been challenged and refuted (Saxe et al. (2018); Goldfeld et al. (2019b)), their idea of using mutual information to study DNN remains.

For most practical problems, mutual information does not have an analytical solution. In addition, its sample complexity scales exponentially with the dimension of the variables (Goldfeld et al. (2019a)). Most importantly, for a DNN that has a deterministic and injective mapping, measuring $I(X;T_l)$ is meaningless. In this scenario, it can be mathematically shown that $I(X;T_l)$ is constant if X is discrete. On the other hand, $I(X;T_l)$ is infinite when X is continuous. Both (Shwartz-Ziv & Tishby (2017); Saxe et al. (2018)), used a binning approach on the representation of the DNN $I(X;\text{Bin}(T_l))$ to overcome this hurdle. However, it is important to note that MI measures computed with binned representations induce an artificial non-injective mapping that doesn't reflect the true operation of these DNNs. Further, with this non-injective mapping, this quantity now depends on the system parameters. So what they are measuring in fact is a system-dependent approximation for a system-independent quantity.

Furthermore, this technique is very sensitive to the bin size and can cause errors in the estimation as shown by Goldfeld et al. (2019b). In the same work, the authors also proposed a rigorous framework to estimate $I(X;T_l)$ through the addition of Gaussian noise in the latent representation of the DNNs. Their approach introduces system dependency into the mutual information measurement. In other words, changing the variance of the noise influences the DNN prediction, thereby allowing Goldfeld *et al.* to investigate what the mutual information is capturing within the network. They found that both estimation of $I(X;T_l)$ in a noisy network or $I(X;\text{Bin}(T_l))$ in deterministic net-

works, track the amount of clustering in the the latent representation. Further, they attributed the intriguing observations of compression in DNNs, as reported by Schwarz-Ziv and Tishby, to this geometric phenomenon of clustering. Compression refers to the phenomenon of decrease in MI between the input and internal representations over the course of training, and was initially proposed as a plausible causal factor for the generalization ability of DNNs.

As a result, in order to gain more insight into the DNN's representation, we need to study the geometry of the latent space distribution. Through this analysis, we hope to facilitate future research into the dynamics of DNN learning. To this end, we propose to study the clustering of data points in the latent space through the use of density-based spatial clustering of applications with noise (DBSCAN, Ester et al. (1996)). By using a distance threshold, we can connect points in the latent space in order to form a graphical representation of the data. DBSCAN can then be used to find clusters within the graphs. Finally, by measuring how clustered the data points are, we can have a proxy for mutual information. The advantage of working with graphs is that we can remove the dependency of dimension of the latent space in estimating the mutual information, allowing us to start investigating more real-world datasets. We applied our method to Shwartz-Ziv Tishby's network Shwartz-Ziv & Tishby (2017) and the MNIST digits dataset. Our estimate of mutual information shows similar trend to that of Goldfeld et al. (2019b) while taking much less computational time.

2 METHODOLOGY

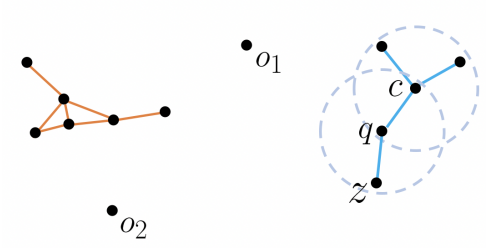

Figure 1: Illustration of DBSCAN with 4 *minPts*. Point c is a core point. The ϵ-radius for some points are shown as dashed circle. Point q is directly reachable by c. Whereas z is reachable by c. Outliers cannot be reached by core points and does not *minPts*, they are displayed as $o_{1,2}$ Different colors represent distinct clusters.

In order to estimate mutual information $I(X;T_l)$ in a meaningful manner, we can add noise into the representation of the neural network (Goldfeld et al. (2019b)). This measurement, as shown previously, tracks the degree with which the latent space is clustered. On the other hand, even though binned mutual information is not a good estimate of the actual mutual information, it also tracks the amount of how well the representation is clustered in the same space. As a result, it is natural to implement a more direct way to identify clustering as a proxy to estimate mutual information. We propose to use a density-based clustering technique, known as DBSCAN (Ester et al. (1996)). Density-based clustering methods, more precisely DBSCAN, provides a principled way of formalizing the notion of what we call a cluster. DBSCAN finds the number of dense regions in the latent space based solely on its neighborhood graph. Suppose that the latent space is a constellation with dense regions separated by sparse areas. DBSCAN discovers clusters by giving the same labels to closely packed points and marking sparse points as outliers.

Quantifying density of the space is controlled by two main hyperparameters, namely ϵ-neighbor or -radius and minimum number of points (*minPts*). A single object is called a core-point c if it contains at least *minPts* points within an ϵ-radius, counting itself. Objects that are within ϵ-radius of a reference point p are said to be directly reachable by p. A point z is said to be reachable by c if there is a path from the core point. In other words, for a given sequence of points $(p_0, ..., p_i, p_{i+1}, .., p_z)$, we must have $p_0 = c$, $p_z = z$, and p_{i+1} must be directly reachable by p_i. This is illustrated in fig 1, where point z is reachable since q bridges a connection from the core point c. At the beginning of the DBSCAN, randomly picks a point in the latent constellation that has not been visited. It checked whether it is a core point. If it is a core point, then it checks which other points are reachable by c.

Every object within this group is given a cluster label. Points which are not core or reachable by any core points are labeled as noise.

The time complexity of DBSCAN is at most $\mathcal{O}(n^2)$, where n is the number of data point. If an indexing structure is used, such as kd-tree, the time complexity can be as low as $\mathcal{O}(n \log n)$ (Han et al. (2011)). For the purpose of estimating mutual information, we should not discard any noise points as they contribute to the measurement by increasing the entropy of the representation. Therefore, we set the *minPts* to be 1, which allows each isolated point to be a core point even if there is no other point within its ϵ-radius. Thus, each isolated point increases the cluster count by 1. *It is important to note that with the minPts parameter set to 1, the number of clusters identified by DBSCAN is simply equal to the number of connected components in an ϵ - neighborhood graph*. Because of efficient neighborhood indexing in DBSCAN, the number of clusters and clustering labels are estimated much faster than traditional graph-based methods for computing connected components. Once the clusters are found, we can estimate the entropy.

To measure the amount of clustering, which we use as the proxy of $I(X;T_l)$, we propose two methods. The first one is simply $\log(\#\, clusters)$. For the second estimate, we first find the maximum distance d_i between points within a cluster for the ith label. If the cluster label only contains 1 point, then we assign $d_i = \epsilon$. We assign a probability of $p_i = d_i / \Sigma_i d_i$ for each ith labeled cluster. Finally, we compute the entropy $H_{clusters} = -\Sigma_i p_i \log p_i$. We compare our estimates against binned mutual information at different bin sizes over the course of training.

3 RESULTS

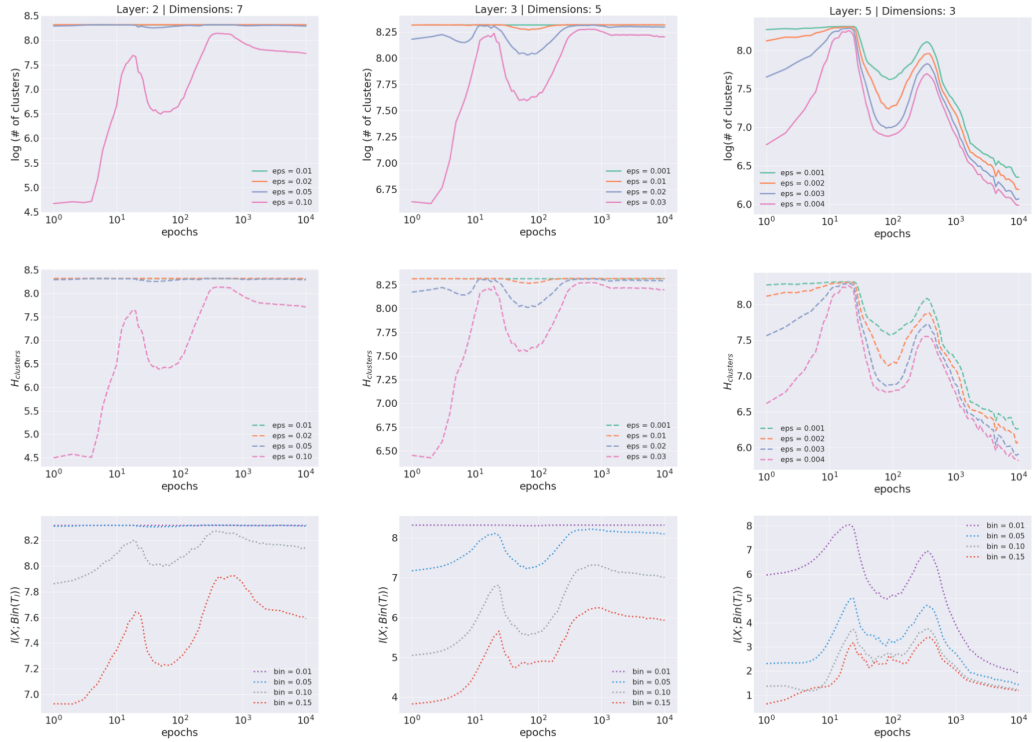

Figure 2: Evolution of approximate MI measures across training epochs for the Shwarz-Tishby model. Each column represents a different layer of the network, whereas each row represents a different approximation technique for Mutual Information. Top row depicts log(# of clusters), middle row shows the clustering entropy and bottom row shows the binned MI estimates. Left to right, the columns represents Layer 2,3 and 5, containing 7,5 and 3 hidden neurons respectively.

4 EXPERIMENTS

We conducted our experiments on two different datasets. The first one uses the same data set as that of Shwartz-Ziv & Tishby (2017) for which the mutual information trajectories over the course of training have been well-characterized previously. It consists of a binary classification task on a 12-dimensional input using a fully connected DNN. The architecture has six layers (12–10–7–5–4–3–2) and $tanh$ non-linearities only. We measured the mutual information at several intermediate layers of the network. We compared our estimates to $I(X; \text{Bin}(T_l))$, which we know the behavior of.

For the second experiment, we utilized the MNIST dataset. We followed the experimental protocol of Saxe et al., who trained a fully-connected DNN on MNIST with $tanh$ non-linearities. The architecture comprised five dense layers of size 784-1024-20-20-20-10 successively.

4.1 SZT MODEL

Figure 2 depicts the estimated clusterability measures, namely (i) logarithm of the number of densely connected components or clusters in the representation and (ii) entropy of clusters, over the course of training at different epsilon thresholds. As can be seen from the figure, both these measures capture the trends of binned MI fairly accurately. Consistent with Tishby's observations, binned MI does exhibit a compression phenomemon, where it gradually begins to decrease during training. However, as pointed out by Ziv et al., this compression is simply a result of progressive geometric clustering of internal representations. This geometric phenomenon is visually depicted in Figure 3.

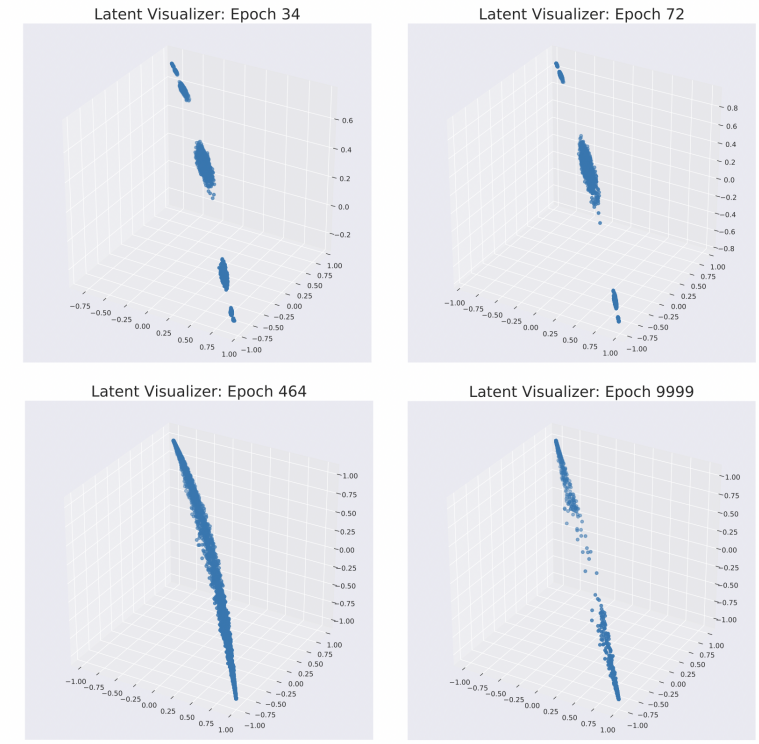

Figure 3: Scatter plots during different training epochs illustrating the clustering phenomenon in Shwarz-Tishby model

4.2 MNIST FULLY-CONNECTED NETWORK

Figure 4 shows the mutual information trajectories of the penultimate layer containing 20 hidden neurons. Again, we observe a similar trend. Beyond 1000 epochs, both the binned MI and the clusterability-based MI estimates start to gradually drop. This provides further evidence for the

observation that evolution of binned MI or the entropy of internal representations is simply tracking clusterability of internal representations.

A critical hyperparameter of the proposed approach is ϵ, which determines the neighborhood radius. At present, we do not have a principled way of choosing this parameter. Heuristically, we examined the histogram of pairwise distances between the representations at the zero-eth epoch to identify the lowest 5 percentile cut-off and estimated the clusterability measures in the vicinity of this threshold. However, a more formal approach for determining ϵ is needed. We leave this exploration for future work.

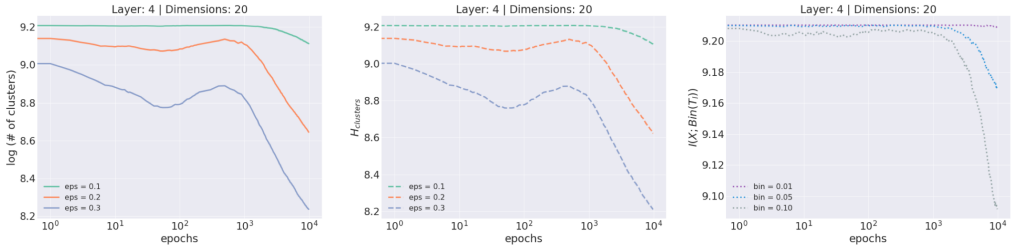

Figure 4: Evolution of approximate MI measures across training epochs for the Saxe et al. MNIST network in the penultimate layer with 20 hidden neurons

5 Conclusion

In this work, we presented a novel approach to estimate mutual information that relies solely on the neighborhood graph of internal representations. We employed density-based clustering to quantify clusterability of the induced graph structure. We demonstrated that notions of clusterability captured by this graphical perspective closely follow the trajectory of mutual information over the course of training. More importantly, the proposed method scales only linearly in dimensionality of hidden representations and is thus likely to scale much better on real-world datasets. Approximate MI estimation methods can afford an information-theoretic understanding of learning dynamics in larger networks. In future, we want to investigate the feasibility of using this estimation technique on state-of-the-art networks and real-world recognition datasets. Consequently, with this method, we hope to probe open questions about the relationship between compression/clustering and generalization.

References

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, pp. 226–231. AAAI Press, 1996. URL `http://dl.acm.org/citation.cfm?id=3001460.3001507`.

Ziv Goldfeld, Kristjan Greenewald, Yury Polyanskiy, and Jonathan Weed. Convergence of smoothed empirical measures with applications to entropy estimation. *arXiv preprint arXiv:1905.13576*, 2019a.

Ziv Goldfeld, Ewout Van Den Berg, Kristjan Greenewald, Igor Melnyk, Nam Nguyen, Brian Kingsbury, and Yury Polyanskiy. Estimating information flow in deep neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 2299–2308, 2019b.

Jiawei Han, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

Andrew Michael Saxe, Yamini Bansal, Joel Dapello, Madhu Advani, Artemy Kolchinsky, Brendan Daniel Tracey, and David Daniel Cox. On the information bottleneck theory of deep learning. 2018.

Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.