SCALABLE MUTUAL INFORMATION ESTIMATION USING DEPENDENCE GRAPHS

*Morteza Noshad, Yu Zeng, Alfred O. Hero III**

University of Michigan, Electrical Engineering and Computer Science, Ann Arbor, Michigan, U.S.A

ABSTRACT

The Mutual Information (MI) is an often used measure of dependency between two random variables utilized in information theory, statistics and machine learning. Recently several MI estimators have been proposed that can achieve parametric MSE convergence rate. However, most of the previously proposed estimators have high computational complexity of at least $O(N^2)$. We propose a unified method for empirical non-parametric estimation of general MI function between random vectors in $\mathbb{R}^d$ based on N i.i.d. samples. The reduced complexity MI estimator, called the ensemble dependency graph estimator (EDGE), combines randomized locality sensitive hashing (LSH), dependency graphs, and ensemble bias-reduction methods. We prove that EDGE achieves optimal computational complexity $O(N)$, and can achieve the optimal parametric MSE rate of $O(1/N)$ if the density is d times differentiable. To the best of our knowledge EDGE is the first non-parametric MI estimator that can achieve parametric MSE rates with linear time complexity. We illustrate the utility of EDGE for the analysis of the information plane (IP) in deep learning. Using EDGE we shed light on a controversy on whether or not the compression property of information bottleneck (IB) in fact holds for ReLu and other rectification functions in deep neural networks (DNN).

1. INTRODUCTION

The Mutual Information (MI) is an often used measure of dependency between two random variables or vectors [1], and it has a wide range of applications in information theory [1] and machine learning [2, 3]. Non-parametric MI estimation methods have been studied that use estimation strategies including KSG [4], KDE [5] and Parzen window density estimation [6]. The performance of these estimators has been evaluated and compared based on both empirical studies [7] and asymptotic analysis [8]. Recently several MI estimators have been proposed that can achieve parametric MSE rate of convergence. For example, in [9] a KDE plug-in estimator for Rényi divergence and mutual information achieves the MSE rate of $O(1/N)$ when the densities are at least d times differentiable. Another KDE based mutual information estimator was proposed in [8] that can achieve the MSE rate of $O(1/N)$ when the densities are $d/2$ times differentiable. Recently Moon et al [10] and Gao et al [11] respectively proposed KDE and KNN based MI estimators for random variables with mixtures of continuous and discrete components. Most of these estimators, however, have high computational cost and require knowledge of the density support boundary.

In this paper we propose a reduced complexity MI estimator called the ensemble dependency graph estimator (EDGE). The estimator combines randomized locality sensitive hashing (LSH), dependency graphs, and ensemble bias-reduction methods. A dependence graph is a bipartite directed graph consisting of two sets of nodes V and U. The data points are mapped to the sets V and U using a randomized LSH function H that depends on a hash parameter ϵ. Each node is assigned a weight that is proportional to the number of hash collisions. Likewise, each edge between the vertices v_i and u_j has a weight proportional to the number of (X_k, Y_k) pairs mapped to the node pairs (v_i, u_j). For a given value of the hash parameter ϵ, a base estimator of MI is proposed as a weighted average of non-linearly transformed of the edge weights. The proposed EDGE estimator of MI is obtained by applying the method of weighted ensemble bias reduction [10, 12] to a set of base estimators with different hash parameters. This estimator is a non-trivial extension of the LSH divergence estimator defined in [13]. LSH-based methods have previously been used for KNN search and graph constructions problems [14, 15], and they result in fast and low complexity algorithms.

Recently, Shwartz-Ziv and Tishby utilized MI to study the training process in Deep Neural Networks (DNN) [16]. Let X, T and Y respectively denote the input, hidden and output layers. The authors of [16] introduced the information bottleneck (IB) that represents the tradeoff between two mutual information measures: $I(X, T)$ and $I(T, Y)$. They observed that the training process of a DNN consists of two distinct phases; 1) an initial fitting phase in which $I(T, Y)$ increases, and 2) a subsequent compression phase in which $I(X, T)$ decreases. Saxe *et al* in [17] countered the claim of [16], asserting that this compression property is not universal, rather it depends on the specific activation function. Specifically, they claimed that the compression property does not hold for ReLu activation functions. The authors of [16] challenged these claims, arguing that the authors of [17] had not observed compression due to poor estimates of the MI. We use our proposed rate-optimal ensemble MI estimator to explore this

*This research was partially supported by ARO grant W911NF-15-1-0479.

controversy, observing that our estimator of MI does exhibit the compression phenomenon in the ReLU network studied by [17]. Our contributions are as follows:

- To the best of our knowledge the proposed MI estimator is the first estimator to have linear complexity and can achieve the optimal MSE rate of $O(1/N)$.
- The proposed MI estimator provides a simplified and unified treatment of mixed continuous-discrete variables. This is due to the hash function approach that is adopted.
- EDGE is applied to IB theory of deep learning, and provides evidence that the compression property does indeed occur in ReLu DNNs, contrary to the claims of [17].

The rest of the paper is organized as follows. In Section2, we introduce the general definition of MI and define the dependence graph. In Section 3, we introduce the hash based MI estimator and give theory for the bias and variance. In section 4 we introduce the ensemble dependence graph MI estimator (EDGE) and show how the ensemble estimation method can be used to improve the convergence rates. Finally, in Section 5 we provide numerical results as well as study the IP in DNNs.

2. MUTUAL INFORMATION

In this section, we introduce the general mutual information function based on the f-divergence measure. Then, we define a consistent estimator for the mutual information function. Consider the probability measures P and Q on a Euclidean space $\mathcal{X}$. Let $g : (0,\infty) \to \mathbb{R}$ be a convex function with $g(1) = 0$. The f-divergence between P and Q can be defined as follows [18, 19].

$$D(P\|Q) := \mathbb{E}_Q\left[g\left(\frac{dP}{dQ}\right)\right]. \tag{1}$$

Mutual Information: Let $\mathcal{X}$ and $\mathcal{Y}$ be Euclidean spaces and let P_{XY} be a probability measure on the space $\mathcal{X} \times \mathcal{Y}$. For any measurable sets $A \subseteq \mathcal{X}$ and $B \subseteq \mathcal{Y}$, we define the marginal probability measures $P_X(A) := P_{XY}(A \times \mathcal{Y})$ and $P_Y(B) := P_{XY}(\mathcal{X} \times B)$. Similar to [11, 18], the general MI denoted by $I(X,Y)$ is defined as

$$D(P_{XY}\|P_XP_Y) = \mathop{\mathbb{E}}_{P_XP_Y}\left[g\left(\frac{dP_{XY}}{dP_XP_Y}\right)\right], \tag{2}$$

where $\frac{dP_{XY}}{dP_XP_Y}$ is the Radon-Nikodym derivative, and $g : (0,\infty) \to \mathbb{R}$ is, as in (1) a convex function with $g(1) = 0$. Shannon mutual information is a particular cases of (1) for which $g(x) = x\log x$.

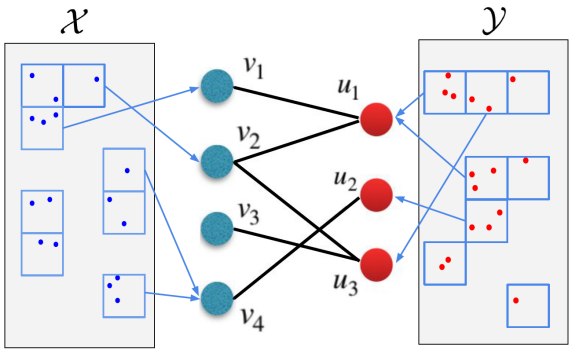

Fig. 1. Sample dependence graph with 4 and 3 respective distinct hash values of $\mathbf{X}$ and $\mathbf{Y}$ data jointly encoded with LSH, and the corresponding dependency edges.

2.1. Dependence Graphs

Consider N i.i.d samples (X_i, Y_i), $1 \le i \le N$ drawn from the probability measure P_{XY}, defined on the space $\mathcal{X} \times \mathcal{Y}$. Define the sets $\mathbf{X} = \{X_1, X_2, ..., X_N\}$ and $\mathbf{Y} = \{Y_1, Y_2, ..., Y_N\}$. The dependence graph $G(\mathbf{X}, \mathbf{Y})$ is a bipartite graph, consisting of two sets of nodes V and U with cardinalities denoted as $|V|$ and $|U|$, and the set of edges E_G. Each point in the sets $\mathbf{X}$ and $\mathbf{Y}$ is mapped to the nodes in the sets U and V, respectively, using the hash function H, described as follows.

A vector valued hash function H is defined in a similar way as defined in [13]. First, define the vector valued hash function $H_1 : \mathbb{R}^d \to \mathbb{Z}^d$ as

$$H_1(x) = [h_1(x_1), h_1(x_2), ..., h_1(x_d)], \tag{3}$$

where x_i denotes the ith component of the vector x. In (3), each scalar hash function $h_1(x_i) : \mathbb{R} \to \mathbb{Z}$ is given by

$$h_1(x_i) = \left\lfloor \frac{x_i + b}{\epsilon} \right\rfloor, \tag{4}$$

for a fixed $\epsilon > 0$, where $\lfloor y \rfloor$ denotes the floor function (the smallest integer value less than or equal to y), and b is a fixed random variable in $[0, \epsilon]$. Let $\mathcal{F} := \{1, 2, .., F\}$, where $F := c_H N$ and c_H is a fixed tunable integer. We define a random hash function $H_2 : \mathbb{Z}^d \to \mathcal{F}$ with a uniform density on the output and consider the combined hashing function

$$H(x) := H_2(H_1(x)), \tag{5}$$

which maps the points in $\mathbb{R}^d$ to $\mathcal{F}$.

$H(x)$ reveals the index of the mapped vertex in $G(X,Y)$. The weights ω_i and ω'_j corresponding to the nodes v_i and u_j, and ω_{ij}, the weight of the edge (v_i, u_j), are defined as follows.

$$\omega_i = \frac{N_i}{N}, \qquad \omega'_j = \frac{M_j}{N}, \qquad \omega_{ij} = \frac{N_{ij}N}{N_iM_j}, \tag{6}$$

where N_i and M_j respectively are the the number of hash collisions at the vertices v_i and u_j, and N_{ij} is the number of joint

collisions of the nodes (X_k, Y_k) at the vertex pairs (v_i, u_j). The number of hash collisions is defined as the number of instances of the input variables map to the same output value. In particular,

$$N_{ij} := \#\left\{(X_k, Y_k) \text{ s.t } H(X_k) = i \text{ and } H(Y_k) = j\right\}. \quad (7)$$

Fig. 1 represents a sample dependence graph. Note that the nodes and edges with zero collisions do not show up in the dependence graph.

3. THE BASE ESTIMATOR OF MI

3.1. Assumptions

The following are the assumptions we make on the probability measures and g:

A1. The support sets $\mathcal{X}$ and $\mathcal{Y}$ are bounded.

A2. The following supremum exists and is bounded:

$$\sup_{P_X P_Y} g\left(\frac{dP_{XY}}{dP_X P_Y}\right) \leq U.$$

A3. Let x_D and x_C respectively denote the discrete and continuous components of the vector x. Also let $f_{X_C}(x_C)$ and $p_{X_D}(x_D)$ respectively denote density and pmf functions of these components associated with the probability measure P_X. The density functions $f_{X_C}(x_C)$, $f_{Y_C}(y_C)$, $f_{X_C Y_C}(x_C, y_C)$, and the conditional densities $f_{X_C|X_D}(x_C|x_D)$, $f_{Y_C|Y_D}(y_C|y_D)$, $f_{X_C Y_C|X_D Y_D}(x_C, y_C|x_D, y_D)$ are Hölder continuous.

Hölder continuous functions: Given a support set $\mathcal{X}$, a function $f : \mathcal{X} \to \mathbb{R}$ is called Hölder continuous with parameter $0 < \gamma \leq 1$, if there exists a positive constant G_f, possibly depending on f, such that for every $x \neq y \in \mathcal{X}$,

$$|f(y) - f(x)| \leq G_f \|y - x\|^{\gamma}. \quad (8)$$

A4. Assume that the function g in (2) is Lipschitz continuous; i.e. g is Hölder continuous with $\gamma = 1$.

3.2. The Base Estimator of MI

For a fixed value of the hash parameter ϵ, we propose the following base estimator of MI (2) function based on the dependence graph:

$$\widehat{I}(X, Y) := \sum_{e_{ij} \in E_G} \omega_i \omega'_j \widetilde{g}(\omega_{ij}), \quad (9)$$

where the summation is over all edges $e_{ij} : (v_i \to u_j)$ of $G(X, Y)$ having non-zero weight and $\widetilde{g}(x) := \max\{g(x), U\}$.

When X and Y are strongly dependent, each point X_k hashed into the bucket (vertex) v_i corresponds to a unique hash value for Y_k in U. Therefore, asymptotically $\omega_{ij} \to 1$ and the mutual information estimation in (9) takes its maximum value. On the other hand, when X and Y are independent, each point X_k hashed into the bucket (vertex) v_i may be associated with different values of Y_k, and therefore asymptotically $\omega_{ij} \to \omega_j$ and the Shannon MI estimation tends to 0.

3.3. Convergence Rates

In the following theorems we state upper bounds on the bias and variance rates of the proposed MI estimator (9). The proofs are given in appendices A and B of the arXiv version [20]. We define the notations $\mathbb{B}[\hat{T}] = \mathbb{E}[\hat{T}] - T$ for bias and $\mathbb{V}[\hat{T}] = \mathbb{E}[\hat{T}^2] - \mathbb{E}[\hat{T}]^2$ for variance of $\hat{T}$. The following theorem states an upper bound on the bias.

Theorem 3.1. *Let $d = d_X + d_Y$ be the dimension of the joint random variable (X, Y). Under the aforementioned assumptions **A1-A4**, and assuming that the density functions in **A3** have bounded derivatives up to order $q \geq 0$, the following upper bound on the bias of the estimator in* (9) *holds*

$$\mathbb{B}\left[\widehat{I}(X, Y)\right] = \begin{cases} O(\epsilon^{\gamma}) + O\left(\frac{1}{N\epsilon^d}\right), & q = 0 \\ \sum_{i=1}^{q} C_i \epsilon^i + O(\epsilon^q) + O\left(\frac{1}{N\epsilon^d}\right) & q \geq 1, \end{cases} \quad (10)$$

where ϵ is the hash parameter in (4)*, γ is the smoothness parameter in* (8)*, and C_i are real constants.*

In (10), the hash parameter, ϵ needs to be a function of N to ensure that the bias converges to zero. For the case of $q = 0$, the optimum bias is achieved when $\epsilon = \left(\frac{1}{N}\right)^{\gamma/(\gamma+d)}$. When $q \geq 1$, the optimum bias is achieved for $\epsilon = \left(\frac{1}{N}\right)^{1/(1+d)}$.

Theorem 3.2. *Under the assumptions **A1-A4** the variance of the proposed estimator can be bounded as $\mathbb{V}\left[\widehat{I}(X, Y)\right] \leq O\left(\frac{1}{N}\right)$. Further, the variance of the variable ω_{ij} is also upper bounded by $O(1/N)$.*

4. ENSEMBLE DEPENDENCE GRAPH ESTIMATOR (EDGE)

Given the expression for the bias in Theorem 3.1, the ensemble estimation technique proposed in [12] can be applied to improve the convergence rate of the MI estimator (9). Assume that the densities in **A3** have continuous bounded derivatives up to the order q, where $q \geq d$. Let $\mathcal{T} := \{t_1, ..., t_T\}$ be a set of index values with $t_i < c$, where $c > 0$ is a constant. Let $\epsilon(t) := tN^{-1/2d}$. For a given set of weights $w(t)$ the weighted ensemble estimator is then defined as

$$\widehat{I}_w := \sum_{t \in \mathcal{T}} w(t) \widehat{I}_{\epsilon(t)}, \quad (11)$$

where $\widehat{I}_{\epsilon(t)}$ is the mutual information estimator with the parameter $\epsilon(t)$. Using (10), for $q > 0$ the bias of the weighted ensemble estimator (11) takes the form

$$\mathbb{B}(\hat{I}_w) = \sum_{i=1}^{q} CiN^{-\frac{i}{2d}} \sum_{t \in \mathcal{T}} w(t) t^i + O\left(\frac{t^d}{N^{1/2}}\right) + O\left(\frac{1}{N\epsilon^d}\right) \quad (12)$$

Given the form (12), as long as $T \geq q$, we can select the weights $w(t)$ to force to zero the slowly decaying terms

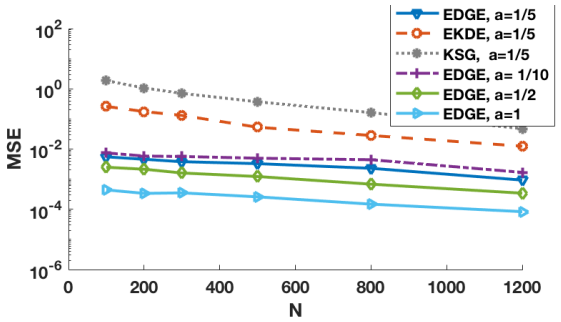

Fig. 2. MSE comparison of EDGE, EDKE and KSG Shannon MI estimators. X is a $2D$ Gaussian random variable with unit covariance matrix. $Y = X + aN_U$, where N_U is a uniform noise. The MSE rates of EDGE, EKDE and KSG are compared for various values of a.

in (12), i.e. $\sum_{t\in\tau} w(t)t^{i/d} = 0$ subject to the constraint that$\sum_{t\in\tau} w(t) = 1$. However, T should be strictly greater than q in order to control the variance, which is upper bounded by the euclidean norm squared of the weights ω. In particular we have the following theorem (the proof is given in Appendix C of the arXiv version [20]):

Theorem 4.1. *For $T > d$ let w_0 be the solution to:*

$$\begin{aligned} \min_{w} \quad & \|w\|_2 \\ \textit{subject to} \quad & \sum_{t\in\mathcal{T}} w(t) = 1, \\ & \sum_{t\in\mathcal{T}} w(t)t^i = 0, i \in \mathbb{N}, i \leq d. \end{aligned} \tag{13}$$

Then the MSE rate of the ensemble estimator $\widehat{I}_{w_0}$ is $O(1/N)$.

5. EXPERIMENTS

We first use simulated data to compare the proposed estimator to the competing MI estimators Ensemble KDE (EKDE) [10], and generalized KSG [11]. Both of these estimators work on mixed continuous-discrete variables.

Fig. 2, shows the MSE estimation rate of Shannon MI between the continuous random variables X and Y having the relation $Y = X + aN_U$, where X is a $2D$ Gaussian random variable with the mean $[0, 0]$ and covariance matrix $C = I_2$. Here I_d denote the d-dimensional identity matrix. N_U is a uniform random vector with the support $\mathcal{N}_U = [0, 1] \times [0, 1]$. We compute the MSE of each estimator for different sample sizes. The MSE rates of EDGE, EKDE and KSG are compared for $a = 1/5$. Further, the MSE rate of EDGE is investigated for noise levels of $a = \{1/10, 1/5, 1/2, 1\}$. As the dependency between X and Y increases the MSE rate becomes slower.

Next, we use EDGE to study the information bottleneck in DNNs. Fig. 3 represents the information plane of a DNN with 4 fully connected hidden layers of width $784 - 1024 - 20 - 20 - 20 - 10$ with tanh and ReLU activations. The network is trained with Adam optimization with a learning rate of

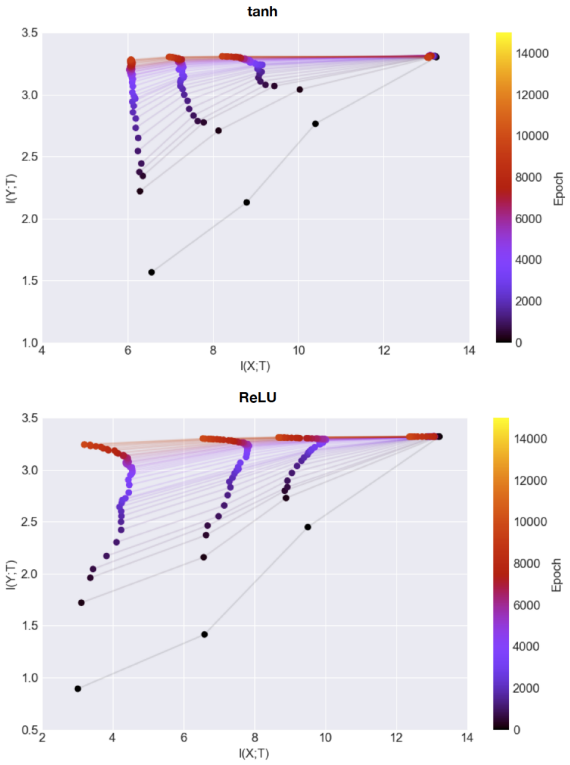

Fig. 3. Information plane estimated using EDGE for a neural network of size $784 - 1024 - 20 - 20 - 20 - 10$ trained on the MNIST dataset with tanh (top) and ReLU (bottom) activations.

0.003 and cross-entropy loss functions to classify the MNIST handwritten-digits dataset. We repeat the experiment for 20 iterations with different randomized initializations and take the average over all experiments. In both cases of ReLU and tanh activations we observe compression in all of the hidden layers. However, the amount of compressions is different for ReLU and tanh activations. The average test accuracy in both of these networks are around 0.98. This network is the same as the one studied in [17], for which it is claimed that no compression happens with a ReLU activation. The base estimator used in [17] provides KDE-based lower and upper bounds on the true MI [21]. According to our experiments (not shown) the upper bound is in some cases twice as large as the lower bound. In contrast, our proposed ensemble method estimates the exact mutual information with significantly higher accuracy. More experiments on simulated and real datasets are provided in the arXiv version [20].

6. CONCLUSION

In this paper we proposed a fast non-parametric estimation method for MI based on random hashing, dependence graphs, and ensemble estimation. Remarkably, the proposed estimator has linear computational complexity and attains optimal (parametric) rates of MSE convergence. We provided bias and variance convergence rate, and validated our results by numerical experiments.

7. REFERENCES

[1] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

[2] H. Liu, L. Liu, and H. Zhang, "Ensemble gene selection for cancer classification," *Pattern Recognition*, vol. 43, no. 8, pp. 2763–2772, 2010.

[3] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

[4] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical Review E*, vol. 69, no. 6, p. 066138, 2004.

[5] Y. Moon, B. Rajagopalan, and U. Lall, "Estimation of mutual information using kernel density estimators," *Physical Review E*, vol. 52, no. 3, p. 2318, 1995.

[6] N. Kwak and C.-H. Choi, "Input feature selection by mutual information based on parzen window," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 12, pp. 1667–1671, 2002.

[7] S. Khan, S. Bandyopadhyay, A. R. Ganguly, S. Saigal, D. J. Erickson III, V. Protopopescu, and G. Ostrouchov, "Relative performance of mutual information estimation methods for quantifying the dependence among short and noisy data," *Physical Review E*, vol. 76, no. 2, p. 026209, 2007.

[8] K. Kandasamy, A. Krishnamurthy, B. Poczos, L. Wasserman, *et al.*, "Nonparametric von mises estimators for entropies, divergences and mutual informations," in *Advances in Neural Information Processing Systems*, pp. 397–405, 2015.

[9] S. Singh and B. Póczos, "Exponential concentration of a density functional estimator," in *Advances in Neural Information Processing Systems*, pp. 3032–3040, 2014.

[10] K. R. Moon, K. Sricharan, and A. O. Hero III, "Ensemble estimation of mutual information," *Proceedings of the IEEE Intl Symp. on Information Theory (ISIT), Aachen, June 2017*.

[11] W. Gao, S. Kannan, S. Oh, and P. Viswanath, "Estimating mutual information for discrete-continuous mixtures," in *Advances in Neural Information Processing Systems*, pp. 5988–5999, 2017.

[12] K. R. Moon, K. Sricharan, K. Greenewald, and A. O. Hero, "Improving convergence of divergence functional ensemble estimators," in *IEEE International Symposium Inf Theory*, pp. 1133–1137, IEEE, 2016.

[13] M. Noshad and A. O. Hero III, "Scalable hash-based estimation of divergence measures," *Proceedings of the 22nd Conference on Artificial Intelligence and Statistics, Canary Islands, March 2018, arXiv:1801.00398*.

[14] Y. Zhang, K. Huang, G. Geng, and C. Liu, "Fast kNN graph construction with locality sensitive hashing," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 660–674, 2013.

[15] Q. Lv, W. Josephson, Z. Wang, M. Charikar, and K. Li, "Multi-probe LSH: efficient indexing for high-dimensional similarity search," in *Proceedings of the 33rd International Conference on Very Large Data Bases*, pp. 950–961, VLDB Endowment, 2007.

[16] R. Shwartz-Ziv and N. Tishby, "Opening the black box of deep neural networks via information," *arXiv preprint arXiv:1703.00810*, 2017.

[17] A. M. Saxe, Y. Bansal, J. Dapello, M. Advani, A. Kolchinsky, B. D. Tracey, and D. D. Cox, "On the information bottleneck theory of deep learning," *ICLR*, 2018.

[18] Y. Polyanskiy and Y. Wu, "Strong data-processing inequalities for channels and bayesian networks," in *Convexity and Concentration*, pp. 211–249, Springer, 2017.

[19] I. Csiszár, "Generalized cutoff rates and renyi's information measures," *IEEE Tran on Information Theory*, vol. 41, no. 1, pp. 26–34, 1995.

[20] M. Noshad and A. O. Hero III, "Scalable mutual information estimation using dependence graphs," *arXiv preprint arXiv:1801.09125*, 2018.

[21] A. Kolchinsky and B. D. Tracey, "Estimating mixture entropy with pairwise distances," *Entropy*, vol. 19, no. 7, p. 361, 2017.