Estimating Information Flow in DNNs

Ziv Goldfeld[1,3], Ewout van den Berg[2,3], Kristjan Greenewald[2,3], Igor Melnyk[2,3]
Brian Kingsbury[2,3], Nam Nguyen[2,3], Yury Polyanskiy[1,3]
[1]MIT, [2]IBM Research, [3]MIT-IBM Watson AI Lab

IBM Research AI
MIT-IBM Watson AI Lab

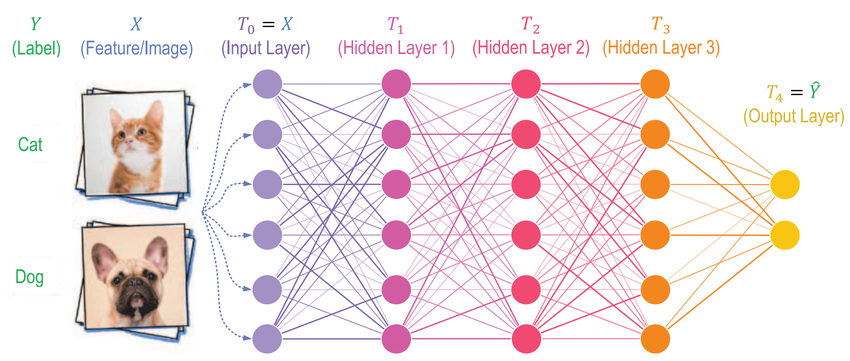
Deep Learning - What's Under the Hood?

- **Lacking Theory:** Macroscopic understanding of Deep Learning
  - What drives the evolution of internal representations?
  - What are properties of learned representations?
  - How fully trained networks process information?
- **Attempts to Understand Effectiveness of DL:**
  - Loss landscape [Saxe *et al.*'14, Choromanska *et al.*'15, Kawaguchi'16, Keskar *et al.*'17]
  - Wavelets and sparse coding [Bruna-Mallat'13, Giryes *et al.*'16, Papyan *et al.*'16]
  - Adversarial examples [Szegedy *et al.*'14, Nguyen *et al.*'17, Liu *et al.*'16, Cisse *et al.*'16]
  - Information Bottleneck Theory [Tishby-Zaslavsky'15, Shwartz-Tishby'17, Saxe *et al.*'18]

★ **Goal:** IB theory mathematical analysis & test 'compression' phenomenon

Information Bottleneck

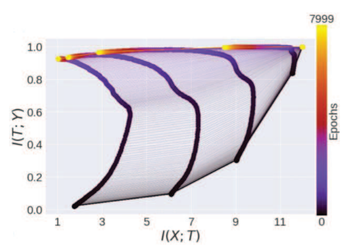
(Deterministic) Feedforward DNN: Each layer $T_\ell = f_\ell(T_{\ell-1})$



- **Information Plane:** Evolution of $\big(I(X;T_\ell), I(Y;T_\ell)\big)$ during training

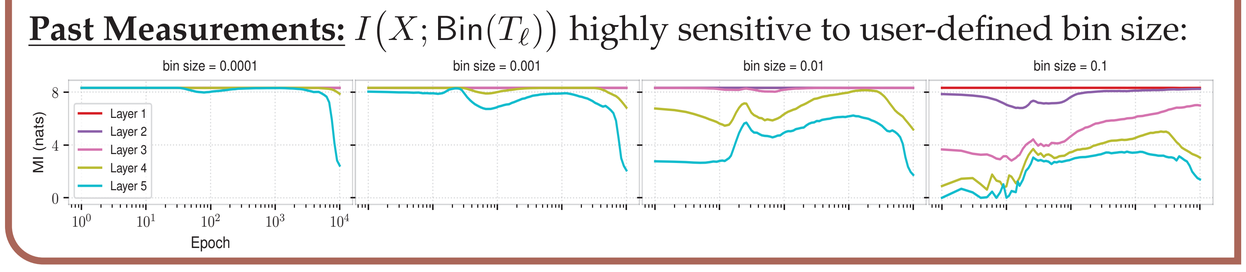
IB Theory Claim: Training comprises 2 phases

1. **Fitting:** $I(Y;T_\ell)$ & $I(X;T_\ell)$ rise (short)
2. **Compression:** $I(X;T_\ell)$ slowly drops (long)



Proposition 1 (Informal). *Det. DNNs with strictly monotone nonlinearities (e.g., tanh or sigmoid)* $\implies$ $I(X;T_\ell)$ *is* ***independent of the DNN parameters***

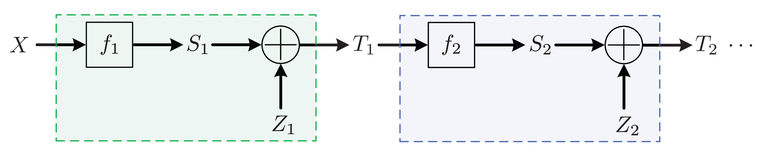
Past Measurements: $I\big(X;\mathrm{Bin}(T_\ell)\big)$ highly sensitive to user-defined bin size:



Noisy Neural Networks

Modification: Inject (small) Gaussian noise to neurons' output

- **Formally:** $T_\ell = S_\ell + Z_\ell$, where $S_\ell \triangleq f_\ell(T_{\ell-1})$ and $Z_\ell \sim \mathcal{N}(0, \sigma^2 \mathrm{I}_d)$



$\implies$ $X \mapsto T_\ell$ is a **parametrized channel** (by DNN's parameters)

$\implies$ $I(X;T_\ell)$ is a **function** of parameters!

Mutual Information Estimation

Setup: Estimate $h(P * \mathcal{N}_\sigma)$ from n i.i.d. samples $S^n \triangleq (S_i)_{i=1}^n$ of $P \in \mathcal{F}_d$.

Theorem 1 (Goldfeld-Greenewald-Polyanskiy-Weed'19). *Sample complexity of any accurate estimator (additive gap η) is* $\Omega\left(\frac{2^d}{\eta d}\right)$.

Structured Estimator: $\hat{h}(S^n, \sigma) \triangleq h(\hat{P}_n * \mathcal{N}_\sigma)$, where $\hat{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{S_i}$

Theorem 2 (GGPW'19). *For* $\mathcal{F}_{d,K}^{(\mathsf{SG})} \triangleq \{P \,|\, P$ *is* K*-subgaussian in* $\mathbb{R}^d\}$, $d \geq 1$ *and* $\sigma > 0$, *we have* $\sup_{P \in \mathcal{F}_{d,K}^{(\mathsf{SG})}} \mathbb{E}_{S^n}\left|h(P * \mathcal{N}_\sigma) - \hat{h}(S^n, \sigma)\right| \leq c_{\sigma,K}^d \cdot n^{-\frac{1}{2}}$.

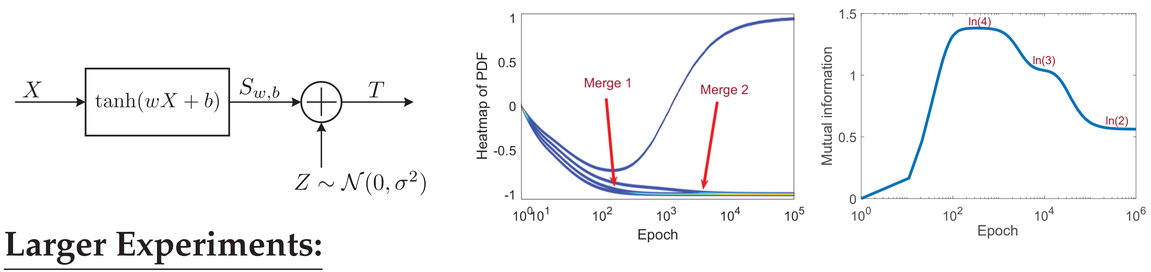
Optimality: $\hat{h}(S^n, \sigma)$ attains sharp dependence on both n and d!

Clustering as the Driver of Compression

Single Neuron Classification:

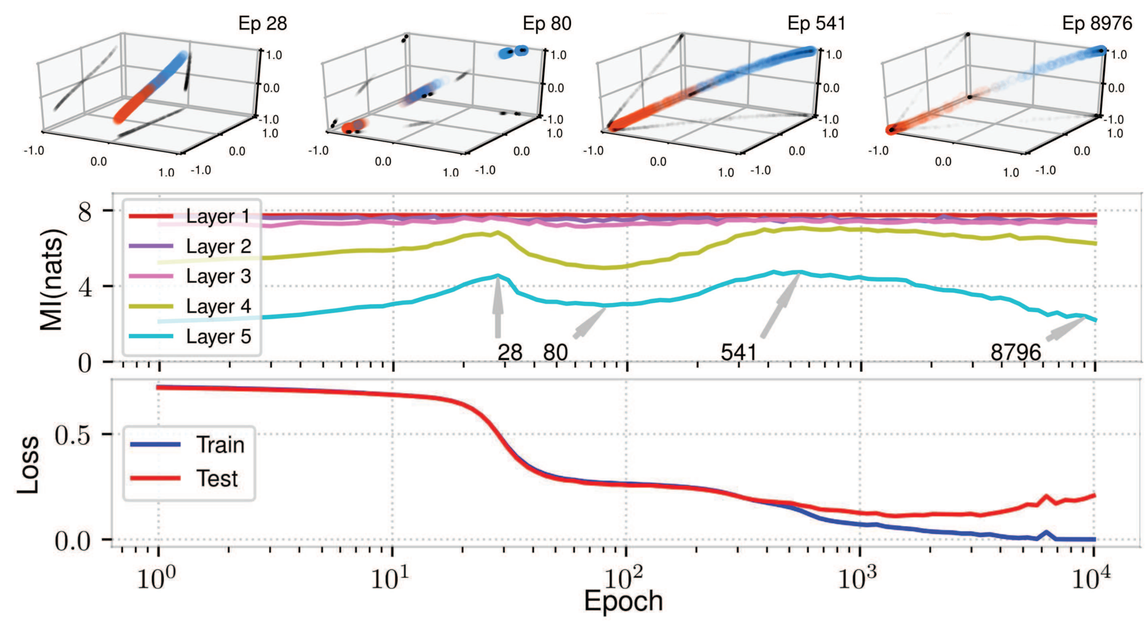
- **Input:** $X \sim \mathsf{Unif}\{\pm 1, \pm 3\}$, $\mathcal{X}_{y=-1} \triangleq \{-3, -1, 1\}$, $\mathcal{X}_{y=1} \triangleq \{3\}$

$\implies$ $I(X;T)$ is # bits (nats) transmittable over AWGN with symbols $\mathcal{S}_{w,b} \triangleq \{\tanh(-3w+b), \tanh(-w+b), \tanh(w+b), \tanh(3w+b)\}$

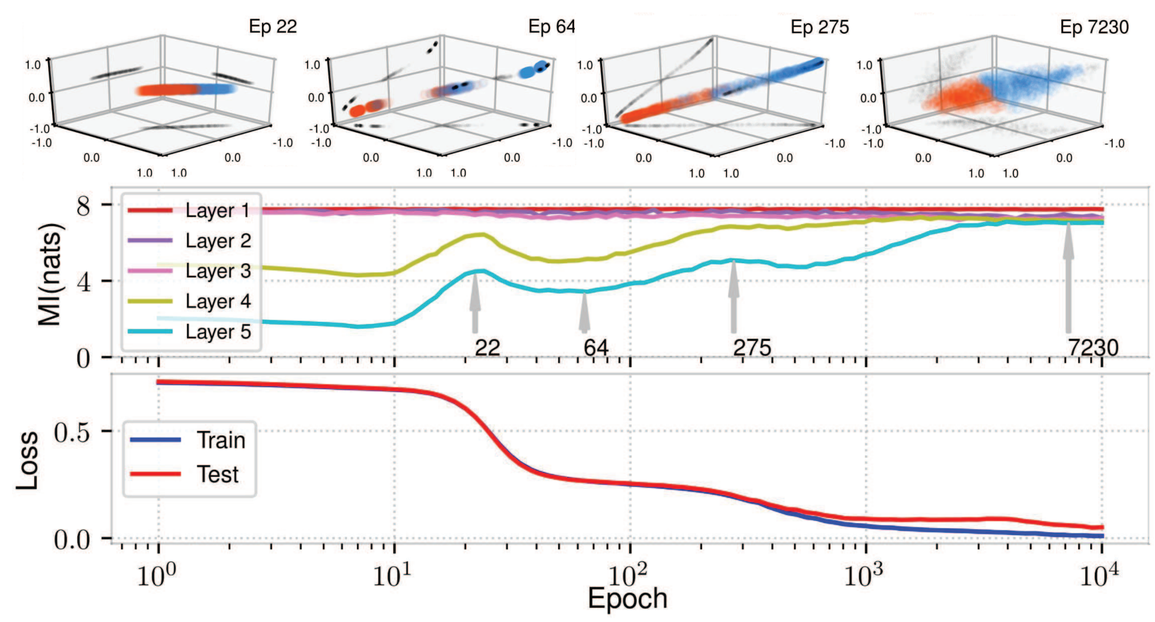


Larger Experiments:

- **Binary Classification:** 12-bit input & 12–**10**–**7**–**5**–**4**–**3**–2 tanh MLP



- Weight orthonormality regularization [Cisse *et al.*'17]:



Relevance to Deterministic Nets

- Noisy DNNs: Compression driven by clustering of representations.
- Clustering is meaningful in deterministic nets and can be measured.
- Binned "mutual information" measured in past works measures clustering.

Conclusion and Future Work

- Geometric clustering of internal representations is the phenomenon underlying observed "information compression".
- Compression is not necessary for generalization.
- Proposed framework for studying IT quantities over DNNs.
  - Optimal estimator (in n and d) for accurate MI estimation.
- Future Work:
  - Methods of tracking geometric clustering in high dimensions.
  - Further exploration of clustering phenomenon.
  - Potential DNN regularization schemes.