Information Flow in Deep Neural Networks

Mohammadamin Kiani   Mohammad Mohammadian   Diba Hadi

Sharif University of Technology
Information Theory, Statistics, and Learning
Instructor: Prof. Yassaee
Project Mentor: Mr. Hadavi

Sep 4, 2025

Outline

Information Plane

- Treat the whole layer T, as a single random variable, characterized by its encoder, $P(T|X)$, and decoder, $P(Y|T)$ distributions.
- The plane of the mutual information values of any other variable with the input variable X and the desired output variable Y.
- **Rational:**
  - invariance to invertible re-parameterization $I(X;Y) = I(\psi(X);\psi(y))$
  - data processing inequality $I(X;Y) \geq I(X;Z)$
- Information path: layers are mapped to K monotonic connected points in the plane.
  - $I(X;Y) \geq I(T_1;y) \geq \cdots \geq I(T_k;Y) \geq I(\hat{Y};Y)$
  - $H(X) \geq I(X;T_1) \geq \cdots \geq I(X;T_k) \geq I(X;\hat{Y})$

Motivation: Understanding DNN Internals

- **Goal:** Characterize information learned in DNN hidden layers.
- **Mutual Information (MI)** $I(\boldsymbol{X}; \boldsymbol{T}_\ell)$ as a measure:
  - $\boldsymbol{X}$: Input, $\boldsymbol{T}_\ell$: Layer ℓ representation.
  - Principled: Invariant to invertible ops, meaningful units (bits/nats).
- **Information Bottleneck (IB) Theory Context**:
  - Learn compressed $\boldsymbol{T}_\ell$ of $\boldsymbol{X}$, informative about target $\boldsymbol{Y}$.
  - Suggests training phases: Fitting ($I(\boldsymbol{T}_\ell; \boldsymbol{Y}) \uparrow$) and Compression ($I(\boldsymbol{X}; \boldsymbol{T}_\ell) \downarrow$).
  - IB Tradeoff: optimal trade-off between compression of X and prediction of Y
    - $T(X) = \min_{p(t|x), p(y|t), p(t)} I(X; T) - \beta I(T; Y)$
    - β determines the level of relevant information captured by the representation T

The Problem with $I(\boldsymbol{X}; \boldsymbol{T}_\ell)$ in Deterministic DNNs

- Deterministic DNN: $\boldsymbol{T}_\ell = f_\ell(\ldots f_1(\boldsymbol{X}))$.
- For common continuous nonlinearities:
  - Continuous $P_X \implies I(\boldsymbol{X}; \boldsymbol{T}_\ell) = \infty$.
  - Discrete P_X (e.g., dataset) $\implies I(\boldsymbol{X}; \boldsymbol{T}_\ell) = H(\boldsymbol{X})$ (constant).
- **Consequence:** True $I(\boldsymbol{X}; \boldsymbol{T}_\ell)$ is often vacuous for observing "compression dynamics."
- Observed changes in $I(\boldsymbol{X}; \mathrm{Bin}(\boldsymbol{T}_\ell))$ (binned MI) in prior work likely stem from the binning approximation, not true MI changes.

A Rigorous Framework: Noisy DNNs

- To make $I(\boldsymbol{X}; \boldsymbol{T}_\ell)$ well-defined & sensitive to parameters, map $\boldsymbol{X} \mapsto \boldsymbol{T}_\ell$ must be stochastic.
- **Proposed Model:** Add i.i.d. Gaussian noise at each hidden layer output.

$$\boldsymbol{T}_\ell = f_\ell(\boldsymbol{T}_{\ell-1}) + \boldsymbol{Z}_\ell, \quad \boldsymbol{Z}_\ell \sim \mathcal{N}(0, \beta^2 \boldsymbol{I}_{d_\ell})$$

  - $\boldsymbol{S}_\ell = f_\ell(\boldsymbol{T}_{\ell-1})$ is the "signal" part before noise.
  - $\boldsymbol{T}_\ell = \boldsymbol{S}_\ell + \boldsymbol{Z}_\ell$.
- This makes $I(\boldsymbol{X}; \boldsymbol{T}_\ell)$ finite and dependent on network weights.
- Data Processing Inequality holds: $\boldsymbol{X} - \boldsymbol{T}_1 - \cdots - \boldsymbol{T}_L$.

Estimating $I(\boldsymbol{X}; \boldsymbol{T}_\ell)$ in Noisy DNNs

- Definition: $I(\boldsymbol{X}; \boldsymbol{T}_\ell) = h(\boldsymbol{T}_\ell) - \mathbb{E}_{\boldsymbol{X}}[h(\boldsymbol{T}_\ell | \boldsymbol{X} = \boldsymbol{x})]$.
- Direct computation of differential entropies $h(\cdot)$ is intractable.
- **Sample Propagation (SP) Estimator Idea**:
  - PDF of $\boldsymbol{T}_\ell$: $p_{\boldsymbol{T}_\ell}(\boldsymbol{t}) = (p_{\boldsymbol{S}_\ell} * \phi_\beta)(\boldsymbol{t})$ (convolution).
    - $p_{\boldsymbol{S}_\ell}$: PDF of signal $\boldsymbol{S}_\ell$. ϕ_β: PDF of noise $\boldsymbol{Z}_\ell$.
  - Estimate $h(p_{\boldsymbol{T}_\ell})$ via $h(\hat{p}_{\boldsymbol{S}_\ell} * \phi_\beta)$, using empirical $\hat{p}_{\boldsymbol{S}_\ell}$ from samples of $\boldsymbol{S}_\ell$.
  - Similar approach for conditional term $h(p_{\boldsymbol{T}_\ell | \boldsymbol{X}=\boldsymbol{x}})$.

The Sample Propagation (SP) Estimator

Given dataset $\mathcal{X} = \{\boldsymbol{x}_i\}_{i=1}^n$:

1. **Estimate** $h(\boldsymbol{T}_\ell)$: Use $h(\hat{p}_{\mathcal{S}_\ell} * \phi_\beta)$.
   - Samples $\mathcal{S}_\ell = \{\boldsymbol{s}_{\ell,i}\}_{i=1}^n$ from $f_\ell(\ldots f_1(\boldsymbol{x}_i))$.
   - This is entropy of a Gaussian mixture centered at $\boldsymbol{s}_{\ell,i}$.
2. **Estimate** $h(\boldsymbol{T}_\ell | \boldsymbol{X} = \boldsymbol{x})$: Use $h(\hat{p}_{\mathcal{S}_\ell^{(\boldsymbol{x})}} * \phi_\beta)$.
   - For each $\boldsymbol{x}$, n_x samples $\mathcal{S}_\ell^{(\boldsymbol{x})}$ by passing $\boldsymbol{x}$ multiple times through $f_\ell(\ldots f_1(\cdot))$.

SP Estimator $\hat{I}_{SP}$:

$$\hat{I}_{SP} = h(\hat{p}_{\mathcal{S}_\ell} * \phi_\beta) - \frac{1}{n} \sum_{\boldsymbol{x} \in \mathcal{X}} h(\hat{p}_{\mathcal{S}_\ell^{(\boldsymbol{x})}} * \phi_\beta)$$

Entropies of Gaussian mixtures often computed via Monte Carlo Integration.

Theoretical Guarantees: Preliminaries

Focus: Estimating $h(P_S * \phi_\beta)$ from N i.i.d. samples $\mathcal{S}_N = \{\boldsymbol{S}_i\}$ from P_S.
Estimator: $h(\hat{P}_{\mathcal{S}_N} * \phi_\beta)$.

- **Minimax Absolute-Error Risk** over distribution class $\mathcal{F}$ for S:

$$R^*(N, \beta, \mathcal{F}) = \inf_{\hat{h}} \sup_{P_S \in \mathcal{F}} \mathbb{E} \left| h(P_S * \phi_\beta) - \hat{h}(\mathcal{S}_N, \beta) \right|$$

  Measures worst-case error for the *best possible* estimator.

Theoretical Guarantees: Preliminaries

Focus: Estimating $h(P_S * \phi_\beta)$ from N i.i.d. samples $\mathcal{S}_N = \{\boldsymbol{S}_i\}$ from P_S.
Estimator: $h(\hat{P}_{\mathcal{S}_N} * \phi_\beta)$.

- **Minimax Absolute-Error Risk** over distribution class $\mathcal{F}$ for S:

$$R^*(N, \beta, \mathcal{F}) = \inf_{\hat{h}} \sup_{P_S \in \mathcal{F}} \mathbb{E} \left| h(P_S * \phi_\beta) - \hat{h}(\mathcal{S}_N, \beta) \right|$$

  Measures worst-case error for the *best possible* estimator.

- **Sample Complexity** $N^*(\eta, \beta, \mathcal{F})$: Smallest N for $R^* \leq \eta$. Minimum samples needed to achieve a target accuracy η.
- Classes considered: $\mathcal{F}_d$ (distributions with support in $[-1, 1]^d$, e.g., for tanh layers), $\mathcal{F}^{(SG)}_{d,\mu,K}$ (subgaussian distributions, e.g., for ReLU layers with subgaussian inputs).

Guarantee 1: Sample Complexity

Statement (Simplified): For fixed noise $\beta > 0$, large dimension d, target error $\eta < \eta_0(\beta)$:

$$N^*(\eta, \beta, \mathcal{F}_d) \geq \Omega\left(\frac{2^{\gamma(\beta)d}}{d\eta}\right)$$

Guarantee 1: Sample Complexity

Statement (Simplified): For fixed noise $\beta > 0$, large dimension d, target error $\eta < \eta_0(\beta)$:

$$N^*(\eta, \beta, \mathcal{F}_d) \geq \Omega\left(\frac{2^{\gamma(\beta)d}}{d\eta}\right)$$

- **Core Implication:** Estimating $h(P_S * \phi_\beta)$ is fundamentally hard in high dimensions.
- The required number of samples N^* grows at least exponentially with dimension d.
- The exponent $\gamma(\beta)$ is monotonically decreasing in β.
  - Larger noise variance β^2 (larger β) $\implies$ smaller $\gamma(\beta)$ $\implies$ (relatively) less severe exponential dependence.
  - More noise "smooths" the distribution, making estimation easier.
- This is a lower bound, applying to *any* estimator, not just the SP one.

Guarantee 2: Risk of $h(\hat{P}_{\mathcal{S}_N} * \phi_\beta)$

Statement (Simplified): For $P_S \in \mathcal{F}^{(SG)}_{d,\mu,K}$ (or $\mathcal{F}_d$):

$$\mathbb{E}\left| h(P_S * \phi_\beta) - h(\hat{P}_{\mathcal{S}_N} * \phi_\beta) \right| \leq C(d, \mu, K, \beta) \cdot \frac{1}{\sqrt{N}}$$

Guarantee 2: Risk of $h(\hat{P}_{\mathcal{S}_N} * \phi_\beta)$

Statement (Simplified): For $P_S \in \mathcal{F}^{(SG)}_{d,\mu,K}$ (or $\mathcal{F}_d$):

$$\mathbb{E}\left|h(P_S * \phi_\beta) - h(\hat{P}_{\mathcal{S}_N} * \phi_\beta)\right| \leq C(d,\mu,K,\beta) \cdot \frac{1}{\sqrt{N}}$$

- **Core Implication:** The specific SP-based estimator $h(\hat{P}_{\mathcal{S}_N} * \phi_\beta)$ achieves the parametric rate of convergence $O(1/\sqrt{N})$ with respect to sample size N.
- This is generally the best possible convergence rate for parametric estimation problems.
- However, the constant $C(d,\mu,K,\beta)$ can be (and often is) exponential in dimension d.
  - $C(d,\mu,K,\beta) \approx \left(\frac{1}{\sqrt{2}} + \frac{K}{\beta}\right)^d \times$ polynomial factors in $d, \mu, K, 1/\beta$.
  - This reflects the curse of dimensionality.

Guarantee 3: MI Estimation Risk

The absolute-error risk of the full SP MI estimator $\hat{I}_{SP}$ (using n samples for unconditional term, and $n_x = n$ for each of n conditional terms):

$$\sup_{P_X} \mathbb{E} \left| I(\boldsymbol{X}; \boldsymbol{T}_\ell) - \hat{I}_{SP} \right| \leq 2\Delta_{\beta,d_\ell}(n) + \frac{d_\ell \log(1 + 1/\beta^2)}{4\sqrt{n}}$$

- $\Delta_{\beta,d_\ell}(n)$ is the risk bound for estimating a single entropy term (i.e., $O(C(d_\ell)/\sqrt{n})$).

Guarantee 3: MI Estimation Risk

The absolute-error risk of the full SP MI estimator $\hat{I}_{SP}$ (using n samples for unconditional term, and $n_x = n$ for each of n conditional terms):

$$\sup_{P_X} \mathbb{E} \left| I(\boldsymbol{X}; \boldsymbol{T}_\ell) - \hat{I}_{SP} \right| \leq 2\Delta_{\beta, d_\ell}(n) + \frac{d_\ell \log(1 + 1/\beta^2)}{4\sqrt{n}}$$

- $\Delta_{\beta,d_\ell}(n)$ is the risk bound for estimating a single entropy term (i.e., $O(C(d_\ell)/\sqrt{n})$).
- **Core Implication:** The overall MI estimation error also converges at the parametric rate $O(1/\sqrt{n})$.
- The bound depends on:
  - Layer dimension d_ℓ (through $\Delta_{\beta,d_\ell}(n)$ and the second term).
  - Noise variance β^2 (larger β can reduce the bound via both terms).
  - Number of samples n.
- The term $1/\beta^2$ relates to the Signal-to-Noise Ratio (SNR) between the signal $\boldsymbol{S}_\ell$ and noise $\boldsymbol{Z}_\ell$.

The Link: Compression & Geometric Clustering

- Consider $I(\boldsymbol{X};\boldsymbol{T}_\ell) = I(\boldsymbol{S}_\ell;\boldsymbol{S}_\ell + \boldsymbol{Z}_\ell)$.
- This is MI over an AWGN-like channel:
  - "Input symbols": deterministic pre-noise activations $\mathcal{S}_\ell = \{\boldsymbol{s}_{\ell,\boldsymbol{x}}\}$.
  - $I(\cdot)$ measures distinguishability of $\boldsymbol{s}_{\ell,\boldsymbol{x}}$ after adding noise $\boldsymbol{Z}_\ell$.
- **Hypothesis: Clustering drives compression.**
  - Training $\implies$ representations $\boldsymbol{s}_{\ell,\boldsymbol{x}}$ of inputs $\boldsymbol{x}$ from the *same class* may cluster.
  - Closer points in $\mathcal{S}_\ell \implies$ Gaussian components in $p_{\boldsymbol{T}_\ell}$ overlap more.
  - More overlap $\implies$ harder to resolve inputs $\implies$ reduction in $I(\boldsymbol{X};\boldsymbol{T}_\ell)$.
- Paper argues: "Compression" in deterministic nets via binned MI was tracking this clustering.

Key Conclusions

- $I(\boldsymbol{X}; \boldsymbol{T}_\ell)$ in *deterministic* DNNs is often ill-defined for studying representation dynamics.
- *Noisy DNN* framework allows rigorous study of $I(\boldsymbol{X}; \boldsymbol{T}_\ell)$.
- Sample Propagation (SP) estimator developed for $I(\boldsymbol{X}; \boldsymbol{T}_\ell)$ in noisy DNNs.
- Detailed theoretical guarantees for SP estimator (risk, sample complexity, bias) show $O(1/\sqrt{N})$ rate but highlight the curse of dimensionality (exponential dependence on d).
- "Compression" ($I(\boldsymbol{X}; \boldsymbol{T}_\ell) \downarrow$) in noisy nets is linked to geometric clustering of representations.

Proposed Work: Input Perturbations

Problem: How does $I(\boldsymbol{X}; \boldsymbol{T}_\ell)$ react to distributional shift $P_X \to P_{X'}$?
Perturbation model: Control the shift the input distribution to another one by some metric
Shift metrics:

- Wasserstein $W_p(P_X, P_{X'})$, Total Variation $\mathrm{TV}(P_X, P_{X'})$, $\mathrm{KL}(P_X \| P_{X'})$

Question:

- Q: Robustness of Estimation — how close is the estimated $I(\boldsymbol{X'}; \boldsymbol{T}_\ell)$ with the real information of original dataset?

Shift Metrics: KL, TV, Wasserstein

KL divergence

$$D_{\mathrm{KL}}(P\|Q) = \int p(x)\,\log\frac{p(x)}{q(x)}\,dx$$

- Asymmetric; $D_{\mathrm{KL}} \geq 0$ with equality iff $P{=}Q$.
- Likelihood-sensitive: large when $q(x){=}0$ where $p(x){>}0$.
- Useful for modeling/calibration where densities are available.

Total Variation (TV)

$$\mathrm{TV}(P,Q) = \tfrac{1}{2}\int |p(x)-q(x)|\,dx = \sup_{A}|P(A)-Q(A)|$$

- Symmetric, bounded: $0 \leq \mathrm{TV} \leq 1$.
- Support-overlap measure; ignores geometry of x.

Wasserstein W_p

$$W_p(P,Q) = \Big(\inf_{\gamma\in\Pi(P,Q)} \mathbb{E}_{(X,Y)\sim\gamma}\|X-Y\|^p\Big)^{1/p}$$

- Geometry-aware: depends on ground metric $\|\cdot\|$.
- Stable under small input shifts; meaningful even with disjoint supports.

Estimation Error under TV Shift

Theorem: Bounding the Estimator's Deviation under TV Shift

Let $\hat{I}_{SP}$ be the estimator built using n samples from P_X. If the data distribution shifts to $P_{X'}$ such that $\mathrm{TV}(P_X, P_{X'}) \le \varepsilon$, the expected deviation is bounded by:

$$\mathbb{E}\left| I(\boldsymbol{X}'; \boldsymbol{T}_\ell) - \hat{I}_{SP} \right| \le \underbrace{\left(\varepsilon \log(N_\ell - 1) + H_b(\varepsilon)\right)}_{\text{Shift Error}} + \underbrace{\left(\frac{8cd_\ell + d_\ell \log(1 + 1/\beta^2)}{4\sqrt{n}}\right)}_{\text{Estimation Error}}$$

where the second term is the explicit risk bound from the paper.

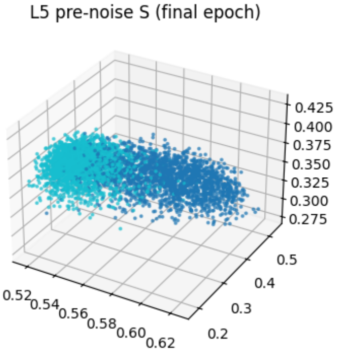
Estimation Error under KL Shift

Theorem: Bounding the Estimator's Deviation under KL Shift

Let $\hat{I}_{SP}$ be the estimator built using n samples from P_X. If the data distribution shifts to $P_{X'}$ such that $D_{\mathrm{KL}}(P_{X'} \| P_X) \leq \varepsilon$, the deviation is bounded by:

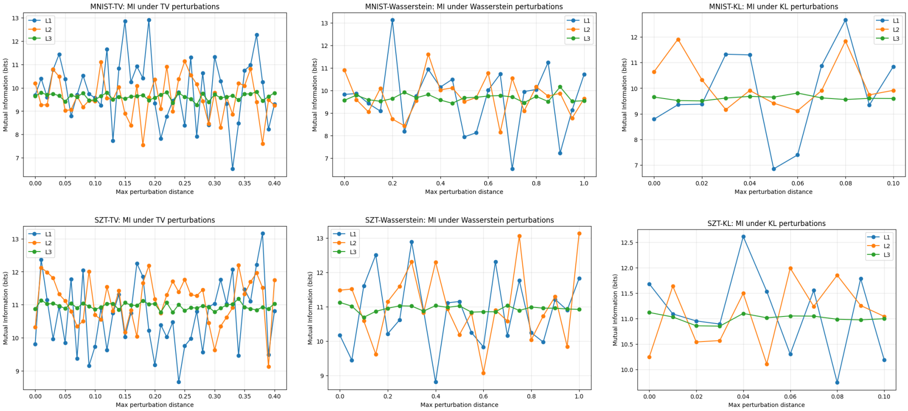
$$\mathbb{E}\left| I(\boldsymbol{X}'; \boldsymbol{T}_\ell) - \hat{I}_{SP} \right| \leq \underbrace{\left(\sqrt{\frac{\varepsilon}{2}} \log(N_\ell - 1) + H_b\left(\sqrt{\frac{\varepsilon}{2}}\right) \right)}_{\text{Shift Error}} + \underbrace{\left(\frac{8cd_\ell + d_\ell \log(1 + 1/\beta^2)}{4\sqrt{n}} \right)}_{\text{Estimation Error}}$$

L5 Pre-noise Representation S_{L5}



- Pre-noise activations S_{L5} reveal emerging class-wise clusters as training proceeds.

MI under shift — Latest SZT & MNIST



- Each panel: $I(\boldsymbol{X}; \boldsymbol{T}_\ell)$ vs. perturbation; legend shows layers.

Key References

[1] Goldfeld, Z., van den Berg, E., Greenewald, K., Melnyk, I., Nguyen, N., Kingsbury, B., & Polyanskiy, Y. (2019).
Estimating Information Flow in Deep Neural Networks.

[2] Shwartz-Ziv, R., & Tishby, N. (2017).
Opening the black box of Deep Neural Networks via Information.

[3] Tishby, N., & Zaslavsky, N. (2015).
Deep Learning and the Information Bottleneck Principle.

[4] Saxe, A. M., Bansal, Y., Dapello, J., Advani, M., Lampinen, A., Teh, B. D., & Ganguli, S. (2018).
On the Information Bottleneck Theory of Deep Learning.

[5] Goldfeld, Z., Greenewald, K., Polyanskiy, Y., & Weed, J. (2019).
Estimating Differential Entropy under Gaussian Convolutions.

Thank You!

Questions?