Deep Learning and Information Theory, and Graph Neural Network

June 2019

Presenter: Derrick Blakely

University of Virginia

https://qdata.github.io/deep2Read/

Table of contents

Information Theory Basics

Information Content

Information content = the amount you learn from an *event* E:

$$I(E) = -\log(Pr(E)) = \log\left(\frac{1}{Pr(E)}\right)$$

- Suppose you know $Pr(E) = 1$
- You don't learn anything when you're told E occurs
- $\implies I(E) = 0$
- Basic intuition: you learn more from surprising (i.e., unlikely) events (hence information content is also called "surprisal")

Weather Example

If it's sunny:

- Reduction in uncertainty $= 1/0.75 = 1.333$
- $I(S) = \log(1.333) = 0.41$

If it's raining:

- Reduction in uncertainty $= 1/0.25 = 4$
- $I(R) = \log(4) = 2$

Entropy

Entropy $=$ expected amount of information:

$$H(X) = -\sum_{x} Pr(x)\log(Pr(x))$$

"Amount of uncertainty about a random variable X"
"Virginia weather is unpredictable" $=$ "Virginia weather has high entropy"

Important Entropy Measures

- Joint entropy: $H(X, Y) = -\sum_{x,y} Pr(x, y) \log(Pr(x, y))$
- Conditional entropy: $H(Y|X) = -\sum_{x,y} Pr(x, y) \log\left(\frac{Pr(x,y)}{Pr(x)}\right)$
- If X and Y are independent: $H(Y|X) = H(Y)$
- If Y is a deterministic function of X: $H(Y|X) = 0$

Mutual Information

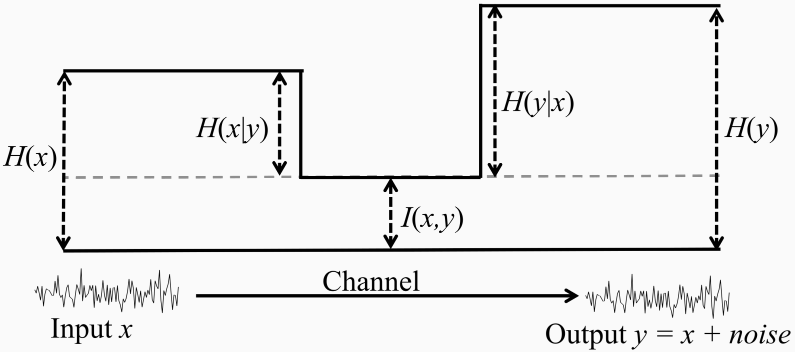
Mutual information:

$$I(X, Y) = H(X) - H(X|Y)$$

$$= -\sum_{x,y} Pr(x, y) \log\left(\frac{Pr(x)Pr(y)}{Pr(x, y)}\right)$$

- Amount of info gained about X when you observe Y
- Reduction in uncertainty about X when you observe Y
- If X and Y are independent, $I(X, Y) = 0$
- If X is a deterministic function of Y, $I(X, Y) = H(X) = H(Y)$

Mutual Information



KL Divergence

$$D_{KL}(P||Q) = -\sum_{x} P(x) \log\left(\frac{P(x)}{Q(x)}\right)$$

- Expected value of the log differences of two distributions
- Also called "relative entropy"
- Measure of difference between two distributions
- Not a distance metric
- Not symmetric

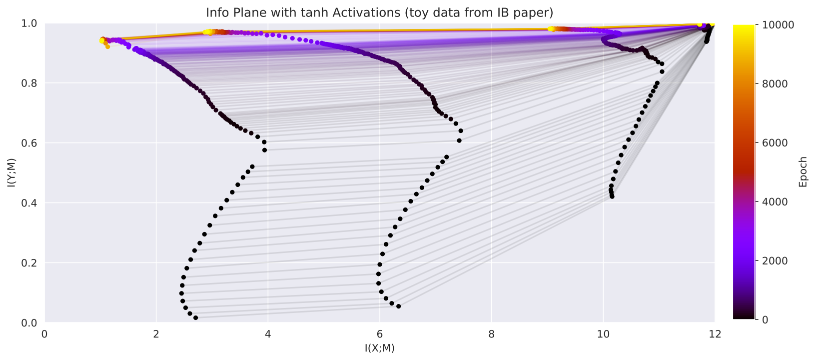
KL Divergence and Mutual Information

$$I(X, Y) = D_{KL}(Pr(X, Y) || Pr(X)Pr(Y))$$

- MI is just KL Divergence of product of marginals from the joint distribution
- I.e., amount of extra information needed if we use the marginals instead of the joint distribution

Information Bottleneck Theory

Information Plane



Information Bottleneck

Developed by Naftali Tishby's group [7, 6]
Uses the idea of the information plane and mutual information to argue:

1. DL uses two phases: (1) initial fitting phase and (2) compression phase
2. Compression phase causes DL's strong generalization performance
3. Compression phase occurs because of the diffusion-like behavior of SGD
4. MI is estimated with binning

Mutual Information Estimation

For each layer h activity, want to compute:

$$I(h; X) = H(h) - H(h|X)$$

The issue: h is not discrete

Continuous Activity Problem

If h is continuous then, let $h = Z$ (because we're already used H for entropy):

$$H(Z) = -\int_{\mathbb{R}} p_Z(z) \log p_Z(z) \mathrm{d}z$$

If X is a delta function (as it is in our case), then p_Z is a delta function, and so $H(Z) = -\infty$

Two Workarounds

To make $H(h)$ finite, we can try two approaches:

1. Discretize h by binning [6]
2. Add noise to convert h into a Gaussian mixture [3, 4, 5]

In both cases, we assume h is a vector of i.i.d dimensions.

Workaround 1: Binning

Do $T = \text{bin}(h)$ and compute p_i = the probability T_i is in bin b_i:

$$H(T) = -\sum_{i}^{N} p_i \log p_i$$

Because $f(X; W) = h$ is a deterministic mapping, we have:

$$H(T|X) = 0$$

Which means:

$$I(T; X) = H(T) - H(T|X) = H(T)$$

About Binning

- Valid way of approximating MI (it's what Tishby does in [6]), but has issues
- How to determine bin width?
- This is a hyperparam that makes a pretty big difference
- The "compression" stage of the IB theory could mostly just be *tanh* tending to map activities to the extreme bins (thus resembling a coin toss)

Workaround 2: Adding Noise

- Assume the observed distribution of samples = true distribution
- Use $T = h + \epsilon$ where $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$
- Aka, T is a mixture of Gaussians, with one Gaussian centered at each sample

Kernel Density Entropy (KDE) Estimation

Kolchinsky et al (2017) [3, 4] MI upper bounds:

$$I(T;X) = H(T) \leq -\frac{1}{P}\sum_i \log \frac{1}{P}\sum_j \exp\left(-\frac{1}{2}\frac{||h_i - h_j||_2^2}{\sigma^2}\right) = H(T)_u$$

And:

$$I(T;Y) = H(T) - H(T|Y)$$

$$\leq H(T)_u - \sum_l^L p_l \left[-\frac{1}{Pl}\sum_{i,Y_i=l} \log \frac{1}{Pl}\sum_{j,Y_j=l} \exp\left(-\frac{1}{2}\frac{||h_i - h_j||_2^2}{\sigma^2}\right)\right]$$

(Lower bounds are the same, except replace σ^2 with $4\sigma^2$)

IB Theory Pros

- Mutual information is a useful tool for exploring the relationships between outputs, inputs, and layers
- Information plane is a useful tool for visualizing training
- Tishby is right that hidden layers compression task-irrelevant information
- Bottleneck bound is probably useful

IB Theory Cons

- Refutation paper: [5]
- There isn't a general DL information plane; it depends greatly on the activations used
- The two-phase idea seems like it's entirely an artifact of using *tanh* layers (which no one uses...)
- No clear connection between compression and generalization; models with poor compression can generalize well
- Compression phase with *tanh* isn't actually caused by SGD
- Compression can occur during the training phase, not some distinct compression phase

Information Theory and the Spectral Domain

Graph Fourier Transform

Classical Fourier Transform:

$$\hat{x}(\zeta) = \langle x, e^{e\pi i\zeta t} \rangle = \int_{-\infty}^{\infty} x(t)e^{-2\pi i\zeta t}\mathrm{d}t = \mathcal{F}\{x(t)\} \tag{1}$$

Graph Fourier Transform:

$$\hat{x}(\lambda_l) = \langle \mathrm{x}, \mathrm{U} \rangle = \sum_{i=0}^{N-1} x(i)u_l^*(i) = \mathcal{F}\{x(i)\} \tag{2}$$

Convolution

Classical Convolution:

$$f(t) = (x * h)(t) = \int_{\mathbb{R}} x(\tau)h(t-\tau)\mathrm{d}\tau \tag{3}$$

Issue: how do you time shift using τ in the vertex domain? Convolution Theorem is useful:

$$\mathcal{F}\{x * h\} = \mathcal{F}\{x(t)\} \cdot \mathcal{F}\{h(t)\} \tag{4}$$

(This is also the theory behind FFT-based and Winograd convolution)

Graph Convolution

Using the convolution theorem and replacing complex exponentials with Laplacian eigenvectors:

$$(x * h)(i) = \sum_{l=0}^{N-1} \hat{x}(\lambda_l)\hat{h}(\lambda_l)u_l(i) \tag{5}$$

Interpretation: vertex-domain convolution = spectral domain element-wise multiplication

Computing Graph Convolutions

Another way of showing graph convolution:

$$h * x = \mathrm{U}\left((\mathrm{U}^\top h) \odot (\mathrm{U}^\top x)\right) = \mathrm{U}\hat{\mathrm{H}}\mathrm{U}^\top x \tag{6}$$

where $\hat{\mathrm{H}} = \mathrm{diag}(\hat{h_1}, ..., \hat{h_n}) = \hat{h}(\Lambda)$ are the spectral filter coefficients.

Computing Graph Convolutions

Approximations:

- Chebynets: Approximate $h * x$ with kth-order Chebyshev polynomials $\rightarrow \hat{h}_i = \hat{h}(\lambda_i) = (2 - \lambda_i)^k$
- GCN: set $k = 1$ and use normalized Laplacian with self-loops $\rightarrow \hat{h}_i = (1 - \lambda_i)^k$; approximate $k > 1$ with multiple layers

GCN:

$$h * x \approx \Theta \left(\widetilde{D}^{-1/2} \widetilde{A} \widetilde{D}^{-1/2} \right) x \tag{7}$$

Cross-Correlation

Classical cross-correlation:

$$R_{xh}(t) = (x \star h)(t) = \int_{\mathrm{R}} x(\tau)^* h(t+\tau)\mathrm{d}\tau \tag{8}$$

Cross-correlation theorem:

$$\mathcal{F}\{x \star h\} = \mathcal{F}\{x(t)\}^* \cdot \mathcal{F}\{h(t)\} \tag{9}$$

Graph cross-correlation:

$$R_{xh}(i) = (x \star h)(i) = \sum_{l=0}^{N-1} \hat{x}(\lambda_l)^* \hat{h}(\lambda_l) u_l(i) \tag{10}$$

(Note the complex conjugate; if $\hat{x}$ not complex, cross-correlation $=$ convolution)

Stationary Time-Series Processes

If $x(t)$ is a (strict) stationary time-series process, then:

1. $E[x_t] = \mu$ for some constant μ
2. $\text{Var}[x_t] = \sigma^2$ for some constant σ^2
3. $\text{Cov}(x_t, x_{t+h})$ is a function of the delay h but not t

Intuitively: $x(t)$ is always the same data-generating process.
Strict stationarity is required for time-series linear regression.

Spectral Density and Autocorrelation

Energy spectral density:

$$S_{xx}(\zeta) = |\hat{x}(\zeta)|^2 \tag{11}$$

Wiener-Khinchin Theorem: if $x(t)$ is a stationary random process:

$$S_{xx}(\zeta) = \mathcal{F}\{R_{xx}\} = \hat{R}_{xx}(\zeta) \tag{12}$$

Spectral Density and Cross-correlation

Spectral Density:

$$S_{xh}(\zeta) = \mathcal{F}\{R_{xh}\} = \mathcal{F}\{(x \star h)(\tau)\} \tag{13}$$

Spectral Entropy

Treat densities as unnormalized scores:

$$P(\lambda_i) = |x(\lambda_i)|^2 = S_{xx}(\lambda_i) = \hat{R}_{xx}(\lambda_i) \tag{14}$$

Normalize to treat as a probability density:

$$p_i = \frac{P(\lambda_i)}{\sum_j P(\lambda_j)} \tag{15}$$

Spectral entropy of $\hat{x}$:

$$H(\hat{x}) = -\sum_i p_i \log p_i \tag{16}$$

Spectral Density and Feature Locality

- Spectral density provides information on locality of feature distribution
- If the power spectrum decays at higher frequencies, it indicates local feature smoothness
- For "natural" images, [2] state:

$$E(|\hat{x}(\zeta)|^2) \sim \zeta^{-2} \tag{17}$$

Mutual Information and Frequency

[1] provides that for a pair of Gaussian stationary time-series processes $x(t)$ and $y(t)$:

$$I(x, y) = -\frac{1}{4\pi} \int_0^{2\pi} \log[1 - |R_{xy}(\lambda)|^2] \mathrm{d}\lambda \tag{18}$$

Can we define something similar for graph signals?

References i

D. R. Brillinger and A. Guha.
Mutual information in the frequency domain.
Journal of statistical planning and inference, 137(3):1076–1084, 2007.

J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun.
Spectral networks and locally connected networks on graphs.
arXiv preprint arXiv:1312.6203, 2013.

A. Kolchinsky and B. Tracey.
Estimating mixture entropy with pairwise distances.
Entropy, 19(7):361, 2017.

A. Kolchinsky, B. D. Tracey, and D. H. Wolpert.
Nonlinear information bottleneck.
arXiv preprint arXiv:1705.02436, 2017.

References ii

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.
2018.

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

N. Tishby, F. C. Pereira, and W. Bialek.
The information bottleneck method.
arXiv preprint physics/0004057, 2000.