Generalization of Deep Neural Networks

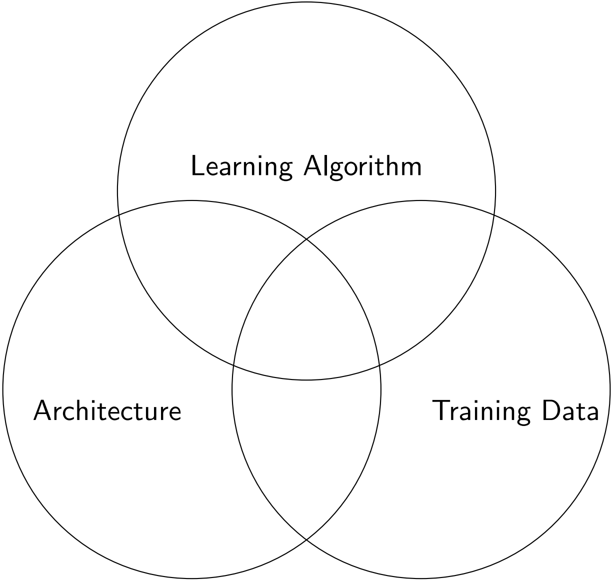
Thomas Walker

Imperial College London

thomas.walker21@imperial.ac.uk

July 23, 2023

The Components of Network Generalizations



Priors Informed by SGD[1]

Theorem

Let $\beta, \delta \in (0,1), n \in \mathbb{N}, \mathcal{D} \in \mathcal{M}_1(Z)$ *and* $P \in \mathcal{M}_1(\mathcal{H})$*. With probability at least* $1-\delta$ *over* $S \sim \mathcal{D}^n$*, for all* $Q \in \mathcal{M}_1(\mathcal{H})$*,*

$$L_{\mathcal{D}}(Q) \leq \Psi_{\beta,\delta}(Q,P;S) := \frac{1}{\beta} L_S(Q) + \frac{\mathrm{KL}(Q,P) + \log\left(\frac{1}{\delta}\right)}{2\beta(1-\beta)|S|}.$$

Require: Stopping criteria $\mathcal{E}$, Prefix fraction α, Batch size b.

function $\textsc{GetBound}(\mathcal{E}, \alpha, T, \sigma_P)$

$S_\alpha \leftarrow \{z_1, \ldots, z_{\alpha|S|} \subset S\}$

$w_\alpha^0 \leftarrow \mathrm{SGD}\left(w_0, S_\alpha, b, \frac{|S_\alpha|}{b}\right)$

$w_S \leftarrow \mathrm{SGD}\left(w_\alpha^0, S, b\infty, \mathcal{E}\right)$

$P \leftarrow \mathcal{N}\left(w_\alpha^G, \sigma_P I_p\right)$

$Q \leftarrow \mathcal{N}\left(w_S, \sigma_P I_p\right)$

Bound$\leftarrow \Psi_\delta^*(Q, P; S \setminus S_\alpha)$

return Bound

end function

[1] Dziugaite, Hsu, Gharbieh, and Roy 2020.

Mutual Information[2]

Definition

For two random variables X and Y, with joint distribution $p(x, y)$, their Mutual Information is defined as,

$$I(X; Y) = \mathrm{KL}(p(x, y), p(x)p(y)) = H(X) - H(X|Y),$$

where $H(X)$ and $H(X|Y)$ are the entropy and conditional entropy of X and Y.

Consider a K-layered deep neural network, with T_i denoting the representation of the i^{th} layer then there is a unique information path,

$$I(X; Y) \geq I(T_1; Y) \geq \cdots \geq I(T_k; Y) \geq I\left(\hat{Y}; Y\right),$$
$$H(X) \geq I(X; T_1) \geq \cdots \geq I(X; T_k) \geq I\left(X; \hat{Y}\right).$$

[2]Shwartz-Ziv and Tishby 2017.

Stiffness[3]

Let $\bar{g} = \nabla_W \mathcal{L}(f_W(X), y)$.

Definition

For two data points (X_1, y_1) and (X_2, y_2) define the sign stiffness to be

$$S_{\mathrm{sign}}\left(\left(X_1, y_1\right), \left(X_2, y_2\right); f\right) = \mathbb{E}\left(\mathrm{sign}\left(\bar{g}_1 \cdot \bar{g}_2\right)\right),$$

and the cosine stiffness to be

$$S_{\cos}\left((X_1, y_1), (X_2, y_2); f\right) = \mathbb{E}\left(\cos\left(\bar{g}_1 \cdot \bar{g}_2\right)\right),$$

where

$$\cos\left(\bar{g}_1 \cdot \bar{g}_2\right) = \frac{\bar{g}_1 \cdot \bar{g}_2}{|\bar{g}_1||\bar{g}_2|}.$$

[3] Fort, Nowak, and Narayanan 2019.

Coherence[4]

Definition

The coherence of a distribution $\mathcal{D}$ is defined to be

$$\alpha(\mathcal{D}) := \frac{\mathbb{E}_{z,z'\sim\mathcal{D}}(g_z \cdot g_{z'})}{\mathbb{E}_{z\sim\mathcal{D}(g_z\cdot g_z)}}.$$

Theorem

If stochastic gradient descent is run for T steps on a training set consisting of m examples drawn from distribution $\mathcal{D}$, then,

$$|\mathrm{gap}(\mathcal{D}, m)| \leq \frac{L^2}{m} \sum_{t=1}^{T} (\eta_k \beta)_{k=t+1}^{T} \cdot \eta_t \cdot \sqrt{2(1-\alpha(w_{t-1}))},$$

where $\mathrm{gap}(\mathcal{D}, m)$ is the expected difference between training and test loss over samples of size m from $\mathcal{D}$.

[4] Chatterjee and Zielinski 2022.

Persistent Homology[5]

Definition

Let δ be a metric on $\mathbb{R}^d$. The Vietoris-Rips complex at scale $\epsilon \geq 0$ on $X \subseteq \mathbb{R}^d$ is the abstract simplicial complex

$$\mathrm{VR}_\epsilon(X) := \{[x_0, \ldots, x_k\} : \delta(x_i, x_j) \leq 2\epsilon, x_0, \ldots, x_k \in X, k = 0, \ldots, n\}.$$

Let X be a sample from a manifold $M \subseteq \mathbb{R}^d$.

- At scale $\epsilon = 0$, then $\mathrm{VR}_0(X) = \{[x] : x \in X\}$, that is VR_0 overfits the data X.
- As $\epsilon \to \infty$ all the points of X become vertices of a single $|X|$-dimensional simplex.

A persistence barcode is an interval $[\epsilon, \epsilon')$ showing where a features emerges then disappears.

[5] Naitzat, Zhitnikov, and Lim 2020.

Architecture Design

Stiffness for to Architecture Design

References

- Shwartz-Ziv, Ravid and Naftali Tishby (2017). "Opening the Black Box of Deep Neural Networks via Information". In: *CoRR*.
- Fort, Stanislav, Pawel Krzysztof Nowak, and Srini Narayanan (2019). "Stiffness: A New Perspective on Generalization in Neural Networks". In: *CoRR*.
- Dziugaite, Gintare Karolina, Kyle Hsu, Waseem Gharbieh, and Daniel M. Roy (2020). "On the role of data in PAC-Bayes bounds". In: *CoRR*.
- Naitzat, Gregory, Andrey Zhitnikov, and Lek-Heng Lim (2020). "Topology of deep neural networks". In: *CoRR*.
- Chatterjee, Satrajit and Piotr Zielinski (2022). *On the Generalization Mystery in Deep Learning*.