On Regularizing Deep Learning using Mutual Information

Xu (Shell) Hu

February 14, 2019

École des Ponts ParisTech

École des Ponts
ParisTech

My research topics

My website: `http://hushell.github.io/`

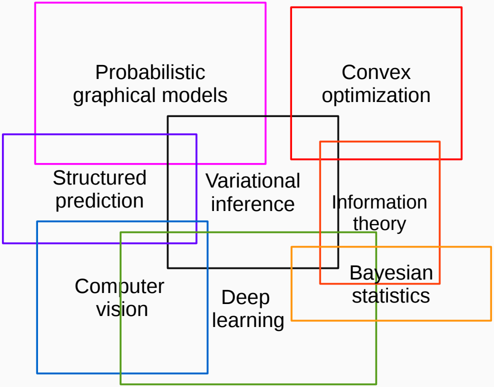


Table of contents

Motivation

Deep neural networks (DNNs) are over-parameterized

Over-parameterization: redundant parameterization (deeper or wider); even more parameters than training points.

Why do we use over-parametrized DNNs?

- Empirically good performance.
- Easier to train (Hinton et al., 2012; Denil et al., 2013).

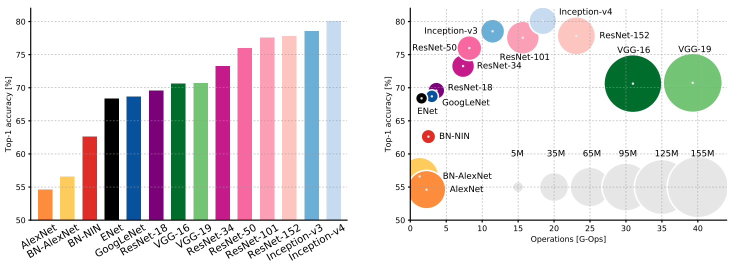
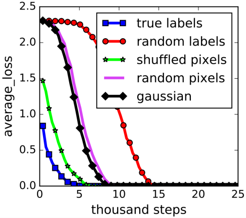


Figure 1: source: Figure 1 & 2 in Canziani et al. (2016).

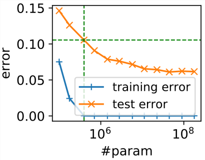
Over-parameterization and the generalization puzzle

Interesting observations on over-parameterized DNNs:

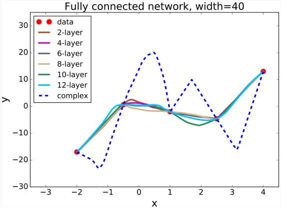
1. High capacity: achieve 0 training errors even with random data.
2. Do not overfit on real datasets as #params increasing.
3. Tend to converge to simple solutions.



Zhang et al. (2016)



Neyshabur et al. (2018)



Wu et al. (2017)

Over-parameterization and the generalization puzzle

Why gradient based optimizers can learn an over-parameterized DNN with small generalization error?

- Is it consistent with the *bias-variance tradeoff*?

  test error = estimator *variance* + squared estimator *bias* + noise.

- PAC learning with VC-dimension cannot explain this:

$$\text{generalization error} \leq \mathcal{O}\Big(\frac{\text{complexity}(\mathcal{H}_{\mathrm{DNN}})}{\sqrt{\#\text{points}}}\Big),$$

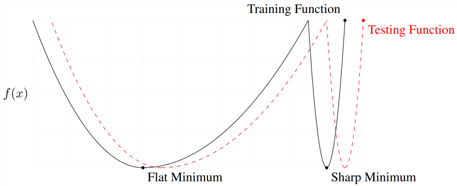
$$\text{VC-dimension}(\mathcal{H}_{\mathrm{DNN}}) = \mathcal{O}\big(\#\text{params} \cdot \log(\#\text{params})\big)$$

  The reasons:

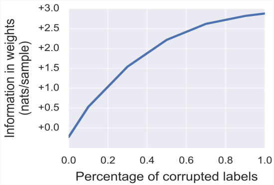
  - Loose bound: usually #points is smaller than #params.
  - Universal bound: it has to hold for all hypotheses in $\mathcal{H}_{\mathrm{DNN}}$.

Need to understand the regularization in deep learning

1. **Over-parameterization eliminates bad local minima** (Soudry and Hoffer, 2017; Kawaguchi, 2016; Lu and Kawaguchi, 2017; Li et al., 2017; Haeffele and Vidal, 2017; Wu et al., 2018).
2. **SGD biases towards low-complexity solutions**:
   - Flat minima conjecture (Keskar et al., 2016; Dinh et al., 2017).
   - Information bottleneck (Tishby et al., 2000): minimal sufficient activation (Tishby and Zaslavsky, 2015), minimal sufficient weights (Achille and Soatto, 2017).



Keskar et al. (2016)



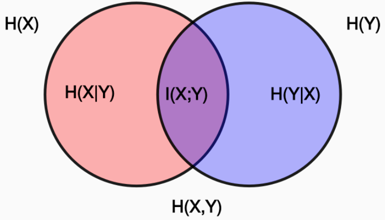
Achille and Soatto (2017)

Mutual Information

Mutual information: a math concept from Shannon

Mutual information measures statistical dependency

$$
\begin{aligned}
I(X;Y) &:= \mathbb{E}_{x,y\sim p(x,y)} \log \frac{p(x,y)}{p(x)p(y)} \\
&= H(X,Y) - H(X|Y) - H(Y|X) \\
&= H(X) - H(X|Y) \\
H(X) = I(X;X) &= \text{ expected amount of information in } X
\end{aligned}
$$



Mutual information is a functional of distributions

If we decompose the joint distribution as $p(x, y) = p(x)q(y|x)$, then the mutual information can be writen as a functional of p and q:

$$I(X; Y) \equiv I(p, q) := \mathbb{E}_{x,y\sim p(x,y)} \log \frac{q(y|x)}{q(y)} = \mathbb{E}_x D_{\mathrm{KL}}\big(q(y|x) \| q(y)\big),$$

$$q(y) := \sum_x p(x)q(y|x).$$

Issue: it is computationally difficult since $q(y|x)$ and $q(y)$ are coupled.

Variational characterization of mutual information

Lemma (Cover and Thomas, 2012, Theorem 10.8.1)

$$I(X;Y) = \max_{\phi(x|y)\in\Delta} \underbrace{\mathbb{E}_{x,y\sim p(x,y)} \log \frac{\phi(x|y)}{p(x)}}_{\tilde{I}(p,q,\phi)}$$

$$I(X;Y) = \min_{m(y)\in\Delta} \underbrace{\mathbb{E}_{x,y\sim p(x,y)} \log \frac{q(y|x)}{m(y)}}_{\hat{I}(p,q,m)}.$$

Learning as lossy compression: a rate-distortion perspective on Bayesian neural networks

Bayesian inference

A brief introduction:

- Bayesians describe data Y through the latent variable model

$$p(Y, w) = p(Y|w)p(w) = p(w) \prod_i p(y_i|w),$$

  assuming the *likelihood* $p(Y|w)$ and the *prior* $p(w)$ are given.

- Bayesians make predictions according to

$$p(y_{\text{new}}|Y) = \int p(y_{\text{new}}|w)p(w|Y)dw,$$

  where $p(w|Y)$ is the *posterior*.

Bayesian neural networks

Vanilla Bayesian neural networks (BNNs) by Hinton and Van Camp (1993); Graves (2011); Blundell et al. (2015):

- Given data S, approximate the posterior $p(w|S)$ by a Gaussian variational distribution $q(w|\theta^*)$ with *mean-field* form:

$$\begin{aligned}\theta^* &= \arg\min_\theta D_{\mathrm{KL}}\big(q(w|\theta)\|p(w|S)\big)\\ &= \arg\min_\theta \int q(w|\theta)\log\frac{q(w|\theta)}{p(w)p(S|w)}dw\\ &= \arg\min_\theta -\mathbb{E}_{q(w|\theta)}[\log p(S|w)] + D_{\mathrm{KL}}(q(w|\theta)\|p(w)).\end{aligned}$$

- Fix the prior $p(w)$ as Gaussian, Laplace, mixture of Gaussians or spike-and-slab distribution.

Rate-distortion tradeoff: a lossy compression framework

To induce a lossy compression of $X \to \hat{X}$, when $p(x)$ is given:

$$\min_{q(\hat{x}|x)\in\Delta} I(p,q)$$
$$\text{s.t.} \underbrace{\sum_{x,\hat{x}} p(x)q(\hat{x}|x)\ d(x,\hat{x})}_{D(p,q)} \leq \text{const.}$$

Plugging variational characterization and fixing the Lagrange multiplier β:

$$\min_{q(\hat{x}|x)\in\Delta} \min_{m(\hat{x})\in\Delta} \hat{I}(p,q,m) + \beta\ D(p,q).$$

An algorithm for rate-distortion tradeoff

The optimization problem of the rate-distortion tradeoff:

$$\min_{q(\hat{x}|x)\in\Delta} \min_{m(\hat{x})\in\Delta} \hat{I}(p, q, m) + \beta\ D(p, q).$$

Alternating projection algorithm (aka Blahut-Arimoto algorithm)

Provided an initial $q_t(\hat{x}|x)$ at $t = 0$. At iteration $t > 0$, taking the following steps:

$$q_t(\hat{x}|x) = \frac{m_t(\hat{x})e^{-\beta d(x,\hat{x})}}{\sum_{\hat{x}'} m_t(\hat{x}')e^{-\beta d(x,\hat{x})}},$$

$$m_{t+1}(\hat{x}) = \sum_x p(x)q_t(\hat{x}|x).$$

Then, the algorithm converges to a global minimum.

Rate-distortion perspective on supervised learning

Supervised learning as a lossy compression for the dataset S:

- We define the joint distribution by the graphical model $S \to w$:

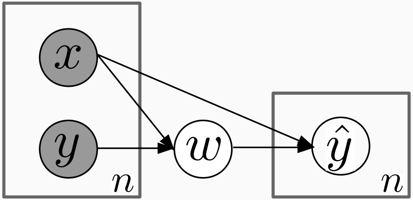
$$p(S, w) = q(w \mid S)p^*(S).$$

- As a comparison, Bayesians use a different decomposition:

$$P(S, w) = p(S \mid w)p(w).$$

- We make predictions according to

$$q(y \mid x, S) := \int p(y \mid x, w)q(w \mid S)dw.$$



Rate-distortion perspective on supervised learning

The lossy-compression induced objective:

$$\min_{q(w|S)\in\Delta} \Big[I\big(q(w|S), p^*(S)\big)\Big] \text{ s.t. } \mathbb{E}_{p^*(S)}\mathbb{E}_{q(w|S)} d(w,S) \leq D$$

$$I\big(q(w|S), p^*(S)\big) \equiv I(w;S) := \mathbb{E}_{p^*(S)}\mathbb{E}_{q(w|S)}\big[\log \frac{q(w|S)}{q(w)}\big],$$

$$d(w,S) := -\sum_{i=1}^{n} \log p(y_i|x_i,w).$$

Applying variational characterization, we obtain

$$I(w;S) \equiv \min_{m(w)\in\Delta} I(q,m) := \mathbb{E}_{p^*(S)}\mathbb{E}_{q(w|S)}\big[\log \frac{q(w|S)}{m(w)}\big].$$

Intuition: $I(w;S)$ is a regularizer, which forces w to contain less information about a particular S. Less memorization implies better generalization.

Approximate Blahut-Arimoto algorithm

1. We use a variational approximation $q(w|\theta)$ for $q(w|S)$ by solving

$$\begin{aligned}\theta(S) &= \arg\min_{\theta} D_{\mathrm{KL}}(q(w|\theta)\|q(w|S)) \\ &= \arg\min_{\theta} D_{\mathrm{KL}}(q(w|\theta)\|m(w)) + \beta\,\mathbb{E}_{q(w|\theta)}\big[d(w,S)\big].\end{aligned}$$

2. $m(w) \approx \sum_S p^*(S)q(w|\theta(S)) \approx \frac{1}{K}\sum_{k=1}^{K} q(w|\theta(B_k)) =: \tilde{m}(w)$, where B_k is a bootstrap sample of size n_b drawn from the empirical distribution $p_S(x,y) = \frac{1}{n}\sum_{i=1}^{n}\delta(x_i = x)\delta(y_i = y)$.

β-BNN

Input: S (dataset), β (coefficient), K (#mixture components), n_b (size of a bootstrap sample).
Initialize: $\Theta = \{\theta_k^{(0)} = (0, I)\}_{k=1}^K$; $\tilde{m}(w) = \frac{1}{K}\sum_{\theta\in\Theta} q(w|\theta)$.
for all $t = 1, \ldots, T$ **do**
    Draw K bootstrap samples $\{B_k\}_{k=1}^K$ of size n_b from $p_S(x, y)$.
    for all $k = 1, \ldots, K$ **do**
        $\theta_k^{(t)} \leftarrow \theta(B_k)$.
        $\Theta = \Theta \cup \{\theta_k^{(t)}\} \setminus \{\theta_k^{(t-1)}\}$.
        if do online update **or** $k = K$ **then**
            $\tilde{m}(w) = \frac{1}{K}\sum_{\theta\in\Theta} q(w|\theta)$.
Output: Θ.

Experiments: colorful MNIST

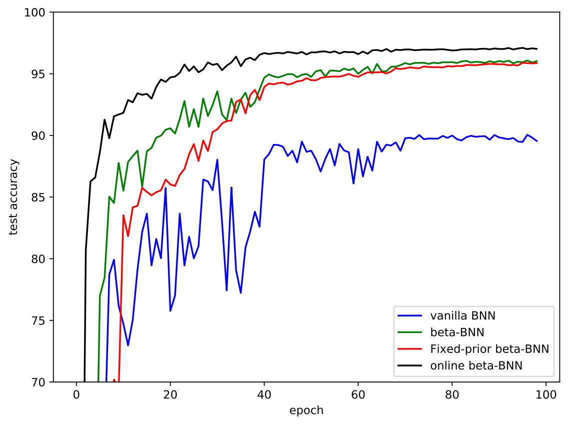
Baselines:

- Vanilla BNN: Blundell et al. (2015).
- Fixed-prior β-BNN: $\tilde{m}(w) \equiv \mathcal{N}(0, I)$.

Algorithm	β^*	Accuracy
Vanilla BNN	$\frac{1}{n}$	90.05
Fixed-prior β-BNN	10^{-10}	95.86
β-BNN	10^{-5}	96.08
Online β-BNN	10^{-3}	97.12

Experiments: colorful MNIST

Test accuracy over training epochs:



Is $I(w; S)$ a good regularizer?

A **bias-variance interpretation**:

$$
\begin{aligned}
I_f(w;S) &= \mathbb{E}_{p^*(S)} \int dw\ q(w|S)\ f\Big(\frac{q(w)}{q(w|S)}\Big) && \text{f-mutual-information} \\
&= \mathbb{E}_{p^*(S)} \int dw\ q(w|S)\Big(\frac{q(w)}{q(w|S)} - 1\Big)^2 && \text{if } f(t) = (t-1)^2 \\
&= \mathbb{E}_{p^*(S)} \mathbb{V}_{q(w|S)}\Big[\frac{q(w)}{q(w|S)}\Big] && \text{since } \mathbb{E}_{q(w|S)}\Big[\tfrac{q(w)}{q(w|S)}\Big] = 1
\end{aligned}
$$

A **PAC learning interpretation** by Xu and Raginsky (2017) if the loss is σ-subgaussian:

$$
\text{gen-error} = \mathbb{E}_{p^*(S)}\mathbb{E}_{q(w|S)}\Big[d(w,p^*) - d(w,S)\Big] \le \sqrt{\frac{2\sigma^2}{n} I(w;S)}.
$$

Moreover, $I(w; S)$ **is upper bounded by sharpness**:
Flat minimum $\Rightarrow$ small $I(w; S)$, but not vice versa.

The same objective can be used for meta learning

One plausible objective for meta learning is to learn a weight generator $q(w|S)$ such that it is a good approximation for all posteriors:

$$\begin{aligned}\min_q \quad & \mathbb{E}_{p^*(S)} D_{\mathrm{KL}}\Big(q(w|S) \| p(w|S)\Big) \\ & = -H(p^*(S)) + D_{\mathrm{KL}}\Big(q(w) \| p(w)\Big) \\ & \quad + \mathbb{E}_{p^*(S)} \mathbb{E}_{q(w|S)} \Big[d(w, S)\Big] + I(w; S).\end{aligned}$$

This is almost identical to the objective for supervised learning.

Variational information distillation for knowledge transfer

Deep learning is data-hungry

Issue: over-parameterized models are often trained with huge data.

- Medical applications is constrained by the number of patients of a particular disease.
- Semantic segmentation requires pixel-level annotation.

A potential **solution**: transfer learning.

- *Finetuning*: initialize with the weights of the source network.
- *Teacher-student knowledge transfer* by Ba and Caruana (2014); Hinton et al. (2015).

Teacher-student knowledge transfer: related work

There is no commonly agreed theory behind knowledge transfer.

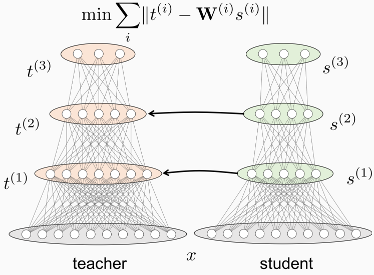


Figure 2: FitNet by Romero et al. (2014).

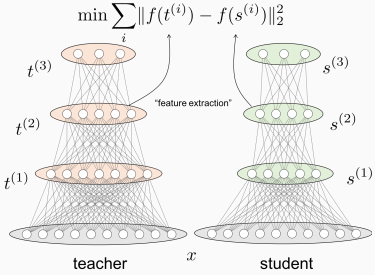
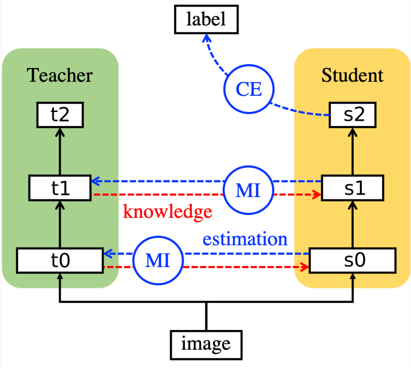


Figure 3: Attention transfer by Zagoruyko and Komodakis (2016).

Mutual information for knowledge transfer

Denote by $\mathbf{t}$ and $\mathbf{s}$ the activations of the teacher and the student respectively. Intuitively, $I(\mathbf{t};\mathbf{s})$ is maximized when $\mathbf{t} = \mathbf{s}$.



Variational information distillation (VID)

Knowledge transfer as a regularization:

$$\mathcal{L} = \mathcal{L}_{\text{task}} - \sum_{k=1}^{K} \lambda_k I(\mathbf{t}^{(k)}, \mathbf{s}^{(k)}),$$

Recall the variational characterization:

$$I(p; q) = \max_{\phi(\mathbf{t}|\mathbf{s})} \tilde{I}(p, q, \phi)$$

Instead of searching for all valid ϕ, we focus on diagonal Gaussians:

$$-\log \phi(\mathbf{t}|\mathbf{s}) = \sum_{n=1}^{N} \log \sigma_n + \frac{(t_n - \mu_n(\mathbf{s}))^2}{2\sigma_n^2} + \text{constant},$$

A related problem: channel capacity estimation

Noisy channel decoding theorem

Given a noisy channel from X to Y with transition $q(y|x)$, the channel capacity is given by

$$\begin{aligned} C &= \max_{p(x)\in\Delta} I(p,q) \\ &= \max_{p(x)\in\Delta} \max_{\phi(x|y)\in\Delta} \tilde{I}(p,q,\phi). \end{aligned}$$

Experiments: transfer from ImageNet to birds

Dataset: Caltech-UCSD Birds 200.

Networks: teacher (ResNet-34), student (ResNet-18).

data per class	≈29.95	20	10	5
Student	37.22	24.33	12.00	7.09
Finetuned	76.69	71.00	59.25	44.07
LwF	55.18	42.13	26.23	14.27
FitNet	66.63	56.63	46.68	31.04
AT	54.62	41.44	28.90	16.55
NST	55.01	41.87	23.76	15.63
VID	**73.25**	**67.20**	**56.86**	**46.21**

Experiments: transfer from ImageNet to indoor scenes

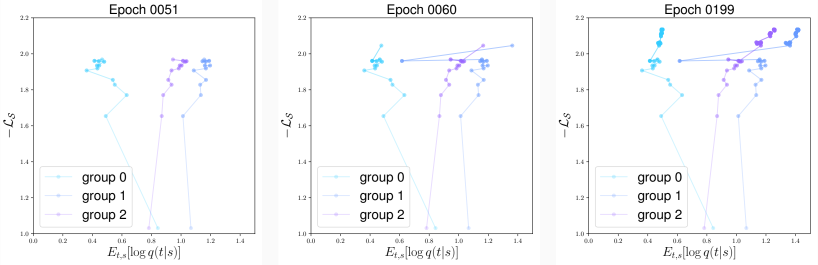
Dataset: MIT-67.

Networks: teacher (ResNet-34), student (VGG-9).

data per class	$\approx$80	50	25	10
Student	53.58	43.96	29.70	15.97
Finetuned	65.97	58.51	51.72	39.63
LwF	60.90	52.01	41.57	27.76
FitNet	70.90	64.70	54.48	40.82
AT	60.90	52.16	42.76	25.60
NST	55.60	46.04	35.22	21.64
VID	**72.01**	**67.01**	**59.33**	**45.90**

Relationship between task loss and VID

Two-stage transition: before epoch 51, only $-\mathcal{L}_{\mathcal{S}}$ increases significantly, $\mathbb{E}_{\mathbf{t},\mathbf{s}}[\log \phi(\mathbf{t}|\mathbf{s})]$ barely changes, so does $I(\mathbf{t};\mathbf{s})$; the first stage ends at epoch 60; at the second stage, $I(\mathbf{t};\mathbf{s})$ slowly increases, which also drives $-\mathcal{L}_{\mathcal{S}}$ increasing.



Experiments: transfer from CNNs to MLPs

Dataset: CIFAR-10.

Networks: teacher (WRN-40-2), student (MLP).

Network	MLP-4096	MLP-2048	MLP-1024
Student	70.60	70.78	70.90
KD	70.42	70.53	70.79
FitNet	76.02	74.08	72.91
VID	**85.18**	**83.47**	**78.57**
Urban et al. (2017)		74.32	
Lin et al. (2015)		78.62	

Questions?

References

Achille, A. and Soatto, S. (2017). Emergence of invariance and disentangling in deep representations. *arXiv preprint arXiv:1706.01350*.

Ba, J. and Caruana, R. (2014). Do deep nets really need to be deep? In *Advances in neural information processing systems*, pages 2654–2662.

Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. (2015). Weight uncertainty in neural networks. *arXiv preprint arXiv:1505.05424*.

Canziani, A., Paszke, A., and Culurciello, E. (2016). An analysis of deep neural network models for practical applications. *arXiv preprint arXiv:1605.07678*.

References ii

Cover, T. M. and Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons.

Denil, M., Shakibi, B., Dinh, L., De Freitas, N., et al. (2013). Predicting parameters in deep learning. In *Advances in neural information processing systems*, pages 2148–2156.

Dinh, L., Pascanu, R., Bengio, S., and Bengio, Y. (2017). Sharp minima can generalize for deep nets. *arXiv preprint arXiv:1703.04933*.

Graves, A. (2011). Practical variational inference for neural networks. In *Advances in neural information processing systems*, pages 2348–2356.

Haeffele, B. D. and Vidal, R. (2017). Global optimality in neural network training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7331–7339.

Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

Hinton, G. E. and Van Camp, D. (1993). Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the sixth annual conference on Computational learning theory*, pages 5–13. ACM.

Kawaguchi, K. (2016). Deep learning without poor local minima. In *Advances in Neural Information Processing Systems*, pages 586–594.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. (2016). On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*.

Li, H., Xu, Z., Taylor, G., and Goldstein, T. (2017). Visualizing the loss landscape of neural nets. *arXiv preprint arXiv:1712.09913*.

References iv

Lin, Z., Memisevic, R., and Konda, K. (2015). How far can we go without convolution: Improving fully-connected networks. *arXiv preprint arXiv:1511.02580*.

Lu, H. and Kawaguchi, K. (2017). Depth creates no bad local minima. *arXiv preprint arXiv:1702.08580*.

Neyshabur, B., Li, Z., Bhojanapalli, S., LeCun, Y., and Srebro, N. (2018). Towards understanding the role of over-parametrization in generalization of neural networks. *arXiv preprint arXiv:1805.12076*.

Romero, A., Ballas, N., Kahou, S. E., Chassang, A., Gatta, C., and Bengio, Y. (2014). Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*.

Soudry, D. and Hoffer, E. (2017). Exponentially vanishing sub-optimal local minima in multilayer neural networks. *arXiv preprint arXiv:1702.05777*.

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

Tishby, N. and Zaslavsky, N. (2015). Deep learning and the information bottleneck principle. In *Information Theory Workshop (ITW), 2015 IEEE*, pages 1–5. IEEE.

Urban, G., Geras, K. J., Kahou, S. E., Aslan, O., Wang, S., Mohamed, A., Philipose, M., Richardson, M., and Caruana, R. (2017). Do deep convolutional nets really need to be deep and convolutional? In *ICLR*.

Wu, C., Luo, J., and Lee, J. D. (2018). No spurious local minima in a two hidden unit relu network.

Wu, L., Zhu, Z., et al. (2017). Towards understanding generalization of deep learning: Perspective of loss landscapes. *arXiv preprint arXiv:1706.10239*.

Xu, A. and Raginsky, M. (2017). Information-theoretic analysis of generalization capability of learning algorithms. In *Advances in Neural Information Processing Systems*, pages 2524–2533.

Zagoruyko, S. and Komodakis, N. (2016). Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2016). Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*.