Градиентная маска и обобщения нейронной сети

Л. Цзян [1]

В рамках практического применения нейронных сетей количество параметров в сети намного больше, чем количество выборок в наборе данных, однако сеть по-прежнему имеет хорошие характеристики обобщения. Традиционно считается, что такие сверхпараметризованные и невыпуклые модели могут легко попадать в локальные минимумы при поиске оптимального решения и показывать плохую производительность обобщения, но на самом деле это не так. Хотя при некоторых условиях регуляризации возможно эффективно контролировать ошибку обобщения сети, по-прежнему трудно объяснить проблему обобщения для больших сетей. В данной статье мы определяем разницу между этапом переобучения и этапом изучения признаков путем количественной оценки влияния обновления одной выборки во время градиентного спуска на весь процесс обучения, выявив, что нейронные сети обычно меньше влияют на другие образцы на этапе переобучения. Кроме того, мы используем информационную матрицу Фишера для маскировки градиента, полученного в процессе обратного распространения, тем самым замедляя поведение нейронной сети при переобучении и улучшая производительность обобщения нейронной сети.

Ключевые слова: Нейронные сети, обобщение, переобучение, информация фишера.

1. Введение: Обобщающая способность нейронных сетей

Обобщение относится к способности модели правильно предсказывать данные, которые она никогда раньше не видела. Изучение способности сверх-параметризованных нейронных сетей к обобщению уже давно представляет интерес в рамках сферы машинного обучения, поскольку идет вразрез с положениями классической теории обучения.

С одной стороны, сверх-параметризованные нейронные сети способны сходиться к нулевым потерям на подавляющем большинстве обучающих наборов данных. Например, работа [1] доказывает, что нейронные

[1] *Цзян Лэй* — аспирант каф. математической теории интеллектуальных систем мех.-мат. ф-та МГУ, e-mail: kiwee@outlook.com

Jiang Lei — graduate student, Lomonosov Moscow State University, Faculty of Mechanics and Mathematics, Chair of Mathematical Theory of Intellectual Systems.

сети способны сходиться даже на случайных данных (зашумленные данные, случайные метки). Это указывает на то, что нейронная сеть достаточно приспособлена к обучающим данным. С другой стороны, нейронные сети, которые особенно подходят для тестового набора данных, могут по-прежнему обеспечивать хорошую производительность теста.

С точки зрения теории, некоторые исследования начинаются с мощности модели нейронных сетей: ошибка обобщения оценивается путем изучения взаимосвязи между мощностью модели и обучающими данными. Типичные методы: размерность Вапника — Червоненкиса (VC dimension) [2, 3], Радемахеровская сложность (Rademacher complexity) [4] и т.д. Другие придерживаются байесовской точки зрения, изучая и измеряя разницу между априорным и апостериорным распределениями модели, чтобы понять ошибку обобщения. Типичная теория — теория Байеса для ВПК-обучения (PAC-bayes) [5, 6, 7]. Напротив, теория PAC-bayes носит общий характер и поэтому применима к различным архитектурам нейронных сетей и наборам данных. Но есть и недостатки, а именно — зависимость от выбора априорного значения [8], в то время как необоснованный априорный выбор приводит к неверным границам обобщения.

На практике использовались различные технические методы для смягчения переобучающего поведения нейронных сетей, тем самым улучшая их способность к обобщению. Типичные технические методы включают:

1) Досрочное завершение: использование проверочного набора для отслеживания переобучения во время обучения и прерывание обучение, когда нейронная сеть входит в стадию доминирования переобучения.

2) Увеличение данных [9]: перед тем, как данные поступают в нейронную сеть, их всячески улучшают с целью увеличить сложность запоминания образцов нейронной сетью.

3) Использование случайности: например, Dropout [10] случайным образом отбрасывает нейроны во время прямого распространения сети.

4) Регуляризация веса [11].

5) Плоскостность: исследования [12, 13, 14] доказали, что гладкость вокруг локального минимума, к которому в итоге сходится нейронная сеть, влияет на производительность обобщения самой сети, а более плоский локальный минимум может улучшить производительность обобщения. Следовательно, явное вынуждение нейронной сети к поиску плоского локального минимума [15] в процессе оптимизации также является техническим методом, который может улучшить производительность обобщения.

2. Вклад

Наш вклад:

1. Мы измеряем влияние обновления одной выборки на другие в процессе градиентного спуска с помощью расстояния Кульбака — Лейблера (РКЛ; KL) и обнаруживаем, что влияние нейронной сети на процесс обучения на этапе выборки памяти (переобучения) в целом меньше, чем на этапе обучения признакам.

2. Мы обнаруживаем, что размер диагональных элементов информационной матрицы Фишера тесно связан с переобучением: во время обратного распространения мы используем информационную матрицу Фишера, чтобы замаскировать градиент веса, замедляя переобучение нейронной сети, тем самым улучшая эффективность обобщения сети.

3. Методология

В качестве примера возьмем задачу множественной классификации: задан набор данных $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, где x_i — данные, y_i — метка one-hot. Пусть $f(x; \theta)$ — это нейронная сеть. Мы рассматриваем нейронную сеть как вероятностную модель, а цель оптимизации состоит в том, чтобы максимизировать функцию правдоподобия:

$$\arg\max_{\theta} = \prod_{i=1}^{n} \prod_{j=1}^{K} f^j(x_i; \theta)^{y_i^j} \tag{1}$$

где K — количество категорий в задаче мультиклассификации, y_i^j — i-ый индекс класса образца, $f^j(x_i; \theta)$ — вероятность j-го класса на выходе нейронной сети для выборки x_i. Максимизация функции правдоподобия эквивалентна минимизации отрицательного логарифмического правдоподобия:

$$\arg\min_{\theta} = -\sum_{i=1}^{n} \sum_{j=1}^{K} y_i^j \log f^j(x_i; \theta) \tag{2}$$

Пусть функция потерь — это $\mathcal{L}(x, y, \theta)$, тогда $\sum_{j=1}^{K} y^j \log f^j(x; \theta) = \mathcal{L}(x, y, \theta)$ — функция потерь перекрестной энтропии.

Сначала мы исследуем, насколько «знания», полученные нейронной сетью из одной выборки, влияют на другие во время градиентного спуска. Пусть случайная выборка — это $(x_i, y_i) \in D$, тогда знания, полученные нейронной сетью из выборки x_i могут быть выражены как градиент — т.е. $g_i = \frac{\nabla \mathcal{L}((x_i, y_i), \theta)}{\nabla \theta}$. Тогда влияние на другие выборки можно измерить по следующей формуле:

$$KL(f(x; \theta) || f(x; \theta - \epsilon g_i)) \tag{3}$$

где KL - это Расстояние Кульбака — Лейблера.

Так как мы считаем нашу нейронную сеть как вероятностную модель с параметрами θ, вследствие этого, информационная матрица Фишер F_θ в θ определяется следующим образом:

$$F_\theta = \mathbb{E}_f[(\nabla \log f(x;\theta))(\nabla \log f(x;\theta))^T]$$

Теорема 1 ([26]). *При ϵ стремится к нулю, имеет место равенство*

$$KL(f(x;\theta)||f(x;\theta-\epsilon g_i)) = \frac{1}{2}\epsilon^2 (g_i)^T F_\theta g_i + O(\epsilon^3) \tag{4}$$

Доказательство. Для упрощения обозначений, пишем f_θ как $f(x;\theta)$, и $f_{\theta'}$ как $f(x;\theta-\epsilon g_i)$. Рассмотрим ряд Тейлора $KL(f(x;\theta)||f(x;\theta-\epsilon g_i))$ в точке θ:

$$\begin{aligned} KL(f_\theta||f_{\theta'}) = \quad & KL \quad (f_\theta||f_\theta) - \epsilon {g_i}^T \nabla_{\theta'} KL(f_\theta||f_{\theta'})|_{\theta'=\theta} \\ + \quad & \frac{1}{2}\epsilon^2 {g_i}^T [\nabla^2_{\theta'} KL(f_\theta||f_{\theta'})|_{\theta'=\theta}] g_i + O(\epsilon^3) \end{aligned}$$

Ясно, что первый член равен нулю, так как расстояние KL между одинаковыми распределениями равно нулю. Дальше, рассмотрим первую и вторую производную РКЛ в точке $\theta' = \theta$:

$$\nabla_{\theta'} KL(f_\theta||f_{\theta'})|_{\theta'=\theta} = -\mathbb{E}_{f_\theta}[\nabla_{\theta'} \log f_{\theta'}|_{\theta'=\theta}] = -\mathbb{E}_{f_\theta}[\frac{1}{f_\theta}(\nabla_\theta f_\theta|_{\theta'=\theta})] = 0 \tag{5}$$

и

$$\nabla^2_{\theta'} KL(f_\theta||f_{\theta'})|_{\theta'=\theta} = -\mathbb{E}_{f_\theta}[(\nabla^2_{\theta'} \log f_{\theta'}|_{\theta'=\theta})] = \mathbb{E}_{f_\theta}[H_{\log f_\theta}] = F_\theta \tag{6}$$

Из (3.5) и (3.6) получим: $KL(f(x;\theta)||f(x;\theta-\epsilon g_i)) = \frac{1}{2}\epsilon^2 (g_i)^T F_\theta g_i + O(\epsilon^3)$ □

На практике, учитывая набор данных, мы используем эмпирический метод Фишера(Empirical Fisher) для аппроксимации матрицы Фишера [16]:

$$F_\theta = \frac{1}{n}\sum_{i=1}^{n}[(\nabla_\theta \log f(x_i;\theta))(\nabla_\theta \log f(x_i;\theta))^T] \tag{7}$$

Из формулы 4 видно, что влияние информации, полученной нейронной сетью из одной выборки, на все обучение связано с информацией Фишера о текущих параметрах нейронной сети.

В работе [17, 18, 19] были выявлены две фазы обучения нейронной сети: фаза быстрого обучения (rapid learning) и фаза итеративной детализации (iterative refinement). На этапе быстрого обучения сеть сначала

изучает функции «общего назначения» (обучение представлений), которые оказывают существенное влияние на задачу классификации. На этом этапе значение функции потерь быстро уменьшается, а нейронная сеть может быстро и правильно классифицировать большинство простых образцов. По мере обучения нейронная сеть постепенно входит в стадию итеративной детализации: настраивает изученные функции, предоставляя возможность идентифицировать сложные образцы. Работа [20] подчеркивает, что на этапе итеративной детализации также увеличивается риск переобучения нейронной сети. Логично, что нейронная сеть на этапе переобучения запоминает сами образцы данных, а не изучает значимые признаки. Необходимо учитывать, что этап переобучения просто запоминает образцы и не помогает процессу обучения.

Пусть g_o — градиент, соответствующий образцу памяти, а g_f — градиент, соответствующий обучению признаков. По сравнению с обучением по признакам ожидаемое значение формулы 4 в процессе запоминания образцов должно быть меньше ожидаемого значения на этапе обучения по признакам. Имеем:

$$\mathbb{E}_{g_o \sim \text{overfitting}}[\epsilon^2 g_o^T F_\theta g_o] < \mathbb{E}_{g_f \sim \text{feature}}[\epsilon^2 g_f^T F_\theta g_f] \tag{8}$$

Мы проверяем достоверность утверждения, отслеживая формулу 8 во время обучения: эксперименты на наборе данных CIFAR-100 с использованием глубокой остаточной сверточной нейронной сети ResNet-18. Мы задействуем случайные метки, чтобы заставить нейронную сеть запомнить образцы данных. Результаты представлены на рисунке 1.

Во-первых, видим, что в процессе обучения значение формулы 4 постепенно уменьшается с увеличением времени обучения, отражая переход от этапа быстрого обучения к этапу итеративной детализации. Во-вторых, мы обнаружили, что при запоминании сетью случайных меток, влияние на другие образцы меньше, чем при обычном обучении.

В формуле 4 , поскольку масштаб информационной матрицы Фишера равен квадрату числа параметров сети, это делает неприемлемым расчет и хранение матрицы Фишера в современных больших нейронных сетях. Поэтому мы обращаемся к работе [21, 22]. Возьмем матрицу Фишера, используемую в формуле 4 , и используем ее диагональ в качестве аппроксимации, обозначив диагональные элементы F_θ как $a = (a_{11}, \ldots, a_{mm})$. Тогда формулу 4 можно записать следующим образом:

$$\epsilon^2 g_i^T F_\theta g_i = \sum_{j=1}^{m} a_{jj} g_{ij}^2 \tag{9}$$

Мы можем использовать формулу 8 для анализа того, сколько информации об образце памяти содержится в каждом компоненте гради-

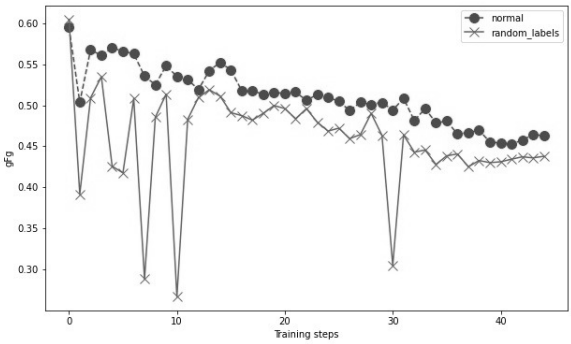


Рис. 1. Значения формулы 4 , соответствующие сетям, обученным с использованием случайных, а также обычных меток

ента g_i в процессе обратного распространения. Формулу 8 можно рассматривать как скалярное произведение вектора $g_i \odot g_i$ (здесь символ $\odot$ - произведение Адамара) и вектора, образованного диагональю матрицы Фишера. Ниже, компонент градиента g_i, соответствующий низкой информации Фишера, с большей вероятностью будет смещен в сторону образца памяти. То есть, мы можем предотвратить переобучение, замаскировав некоторые из этих обновлений параметров.

4. Эксперимент

В данном разделе мы сначала проверяем, может ли механизм маски замедлить явление переобучения, путем экспериментальной количественной оценки взаимосвязи между значением формулы 4 и потерями теста. После этого мы провели эксперименты с различными типами наборов данных и убедились, что механизм маски действительно может улучшить способность сети к обобщению.

4.1. Численный анализ

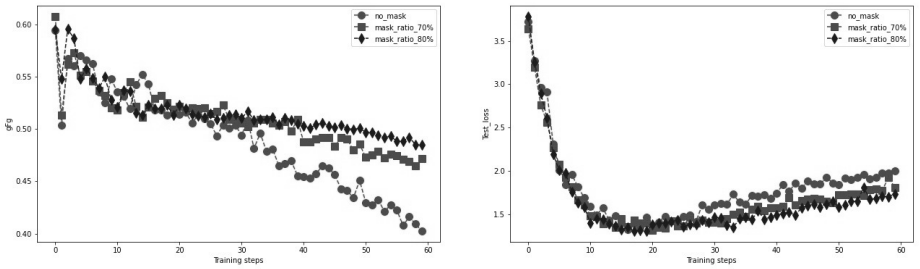
В предыдущем разделе мы обсуждали связь информационной матрицы Фишера с переобучением, а в этом экспериментируем с реальными наборами данных. Мы используем остаточную нейронную сеть ResNet-18 [23], обученную на наборе данных CIFAR-100 [24], который содержит 50 000 обучающих изображений и 10 000 тестовых изображений (всего 100 категорий). Мы маскируем некоторые компоненты в градиенте весов, которые называем механизмом маски градиента, и используем, чтобы наблюдать, может ли механизм маски смягчить явление переобучения.

Сначала сортируем диагональные элементы матрицы Фишера от больших к меньшим, а затем выбираем порог $k \in \{a_{11}, \ldots, a_{mm}\}$, согласно заданному проценту. Формула обновления для градиентного спуска выражается как:

$$\theta^{t+1} = \theta^t - \alpha(g^t \odot M^t), M_i^t = \begin{cases} 1, a_{ii}^t \geq k \\ 0, a_{ii}^t < k \end{cases}, \quad i = 1, \ldots, m \tag{10}$$

где g^t — это градиент весов в момент времени t, M^t — маска, a_{ii}^t — i-ый элемент диагональной матрицы Фишера в точке t.

В качестве порогов выбираем первые 70% и 80% значения диагональной матрицы Фишера, и обучаем сеть градиентным спуском по формуле 10. На рисунке 2 изображена связь между механизмом маски и переобучением. Мы видим, что все три сети в эксперименте перешли в переобучение примерно через 15 эпох. Но после использования механизма маски значение формулы 4 уменьшается медленнее (рисунка 2(a)), значительно замедляя скорость переобучения (рисунка 2(b)). Это доказывает, что механизм маски может эффективно подавлять память нейронной сети об образцах, заставляя нейронную сеть уделять больше внимания изучению признаков.



(a) (b)

Рис. 2. Слева (а) показаны значения формулы 4 , соответствующие трем сетям в процессе обучения; справа (б) показаны тестовые потери в каждый тренировочный момент

4.2. Механизм маски, улучшающий производительность обобщения

Мы проверяем способность механизма маски к обобщению на нескольких различных сетевых архитектурах, а также на разных наборах данных классификации. Для небольших задач (CIFAR-100) мы используем

остаточную сеть ResNet-18 с примерно 12 миллионами параметров, а для больших задач — ImageNet [25], задействовав остаточную нейронную сеть ResNet-50 [23] с 25 миллионами параметров. Из-за их различной сложности для небольших задач мы используем коэффициент маски 0,75, а для больших — 0,25. В целях экономии вычислительных ресурсов мы определяем диагональную матрицу Фишера после каждой эпохи (epoch) в процессе обучения и используем как основу для выбора маски для следующей эпохи обучения. Результаты, представленные в таблице 1, демонстрируют, что способность сети к обобщению улучшилась после использования механизма маски.

Модель	CIFAR-100	Imagenet
Обычная ResNet-18/50	78.2	75.5
ResNet-18/50 с маской	**80.26**	**76.01**

Таблица 1. Лучшая точность (%) ResNet-18 с/без маски на CIFAR-100 и ResNet-50 с/без маски на Imagenet

5. Заключение

В данной статье мы определяем разницу между этапом переобучения и этапом изучения признаков путем количественной оценки влияния обновления одной выборки во время градиентного спуска на весь процесс обучения, выявив, что нейронные сети обычно меньше влияют на другие образцы на этапе переобучения. Кроме того, мы предлагаем механизм градиентной маски с целью скрыть часть обновления весов переобучения, улучшая производительность обобщения нейронной сети.

Список литературы

[1] Zhang, Chiyuan and Bengio, Samy and Hardt, Moritz and Recht, Benjamin and Vinyals, Oriol, “Understanding deep learning (still) requires rethinking generalization”, *Communications of the ACM*, **64** (2019), 107–115.

[2] Sontag, Eduardo D and others, “VC dimension of neural networks”, *NATO ASI Series F Computer and Systems Sciences*, **168** (1998), 69–96.

[3] Bartlett, Peter L and Harvey, Nick and Liaw, Christopher and Mehrabian, Abbas, “Nearly-tight VC-dimension and pseudodimension

bounds for piecewise linear neural networks", *The Journal of Machine Learning Research*, **20** (2019), 2285–2301.

[4] Bartlett, Peter L and Mendelson, Shahar, "Rademacher and Gaussian complexities: Risk bounds and structural results", *Journal of Machine Learning Research*, **3** (2002), 463–482.

[5] Shawe-Taylor, John and Williamson, Robert C, "A PAC analysis of a Bayesian estimator", *Proceedings of the tenth annual conference on Computational learning theory*, 1997, 2–9.

[6] McAllester, David A, "Some pac-bayesian theorems", *Machine Learning*, **37** (1999), 355–363.

[7] Dziugaite, Gintare Karolina and Roy, Daniel M, "Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data", *arXiv preprint arXiv:1703.11008*, 2017.

[8] Dziugaite, Gintare Karolina and Hsu, Kyle and Gharbieh, Waseem and Arpino, Gabriel and Roy, Daniel, "On the role of data in PAC-Bayes bounds", *International Conference on Artificial Intelligence and Statistics*, 2021, 604–612.

[9] Shorten, Connor and Khoshgoftaar, Taghi M, "A survey on image data augmentation for deep learning", *Journal of big data*, **6** (2019), 1–48.

[10] Hinton, Geoffrey E and Srivastava, Nitish and Krizhevsky, Alex and Sutskever, Ilya and Salakhutdinov, Ruslan R, "Improving neural networks by preventing co-adaptation of feature detectors", *arXiv preprint arXiv:1207.0580*, 2012.

[11] Xie, Zeke and Sato, Issei and Sugiyama, Masashi, "Understanding and scheduling weight decay", *arXiv preprint arXiv:2011.11152*, 2020.

[12] Keskar, Nitish Shirish and Mudigere, Dheevatsa and Nocedal, Jorge and Smelyanskiy, Mikhail and Tang, Ping Tak Peter, "On large-batch training for deep learning: Generalization gap and sharp minima", *arXiv preprint arXiv:1609.04836*, 2016.

[13] Neyshabur, Behnam and Bhojanapalli, Srinadh and McAllester, David and Srebro, Nati, "Exploring generalization in deep learning", *Advances in neural information processing systems*, **30** (2017).

[14] Dinh, Laurent and Pascanu, Razvan and Bengio, Samy and Bengio, Yoshua, "Sharp minima can generalize for deep nets", *International Conference on Machine Learning*, 2017, 1019–1028.

[15] Foret, Pierre and Kleiner, Ariel and Mobahi, Hossein and Neyshabur, Behnam, "Sharpness-aware minimization for efficiently improving generalization", *arXiv preprint arXiv:2010.01412*, 2020.

[16] Park, Hyeyoung and Amari, S-I and Fukumizu, Kenji, "Adaptive natural gradient learning algorithms for various stochastic models", *Neural Networks*, **13** (200), 755–764.

[17] Jastrzębski, Stanisław and Arpit, Devansh and Ballas, Nicolas and Verma, Vikas and Che, Tong and Bengio, Yoshua, "Residual connections encourage iterative inference", *arXiv preprint arXiv:1710.04773*, 2017.

[18] Greff, Klaus and Srivastava, Rupesh K and Schmidhuber, Jürgen, "Highway and residual networks learn unrolled iterative estimation", *arXiv preprint arXiv:1612.07771*, 2016.

[19] Raghu, Maithra and Gilmer, Justin and Yosinski, Jason and Sohl-Dickstein, Jascha, "Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability", *Advances in neural information processing systems*, **30** (2017).

[20] Arpit, Devansh and Jastrzębski, Stanisław and Ballas, Nicolas and Krueger, David and Bengio, Emmanuel and Kanwal, Maxinder S and Maharaj, Tegan and Fischer, Asja and Courville, Aaron and Bengio, Yoshua and others, "A closer look at memorization in deep networks", *International conference on machine learning*, 2017, 233–242.

[21] Martens, James and others, "Deep learning via hessian-free optimization", *ICML*, **27** (2010), 735–742.

[22] Chapelle, Olivier and Erhan, Dumitru and others, "Improved preconditioner for hessian free optimization", *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, **201** (2011).

[23] He, Kaiming and Zhang, Xiangyu and Ren, Shaoqing and Sun, Jian, "Deep residual learning for image recognition", *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, 770–778.

[24] Krizhevsky, Alex and Hinton, Geoffrey and others, "Learning multiple layers of features from tiny images", 2009.

[25] Deng, Jia and Dong, Wei and Socher, Richard and Li, Li-Jia and Li, Kai and Fei-Fei, Li, "Imagenet: A large-scale hierarchical image database", *2009 IEEE conference on computer vision and pattern recognition*, 2009, 248–255.

[26] Guo, Dongning, "Relative entropy and score function: New information-estimation relationships through arbitrary additive perturbation", *2009 IEEE International Symposium on Information Theory*, 2009, 814–818.

Gradient mask and neural network generalization
Jiang Lei

Within the practical application of neural networks, the number of parameters in the network is much larger than the number of samples in the dataset, however, the network still has good generalization characteristics. Traditionally considered that such over-parameterized and non-convex models can easily fall into local minima while searching for the optimal solution and show low generalization performance, but in fact it is not. Although under some regularization conditions it is possible to effectively control the network generalization error, it is still difficult to explain the generalization problem for large networks. In our work, we determine the difference between the overfitting step and the feature learning step by quantifying the impact of updating one sample during gradient descent on the entire training process, revealing that neural networks generally have less impact on other samples during the overfitting step. In addition, we use the Fisher information matrix to mask the gradient produced by the backpropagation process, thereby slowing down the neural network's overfitting behavior and improving the neural network's generalization performance.

Keywords: Neural Networks, Generalization, Overfitting, Fisher Information.

References

[1] Zhang, Chiyuan and Bengio, Samy and Hardt, Moritz and Recht, Benjamin and Vinyals, Oriol, "Understanding deep learning (still) requires rethinking generalization", *Communications of the ACM*, **64** (2019), 107–115.

[2] Sontag, Eduardo D and others, "VC dimension of neural networks", *NATO ASI Series F Computer and Systems Sciences*, **168** (1998), 69–96.

[3] Bartlett, Peter L and Harvey, Nick and Liaw, Christopher and Mehrabian, Abbas, "Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks", *The Journal of Machine Learning Research*, **20** (2019), 2285–2301.

[4] Bartlett, Peter L and Mendelson, Shahar, "Rademacher and Gaussian complexities: Risk bounds and structural results", *Journal of Machine Learning Research*, **3** (2002), 463–482.

[5] Shawe-Taylor, John and Williamson, Robert C, "A PAC analysis of a Bayesian estimator", *Proceedings of the tenth annual conference on Computational learning theory*, 1997, 2–9.

[6] McAllester, David A, "Some pac-bayesian theorems", *Machine Learning*, **37** (1999), 355–363.

[7] Dziugaite, Gintare Karolina and Roy, Daniel M, "Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data", *arXiv preprint arXiv:1703.11008*, 2017.

[8] Dziugaite, Gintare Karolina and Hsu, Kyle and Gharbieh, Waseem and Arpino, Gabriel and Roy, Daniel, "On the role of data in PAC-Bayes bounds", *International Conference on Artificial Intelligence and Statistics*, 2021, 604–612.

[9] Shorten, Connor and Khoshgoftaar, Taghi M, "A survey on image data augmentation for deep learning", *Journal of big data*, **6** (2019), 1–48.

[10] Hinton, Geoffrey E and Srivastava, Nitish and Krizhevsky, Alex and Sutskever, Ilya and Salakhutdinov, Ruslan R, "Improving neural networks by preventing co-adaptation of feature detectors", *arXiv preprint arXiv:1207.0580*, 2012.

[11] Xie, Zeke and Sato, Issei and Sugiyama, Masashi, "Understanding and scheduling weight decay", *arXiv preprint arXiv:2011.11152*, 2020.

[12] Keskar, Nitish Shirish and Mudigere, Dheevatsa and Nocedal, Jorge and Smelyanskiy, Mikhail and Tang, Ping Tak Peter, "On large-batch training for deep learning: Generalization gap and sharp minima", *arXiv preprint arXiv:1609.04836*, 2016.

[13] Neyshabur, Behnam and Bhojanapalli, Srinadh and McAllester, David and Srebro, Nati, "Exploring generalization in deep learning", *Advances in neural information processing systems*, **30** (2017).

[14] Dinh, Laurent and Pascanu, Razvan and Bengio, Samy and Bengio, Yoshua, "Sharp minima can generalize for deep nets", *International Conference on Machine Learning*, 2017, 1019–1028.

[15] Foret, Pierre and Kleiner, Ariel and Mobahi, Hossein and Neyshabur, Behnam, "Sharpness-aware minimization for efficiently improving generalization", *arXiv preprint arXiv:2010.01412*, 2020.

[16] Park, Hyeyoung and Amari, S-I and Fukumizu, Kenji, "Adaptive natural gradient learning algorithms for various stochastic models", *Neural Networks*, **13** (200), 755–764.

[17] Jastrzębski, Stanisław and Arpit, Devansh and Ballas, Nicolas and Verma, Vikas and Che, Tong and Bengio, Yoshua, "Residual connections encourage iterative inference", *arXiv preprint arXiv:1710.04773*, 2017.

[18] Greff, Klaus and Srivastava, Rupesh K and Schmidhuber, Jürgen, "Highway and residual networks learn unrolled iterative estimation", *arXiv preprint arXiv:1612.07771*, 2016.

[19] Raghu, Maithra and Gilmer, Justin and Yosinski, Jason and Sohl-Dickstein, Jascha, "Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability", *Advances in neural information processing systems*, **30** (2017).

[20] Arpit, Devansh and Jastrzębski, Stanisław and Ballas, Nicolas and Krueger, David and Bengio, Emmanuel and Kanwal, Maxinder S and Maharaj, Tegan and Fischer, Asja and Courville, Aaron and Bengio, Yoshua and others, "A closer look at memorization in deep networks", *International conference on machine learning*, 2017, 233–242.

[21] Martens, James and others, "Deep learning via hessian-free optimization", *ICML*, **27** (2010), 735–742.

[22] Chapelle, Olivier and Erhan, Dumitru and others, "Improved preconditioner for hessian free optimization", *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, **201** (2011).

[23] He, Kaiming and Zhang, Xiangyu and Ren, Shaoqing and Sun, Jian, "Deep residual learning for image recognition", *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, 770–778.

[24] Krizhevsky, Alex and Hinton, Geoffrey and others, "Learning multiple layers of features from tiny images", 2009.

[25] Deng, Jia and Dong, Wei and Socher, Richard and Li, Li-Jia and Li, Kai and Fei-Fei, Li, "Imagenet: A large-scale hierarchical image database", *2009 IEEE conference on computer vision and pattern recognition*, 2009, 248–255.

[26] Guo, Dongning, "Relative entropy and score function: New information-estimation relationships through arbitrary additive perturbation", *2009 IEEE International Symposium on Information Theory*, 2009, 814–818.