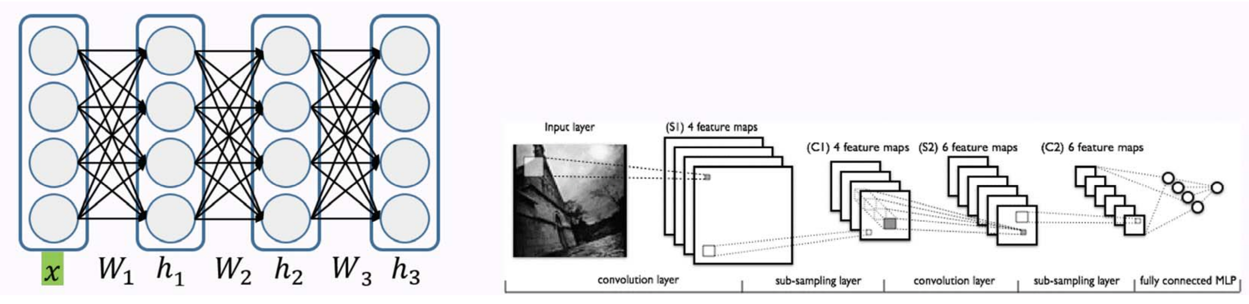
深層学習理論チーム　　鈴木大慈

Deep Learning Theory Team

Taiji Suzuki

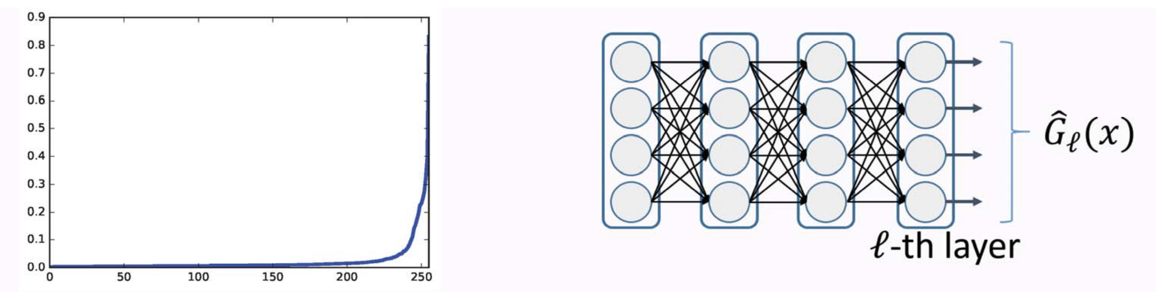
深層学習



深層学習：高い性能 → 理論的解明が世界的な問題に

- パラメータ数＞サンプルサイズでも過学習しない理由
- 構造を自動決定する指針

解析のポイント



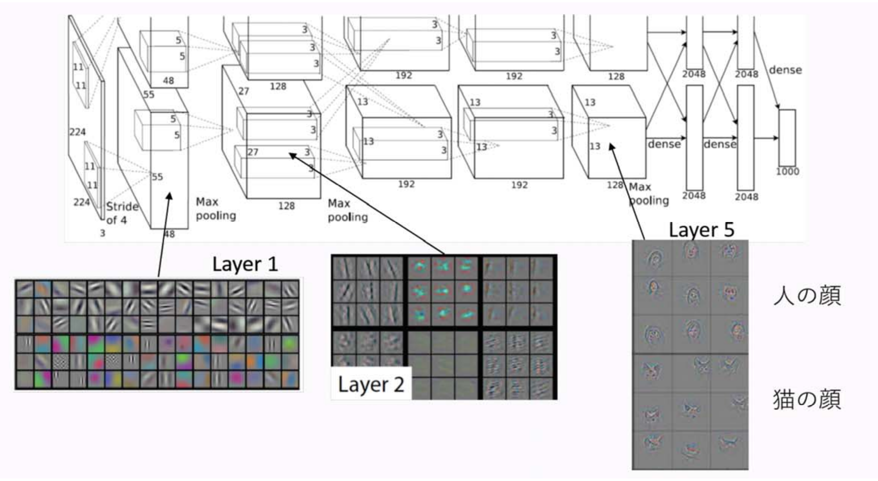
VGG-13の第九層の出力に関する分散共分散行列の固有値の分布 (CIFAR-10). → 沢山の小さい固有値

深層ニューラルネットの積分表現

$\hat{f}(x) = W_L\eta(W_{L-1}\eta(W_{L-2}\cdots \eta(W_1x+b_1)+b_2)\dots))$　　$f^{\rm o}(x) = {\rm g}_L\circ {\rm g}_{L-1}\circ\cdots\circ {\rm g}_1(x)$

$${\rm g}_\ell[F](\tau,x) = \int h^{\rm o}_\ell(\tau,\tau')\eta(F(\tau',x)+b^{\rm o}_\ell(\tau'))\mathrm{d}Q_\ell(\tau')$$

η: 非線形活性化関数（例：ReLU $\eta(x)=\max\{x,0\}$）



- ${\rm g}_\ell[F_{\ell-1}](\tau,x)$は第$\ell$層の特徴量$\tau$を入力$x$がどれだけ含有しているかを表示.
- 積分表現は万有近似能力を示す際に重要(Sonoda and Murata, 2015).

各層の再生核ヒルベルト空間

$F_\ell(\tau,x) = ({\rm g}_\ell\circ{\rm g}_{\ell-1}\circ\cdots\circ{\rm g}_1(x))(\tau)$: 第$\ell$層からの出力.

$${\rm k}_\ell(x,x') = \int \eta(F_{\ell-1}(\tau,x))\eta(F_{\ell-1}(\tau,x'))\mathrm{d}Q_\ell(\tau)$$

$(\mu_j^{(\ell)})_{j=1}^\infty$: カーネルの固有値; ${\rm k}_\ell(x,x') = \sum_{j=1}^\infty \mu_j^{(\ell)}\phi_j^{(\ell)}(x)\phi_j^{(\ell)}(x')$.

自由度

$$N_\ell(\lambda) := \sum_{j=1}^\infty \frac{\mu_j^{(\ell)}}{\mu_j^{(\ell)}+\lambda}$$

汎化誤差の上界

仮定

- 活性化関数はスケール不変: $\eta(au) = a\eta(u)$.　(e.g., ReLU)
- ηは1-Lipschitz連続.
- $\|h^{\rm o}_\ell(\tau,\cdot)\|_{L_2(Q_\ell)} \le R$ $(\forall\tau\in T_\ell)$, $|b^{\rm o}_\ell(\tau)|\le R_b$ $(\forall\tau\in T_\ell)$.
- ある正の実数$\lambda_\ell>0$に対して，横幅m_ℓは以下を満たす:

$$m_\ell \gtrsim N_\ell(\lambda_\ell)\log(N_\ell(\lambda_\ell)).$$

定理 (汎化誤差の上界)

(有限次元近似誤差)　$\hat{\delta}_1 = \sum_{\ell=2}^L 2\sqrt{c_1^{L-\ell-1}}R^{L-\ell}\sqrt{\lambda_\ell}$,

(有限次元モデル内推定誤差)　$\hat{\delta}_2 = R^L\sqrt{\frac{\sum_{\ell=1}^L m_{\ell+1}m_\ell}{n}}$.

- 経験誤差最小化: 高い確率で以下が成り立つ:
$$\|\hat{f}-f^{\rm o}\|_{L_2(\Pi)}^2 \le C(\hat{\delta}_1^2+\hat{\delta}_2^2).$$
- ベイズ推定量: The posterior tail is bounded as
$$\mathrm{E}[\Pi(f:\|f-f^{\rm o}\|_{L_2(\Pi)}\ge r(\hat{\delta}_1+\hat{\delta}_2)|D_n)] \le \exp(-C_1r^2)$$

解析の意味

意義: 有限近似によるバイアス-バリアンスのトレードオフを導出.

- 固有値が速く減衰 → より小さいネットワークで近似できる→高い汎化性能

汎化誤差とネットワークの構造決定

大きめの学習済みネットワーク$\hat{f}$を$f^\#$に圧縮する：

$$\hat{f}\to f^\#$$

$f^\#$の第ℓ層の横幅を$m_\ell^\sharp$とする.

$$\hat{\delta}_{2,n}^2(m^\sharp) = \frac{1}{n}\sum_{\ell=1}^L m_\ell^\sharp m_{\ell+1}^\sharp \log_+(n).$$

問題： 横幅$m_\ell^\sharp$はどれくらいに設定すれば良いか？

圧縮と汎化誤差の関係

$\hat{N}_\ell(\lambda_\ell) = \sum_j \frac{\hat{\mu}_j^{(\ell)}}{\hat{\mu}_j^{(\ell)}+\lambda}$: 経験分布から決まる自由度.

学習したネットワーク$\hat{f}$を第ℓ-層を以下の幅まで圧縮する:

$$m_\ell^\sharp \ge 5\hat{N}_\ell(\lambda_\ell)\log(80\hat{N}_\ell(\lambda_\ell)).$$

すると圧縮したネットワークは，以下の汎化性能を示す：確率$1-5e^{-t}$で

$$\|f^\#-f^{\rm o}\|_{L_2(\Pi)}^2 \le C\left\{\hat{\delta}_{1,n}^2+(\sigma^2+\hat{R}_\infty^2)\hat{\delta}_{2,n}^2(m^\sharp)+R_{n,t}\right\}$$

ただし，$R_{n,t} = \frac{(\hat{R}_\infty^2+\sigma^2)}{n}\left(\log\log\left[\frac{\sqrt{n}}{\frac{\sigma}{\hat{R}_\infty}\wedge 1}\right]+t+\sum_{\ell=2}^L\log(m_\ell)\right)$.

さらに，追加の条件が満たされていれば，元のネットワーク$\hat{f}$も同様の汎化誤差を達成する.

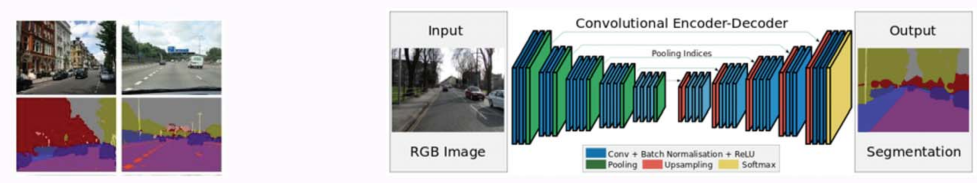
- 固有値が速く落ちるなら，小さいネットワークに圧縮できる.
- $\hat{f}$は圧縮後のネットワークのサイズのみに依存した汎化誤差を示す.

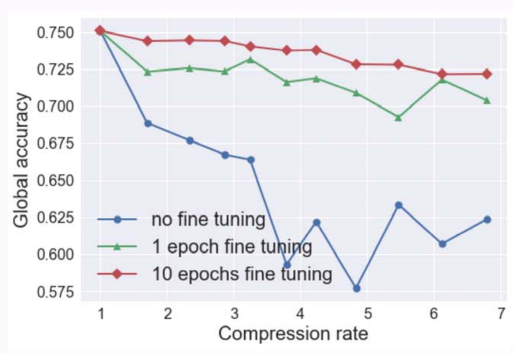
固有値の減衰が速いネットワークは，小さなネットワークに圧縮でき，かつ汎化性能が高い.

構造の自動決定とモデル圧縮への応用

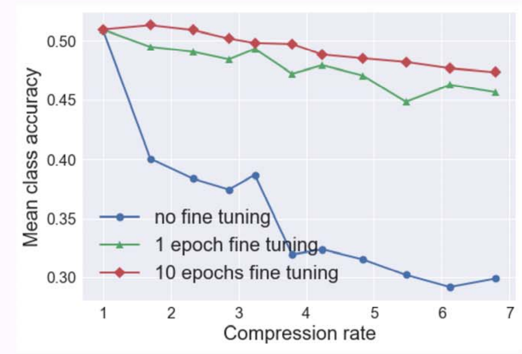
各レイヤーの横幅をデータから適応的に決定.

メモリの効率化＋予測値計算の高速化



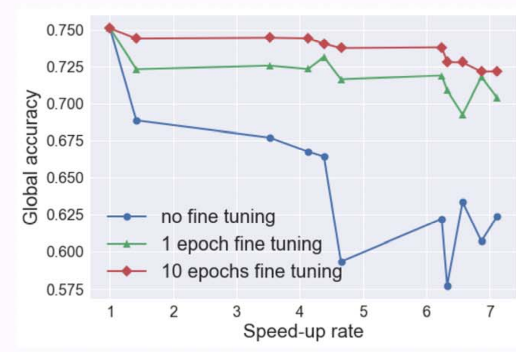


(a) Global accuracy vs. parameter compression rate.



(b) Mean class accuracy vs. parameter compression rate.

精度向上させせつつ処理速度4倍



(c) Global accuracy vs. computational speed up.

(d) Mean class accuracy vs. computation speed up.

圧縮アルゴリズム

$$\min\ |J|\quad (J\subset\{1,\dots,m_\ell\})$$
$$\text{s.t.}\quad \underbrace{\frac{\mathrm{Tr}[\hat{\Sigma}_{F,J}\hat{\Sigma}_{J,J}^{-1}\hat{\Sigma}_{J,F}]}{\mathrm{Tr}[\hat{\Sigma}_{F,F}]}}_{\text{Information ratio}} \ge \alpha$$

ただし，$\hat{\Sigma}$は$\hat{f}$の中間層の経験分布による分散共分散行列，$F=\{1,\dots,m_\ell\}$.

上記の汎化性能を満たす圧縮方法を近似することで導出.

ImageNetデータセット

比較対象：APoZ-2 Hu et al. (2016), SqueezeNet Iandola et al. (2016), and ThiNet Luo et al. (2017). Our method is indicated as "Ours-(type)."

Model	Top-1	Top-5	# Param.	FLOPs
Original VGG	68.34%	88.44%	138.34M	30.94B
APoZ-2	70.15%	89.69%	51.24M	30.94B
SqueezeNet	57.67%	80.39%	1.24M	1.72B
ThiNet-Conv	69.80%	89.53%	131.44M	9.58B
ThiNet-GAP	67.34%	87.92%	8.32M	9.34B
ThiNet-Tiny	59.34%	81.97%	1.32M	2.01B
Ours-Conv	69.61%	89.34%	113.92M	9.71B
Ours-Conv-FC	68.66%	88.90%	45.77M	9.58B
Ours-GAP	66.49%	87.62%	7.97M	9.54B
Ours-Tiny	60.10%	82.89%	2.31M	2.07B

提案手法

サイズが約三分の一でも精度向上

Taiji Suzuki: Fast generalization error bound of deep learning from a kernel perspective. AISTATS2018.