第3回 機械学習勉強会

森研M1　芳賀夢久

前回までのおさらい

- 誤差逆伝播法
  - 順伝搬型ネットワークの学習において，重みとバイアスに関する誤差関数の微分を効率よく計算する手法
  - 勾配消失問題
    - 活性化関数によっては(線形計算など)各層を伝搬するうちに勾配が発散したり0に消失する場合がある
    - 重みの更新がうまくできなくなり，学習が困難になる

前回までのおさらい

- 順伝搬型ネットワークの教師あり学習では，学習開始時の重みは通常ランダムに初期化する

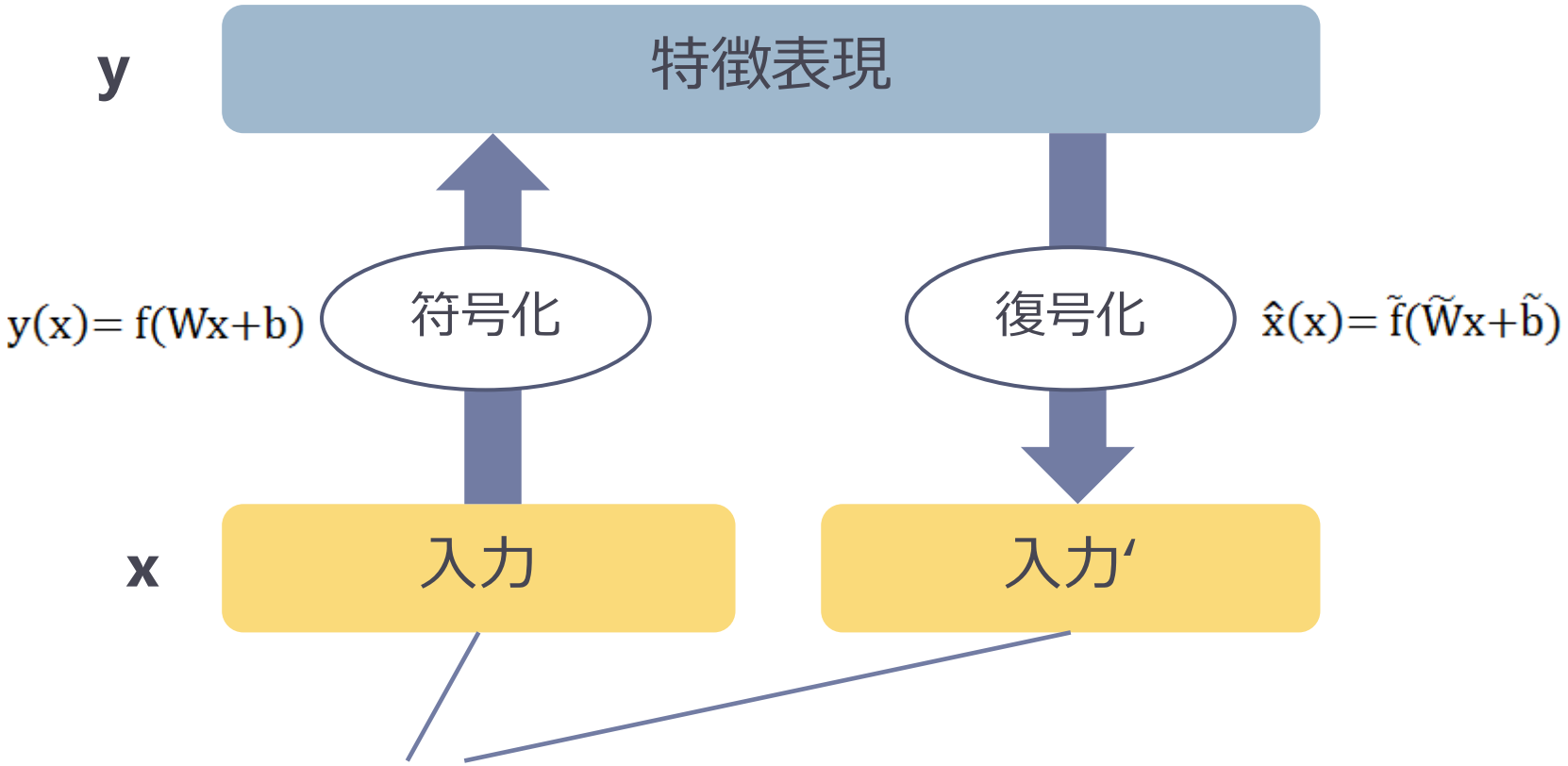
- 事前学習によって初期値を最も良い方法で選定すれば学習がうまくいき，勾配消失を回避できる

- 自己符号化器を使用

第5章　自己符号化器

- 概要
- ネットワークの設計
- 自己符号化器の働き
- スパース正規化
- データの白色化
- ディープネットの事前学習
- その他の自己符号化器

自己符号化器の概要（p.55-57）



元の入力が忠実に再現できるような符号化の方法を定めるのが狙い

活性化関数と誤差関数について（p.57）

- 中間層 f は自由に決めることができ，通常，非線形関数を選ぶ

- 出力層 ~f は目標出力が入力した x 自身になるように入力データの種類に応じて選ぶ

  - x の各成分が実数値でその値に制約がないとき

    → ~f：恒等写像，誤差関数：二乗誤差の総和

  - x の成分が0と1の二値をとるとき

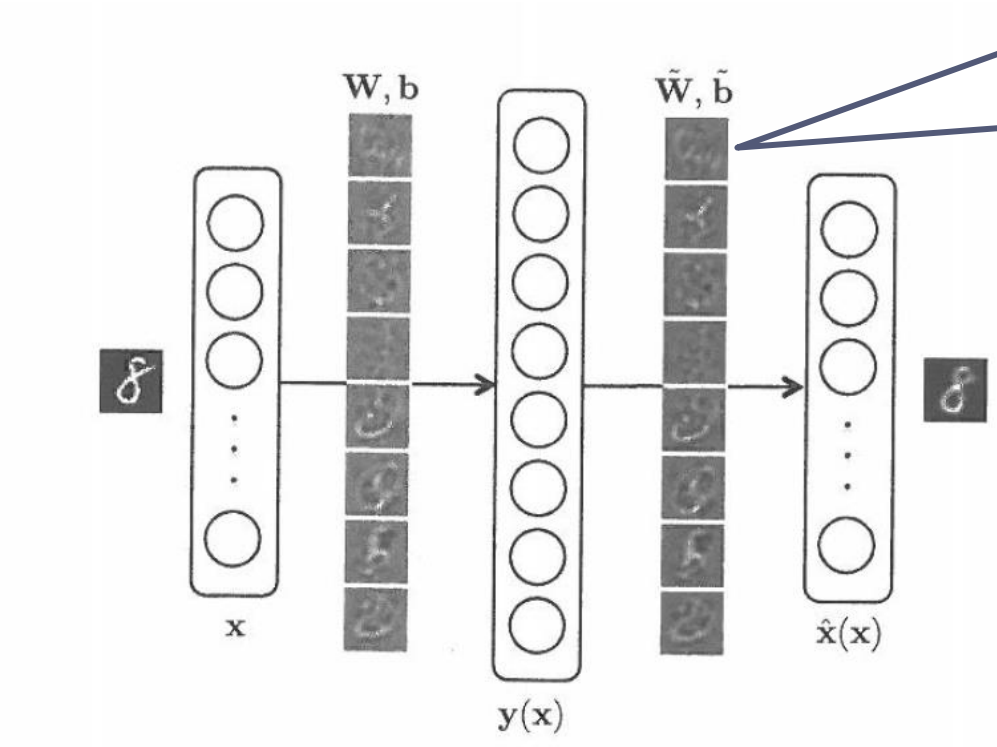
    → ~f：ロジスティック関数，誤差関数：交差エントロピー

基本的な流れ

1. 入力xを符号化器で符号化する
2. １の出力yを復号化器に入力させて入力を再現する
3. 誤差関数を用いて２の結果と元の入力を比較する
4. ３の誤差が最小となるようにパラメータ(W,b)を更新
   - 誤差をパラメータで微分し，それをゼロとなるようなパラメータ値を求める（確率的勾配法）

データを表す特徴の学習（p.58-60）

- 学習を通じてサンプル x の別な表現である y を得ることが目標



各ユニットの重みを画像化したもの
（手書き数字の成分が見て取れる）

図5.2　自己符号化器の動作の様子

主成分分析との関係（p.60-62）

- 入力層のユニット数がDx，中間層のユニット数がDyのとき
- Dy≧Dxのとき，x→ ^xが恒等写像になり得る
  - 誤差関数E(x)の値がいつも０になる

- 活性化関数が線形関数の場合，意味のある結果を得るには Dy≦Dxであることが必須条件

主成分分析との関係（p.60-62）

- Dy≦Dxのとき，E(x)を最小にするWおよび~Wは訓練データの**主成分分析**で得られるものと実質的に同じ

$$\min_{\boldsymbol{\xi},\boldsymbol{\Gamma}} \sum_{n=1}^{N} \left\| (\mathbf{x}_n - \boldsymbol{\xi}) - \boldsymbol{\Gamma}^{\top}\boldsymbol{\Gamma}(\mathbf{x}_n - \boldsymbol{\xi}) \right\|^2$$

誤差関数

$$E(\mathbf{w}) = \sum_{n=1}^{N} \|\mathbf{x}_n - \hat{\mathbf{x}}(\mathbf{x}_n)\|^2$$

スパース正則化（p.62-63）

- 入力データの不要な情報をそぎ落とし，その本質だけを取り出す

- 入力層のユニット数よりも中間層のユニット数の方が多い(Dy≧Dx)場合でも自己符号化器で意味のある表現を実現できる

  →スパース自己符号化器

スパース正則化（p.61-63）

- **スパース(sparse)**な表現とは
  - どの時点でも(どのような入力に対しても)，ほとんどの特徴の値は0であるような表現

- 基本的な考え方
  - 個々の訓練サンプルをなるべく少ない数の中間層のユニットを使って再現できるようにパラメータを設定する

スパース正則化（p.61-63）

- 誤差関数E(x)を拡張する
  - 活性化の平均値を小さくするように制約し，各サンプルを表現するのに使われる中間層のユニットの数が少なくなるようにする

正則化項

$$\tilde{E}(\mathbf{w}) \equiv E(\mathbf{w}) + \beta \sum_{j=1}^{D_y} \mathrm{KL}(\rho \| \hat{\rho}_j)$$

ρ：活性度の目標値
β：目標のバランスを変えるパラメータ

$$\hat{\rho}_j = \frac{1}{N} \sum_{n=1}^{N} y_j(\mathbf{x}_n)$$

中間層のユニットjの
平均活性度の推定値

$$\mathrm{KL}(\rho \| \hat{\rho}_j) = \rho \log\left(\frac{\rho}{\hat{\rho}_j}\right) + (1-\rho)\log\left(\frac{1-\rho}{1-\hat{\rho}_j}\right)$$

ρと^ρjの近さを表す
(図5.5 参照)

最適化（p.64-66）

- 拡張した誤差関数~E(w)を勾配効果法で最小化する際，スパース正則化項を考慮する必要がある

スパース正則化項のみの微分を考えると

正則化項

$$\frac{\partial}{\partial u_j^{(l)}}\left(\beta\sum_{j=1}^{D_l}\mathrm{KL}(\rho\|\hat{\rho}_j)\right)=\beta\frac{\partial}{\partial\hat{\rho}_j}\mathrm{KL}(\rho\|\hat{\rho}_j)\cdot\frac{\partial\hat{\rho}_j}{\partial u_j^{(l)}}$$

$$=\left(-\frac{\rho}{\hat{\rho}_j}+\frac{1-\rho}{1-\hat{\rho}_j}\right)$$

$$=f'(u_j^{(l)})$$

デルタの式 $\delta_j^{(l)}=\sum_k\delta_k^{(l+1)}\left(w_{kj}^{(l+1)}f'(u_j^{(l)})\right)$ より（p.47）

$$\delta_j^{(l)}=\left\{\sum_k\delta_k^{(l+1)}w_{kj}^{(l+1)}+\left(-\frac{\rho}{\hat{\rho}_j}+\frac{1-\rho}{1-\hat{\rho}_j}\right)\right\}f'(u_j^{(l)}) \qquad (5.4)$$

最適化（p.64-66）

- 式(5.4)でデルタを計算する際，全サンプルについてユニットの平均活性度^ρjを計算する必要がある

- ミニバッチ(p.26)を用いる場合，全サンプルの順伝搬計算を行うのはとても非効率

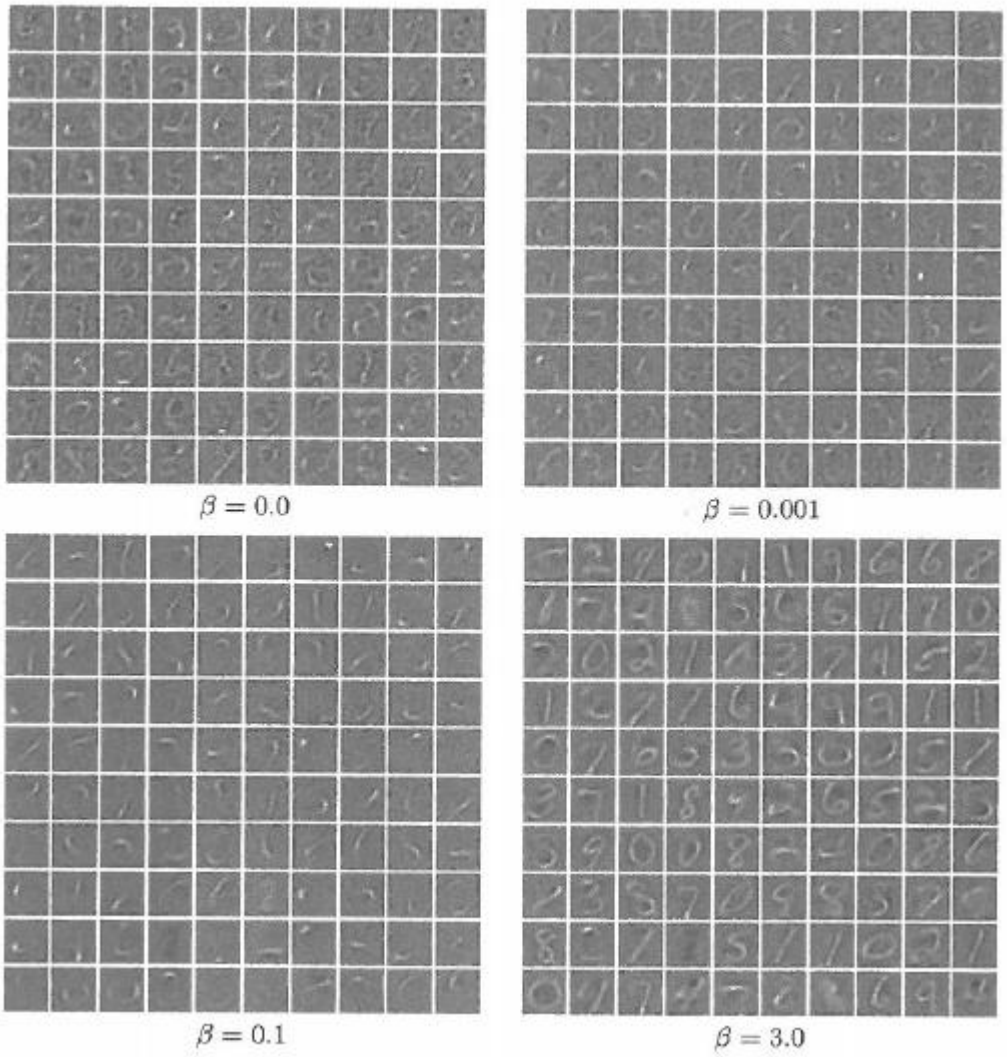
- ミニバッチが含むサンプル集合についてのみ平均活性度を求める

$$\hat{\rho}_j^{(t)} = \lambda \hat{\rho}_j^{(t-1)} + (1-\lambda)\hat{\rho}_j$$

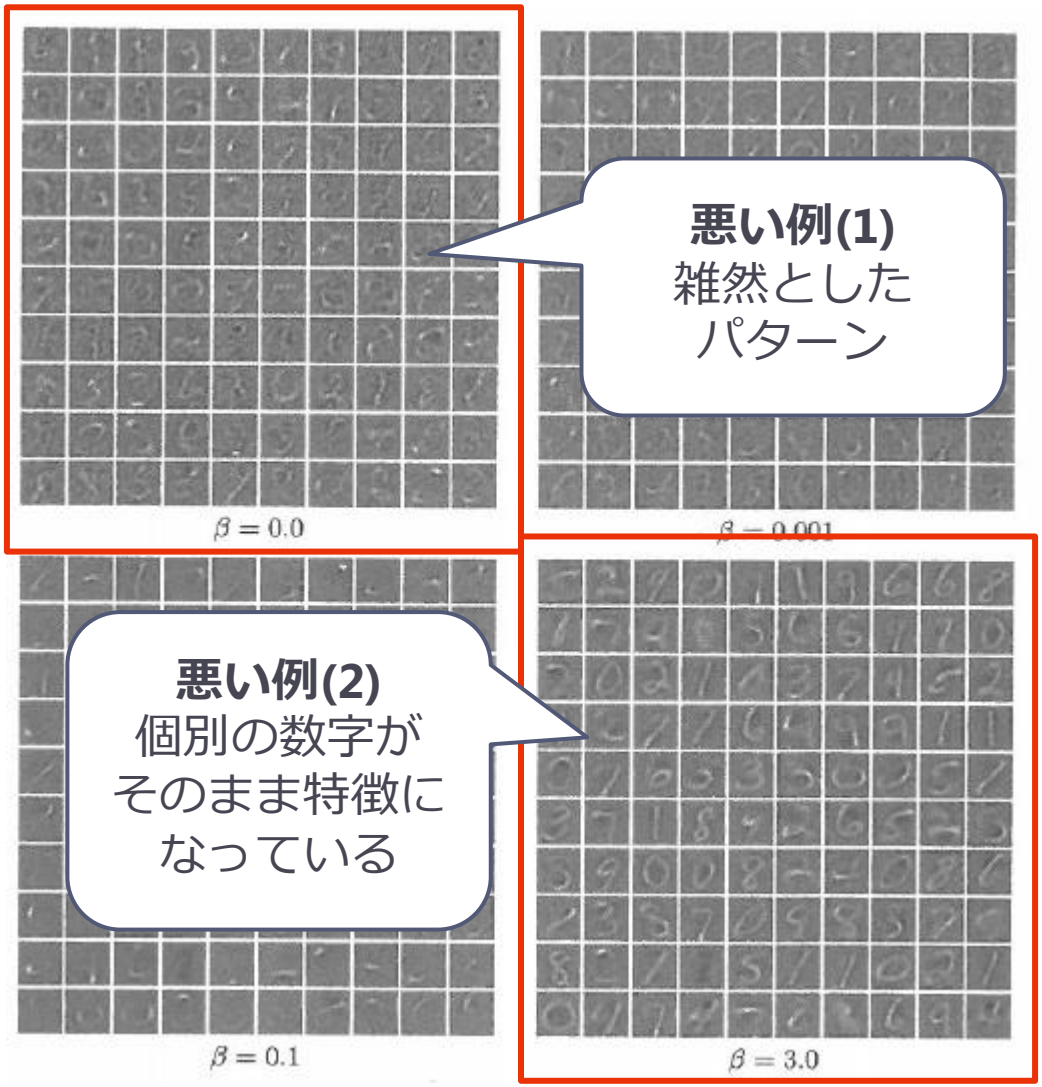
λ：平均の重み

前ミニバッチで使用したユニットjの平均活性度

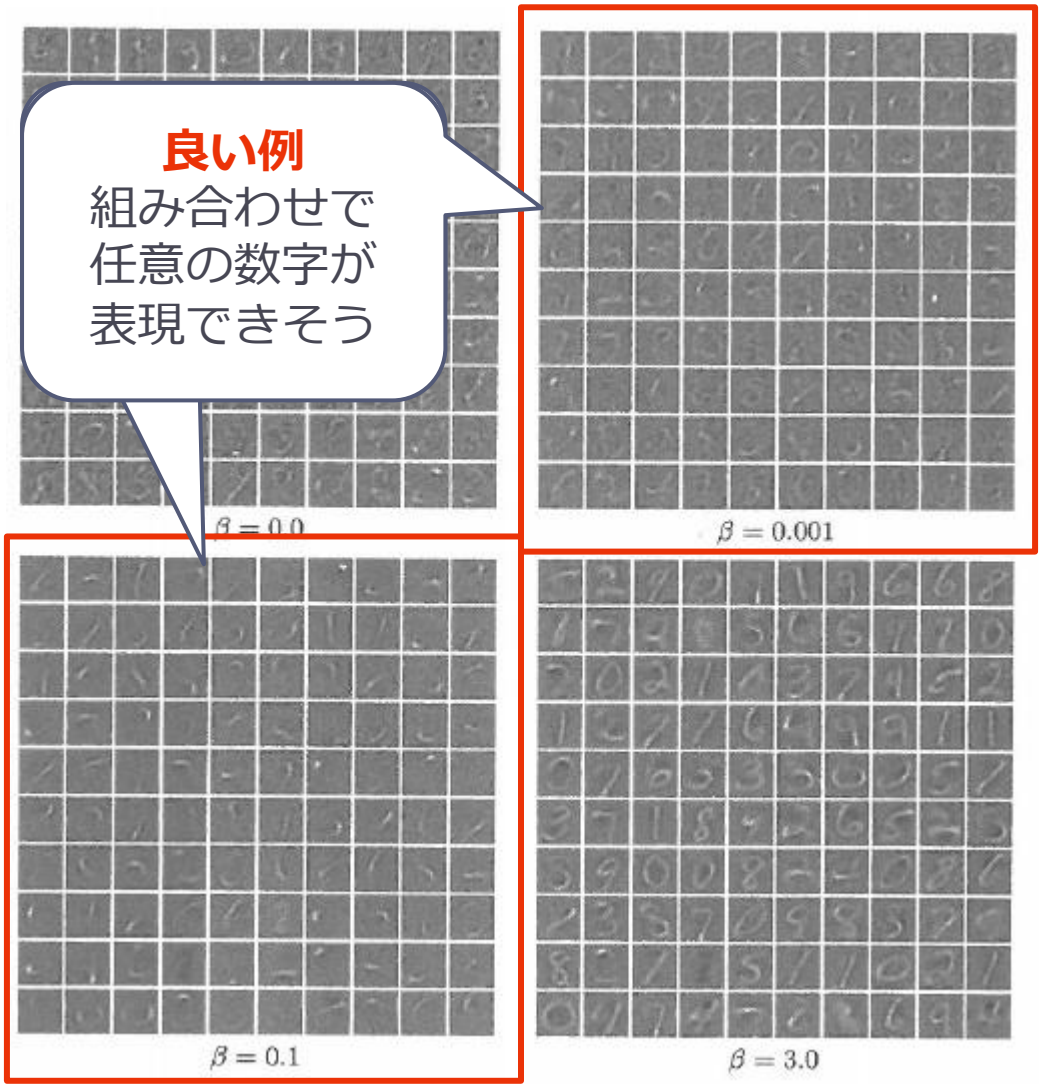
スパース正則化の効果（p.67）



スパース正則化の効果（p.67）



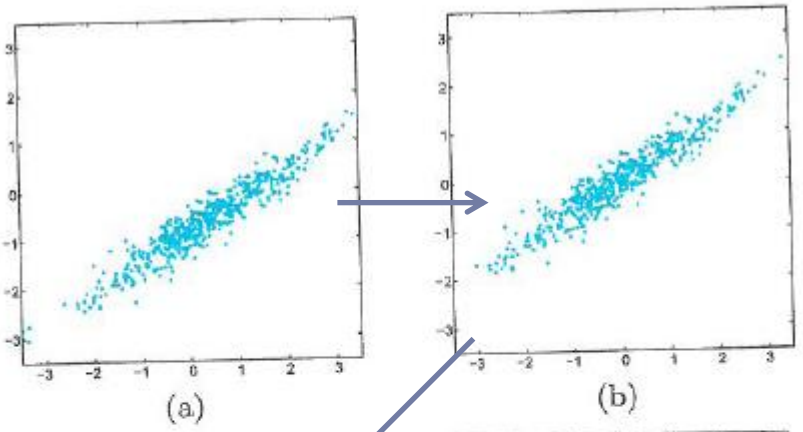
スパース正則化の効果（p.67）



白色化（p.67-72）

- 学習前に訓練データに処理を施し，偏りを除去する手法
  - 狙い：訓練サンプルの成分間の相関をなくす

元のサンプル



平均を取る

各成分の分散を
１にする

白色化

図3.4（p.34）

白色化（p.67-72）

- 各サンプルから平均を引いたものをx1,x2,,,,xNとするとき，成分間の相関は以下の共分散行列で表される

$$\mathbf{\Phi}_X \equiv \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^\top = \frac{1}{N}\mathbf{X}\mathbf{X}^\top$$

- 成分間の相関がない状態とは：
  →共分散行列Φxが対角行列になっているとき

白色化（p.67-72）

すなわち,

$$\mathbf{u}_n = \mathbf{P}\mathbf{x}_n \quad (n = 1, \ldots, N)$$

としたとき

$$\mathbf{\Phi}_U \equiv \frac{1}{N}\sum_{n=1}^{N} \mathbf{u}_n \mathbf{u}_n^\top = \frac{1}{N}\mathbf{U}\mathbf{U}^\top$$

が対角行列となるような**P**を求めればよい（教科書p.68,69参照）

白色化（p.67-72）

- **PCA白色化**
  - 共分散行列の固有ベクトルを利用することがサンプル集合の主成分分析(PCA)を行うことに通じる
- **ZCA白色化**
  - Pを対称行列に制限する考え方

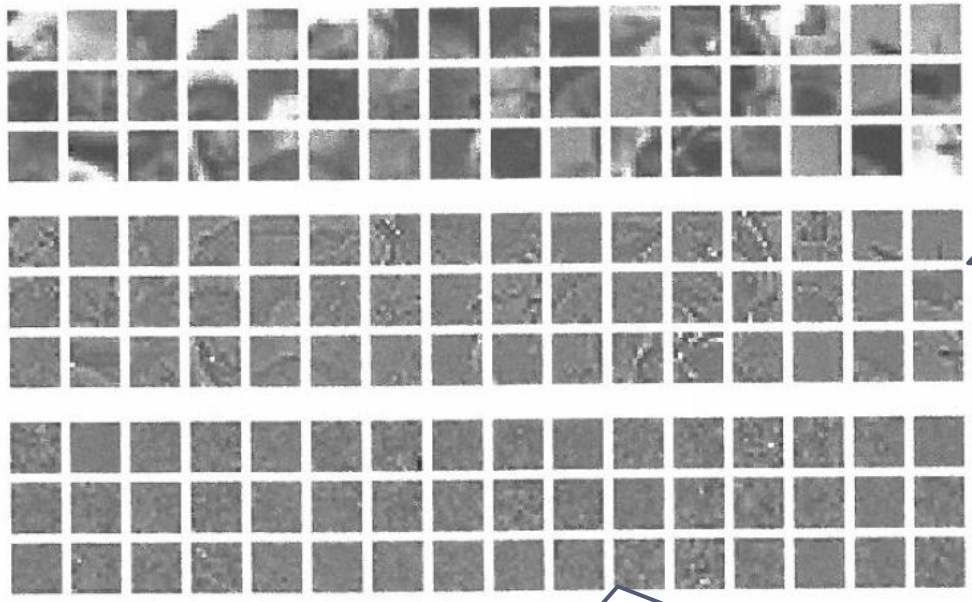
白色化（p.67-72）

- 白色化の結果（p.70）

元の
サンプル

ZCA白色化

PCA白色化

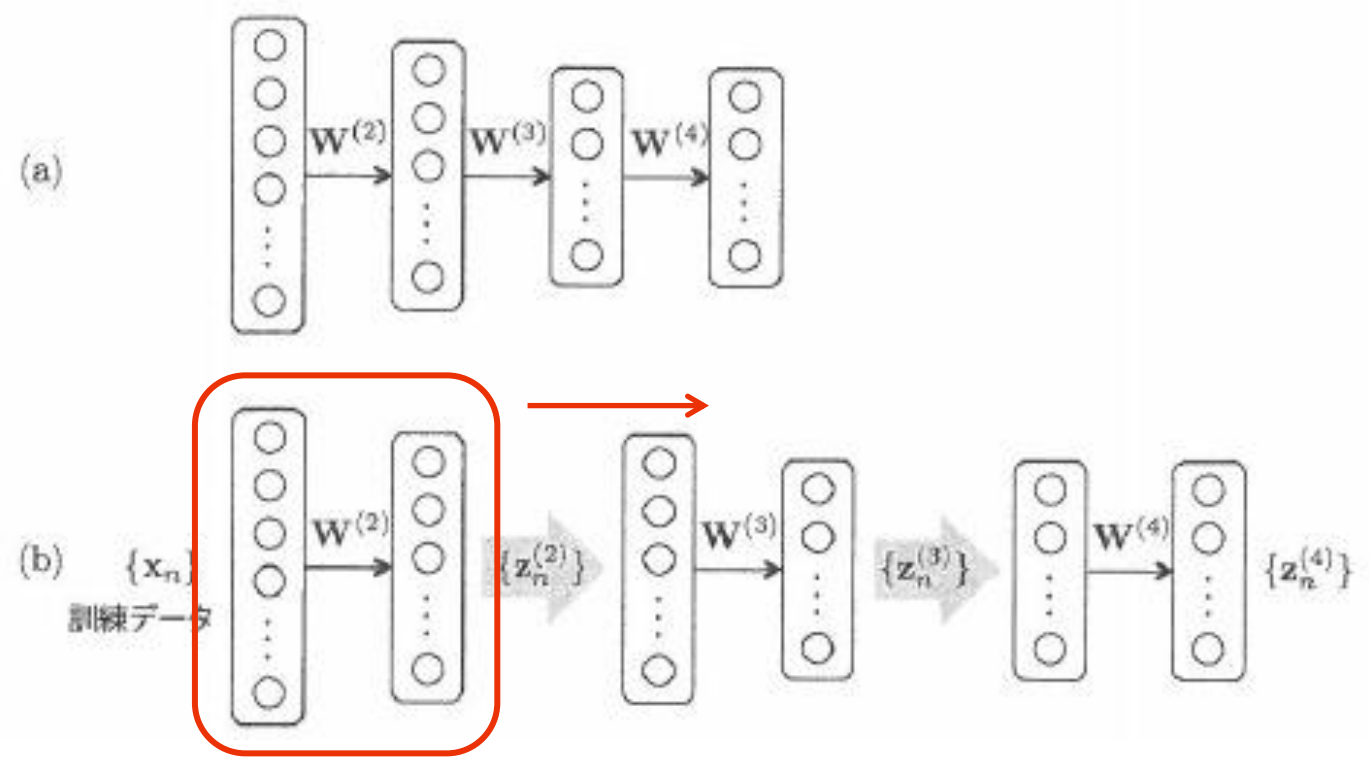


直流成分(明るさ)が取り除かれ，画像のエッジ(濃淡変化の大きい点)が強調される

高周波(明るさの強弱の振動回数が多い)成分が一律に強調される

ディープネットの事前学習(p.72-74)

- 自己符号化器を用いた事前学習（積層自己符号化器）

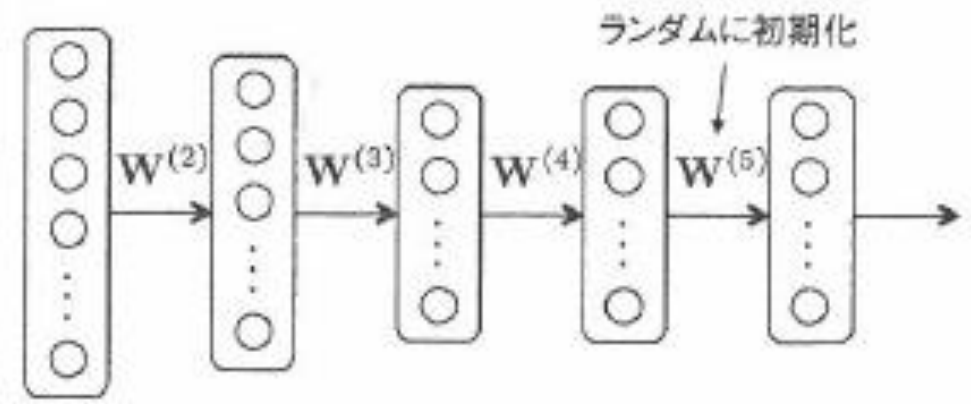


入力層から1層ずつ
ずらしていくイメージ

ディープネットの事前学習(p.72-74)

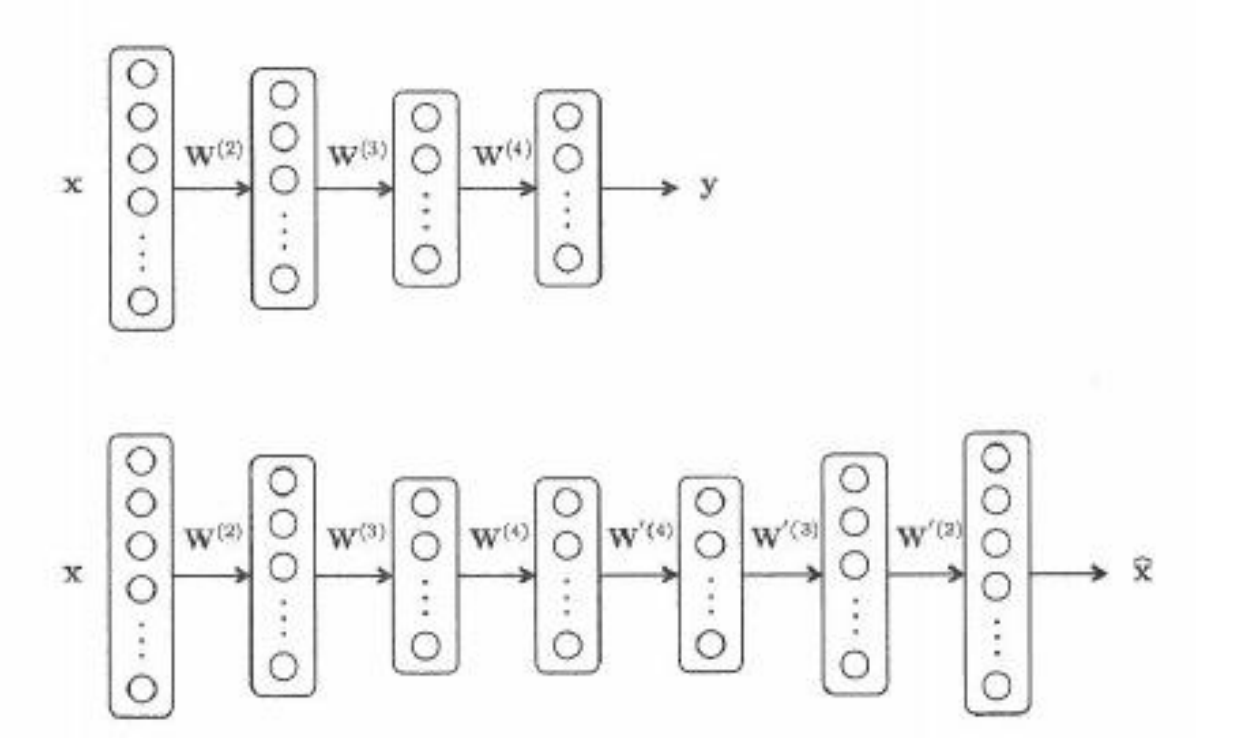
- 最上位層に1層以上追加(重みはランダムに初期化)し，教師あり学習を実行する

  →重みを全てランダムに決めるときに比べ勾配消失が起こる可能性はずっと小さくなる



その他の自己符号化器（p.74-77）

- 多層自己符号化器
  - 多層の順伝播型ネットワークを元に，これを出力層で折り返す
  - 学習が難しいという問題



その他の自己符号化器（p.74-77）

- デノイジング自己符号化器
  - 自己符号化器を拡張し，確率的な要素（制約ボルツマンマシンの考え方）を取り入れる
  - 訓練サンプル x にランダムノイズδxを付加する

$$\tilde{\mathbf{x}} = \mathbf{x} + \delta x \qquad \delta x \sim N(0,\, \sigma^2\, \mathbf{I})$$

  - ランダムノイズを付加した~xを自己符号化期に入力する

$$\hat{\mathbf{x}}(\tilde{\mathbf{x}}) = \tilde{\mathbf{f}}(\tilde{\mathbf{W}}\mathbf{f}(\mathbf{W}\tilde{\mathbf{x}} + \mathbf{b}) + \tilde{\mathbf{b}})$$

参考

- 自然言語処理のための深層学習
  - http://cgi.csc.liv.ac.uk/~danushka/papers/DeepNLP.pdf
- 情報意味論 (第15回) Deep Learning
  - http://www.sakurai.comp.ae.keio.ac.jp/classes/infosem-class/2012/15DeepLearning.pdf
- Deep Learningと画像認識 ～歴史・理論・実践～
  - http://www.slideshare.net/nlab_utokyo/deep-learning-40959442
- 深層学習入門
  - http://www.slideshare.net/bollegala/ss-39065162
- Deep Learningの技術と未来
  - http://www.slideshare.net/beam2d/deep-learning-22544096