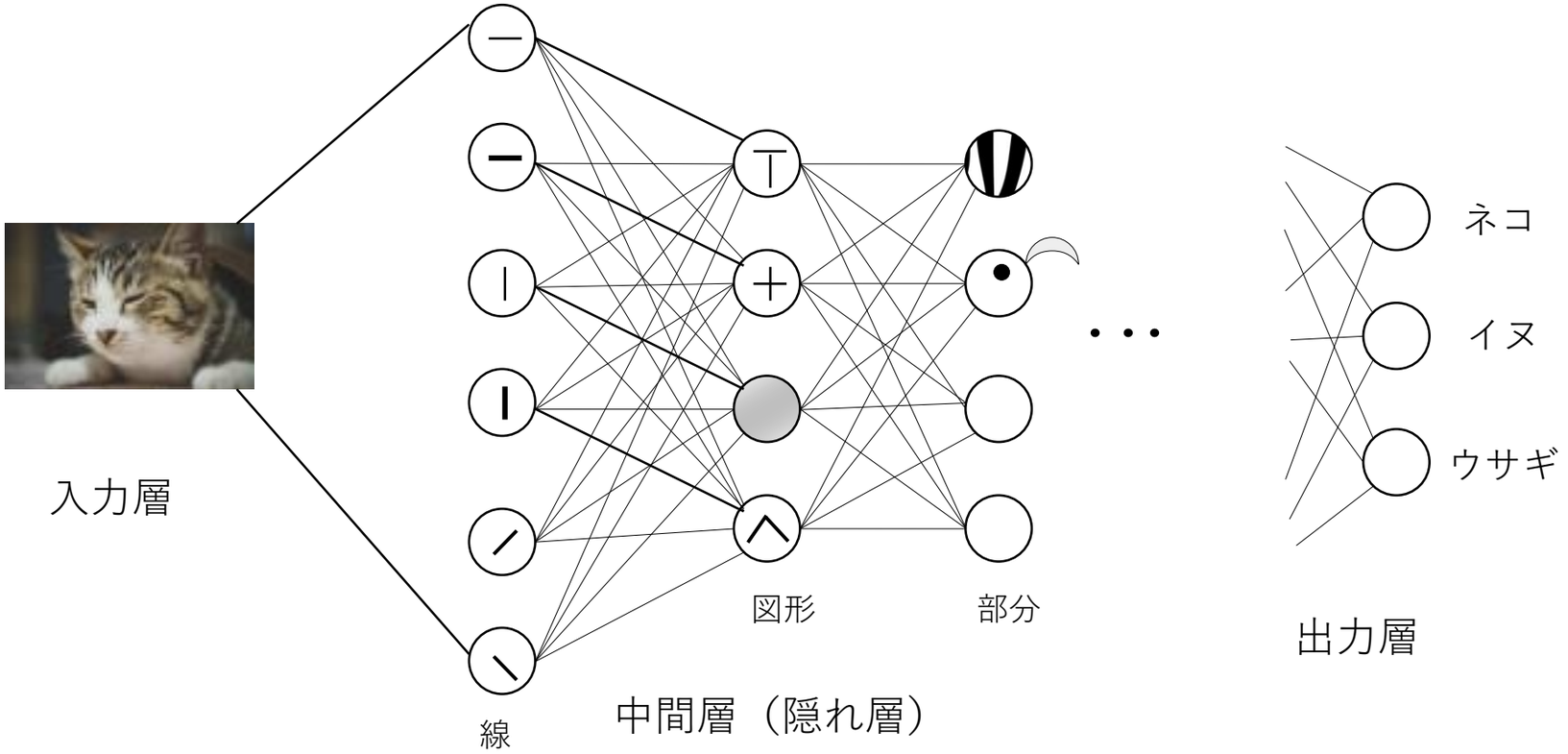
9. 深層学習

- 本章の説明手順

1. 深層学習の定義と学習の枠組み
2. 多階層学習における工夫
3. 深層学習における現実と理論のギャップ
4. 画像認識に適したネットワーク構造
5. 自然言語処理に適したネットワーク構造

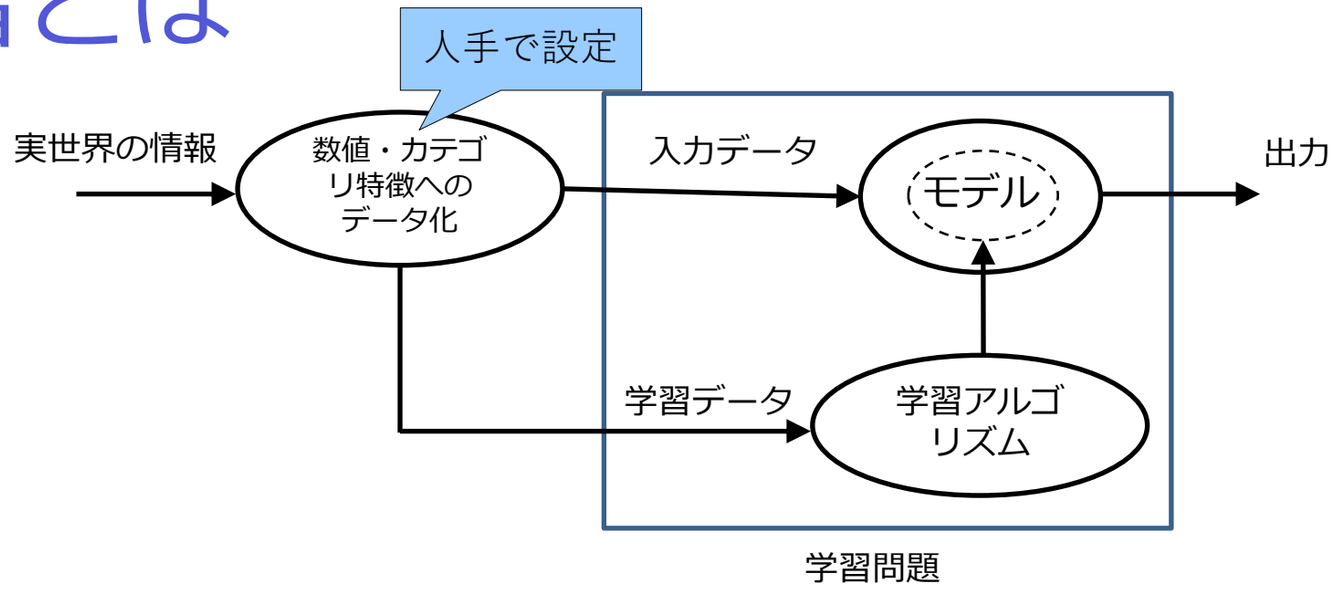
9.1 深層学習とは

- 深層学習 (deep learning)の定義
  - 多階層ニューラルネットワークの学習
  - 画像認識の層数の例：2012年8層 → 2016年1000層以上

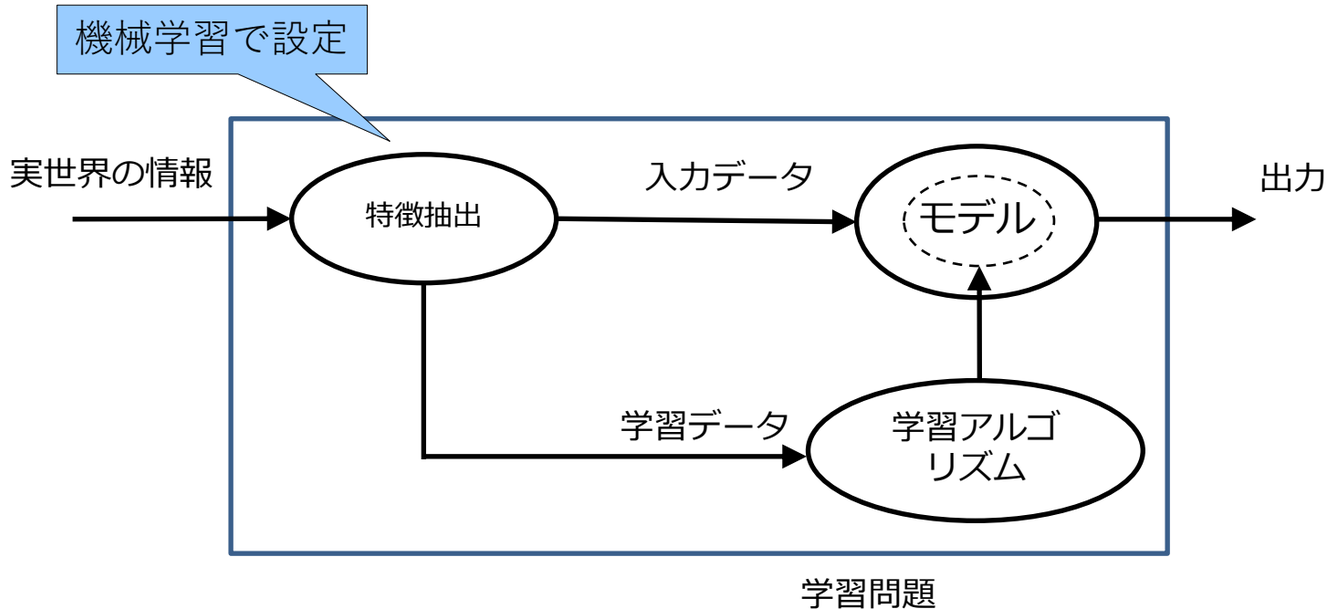


9.1 深層学習とは

これまでに説明した機械学習（識別）

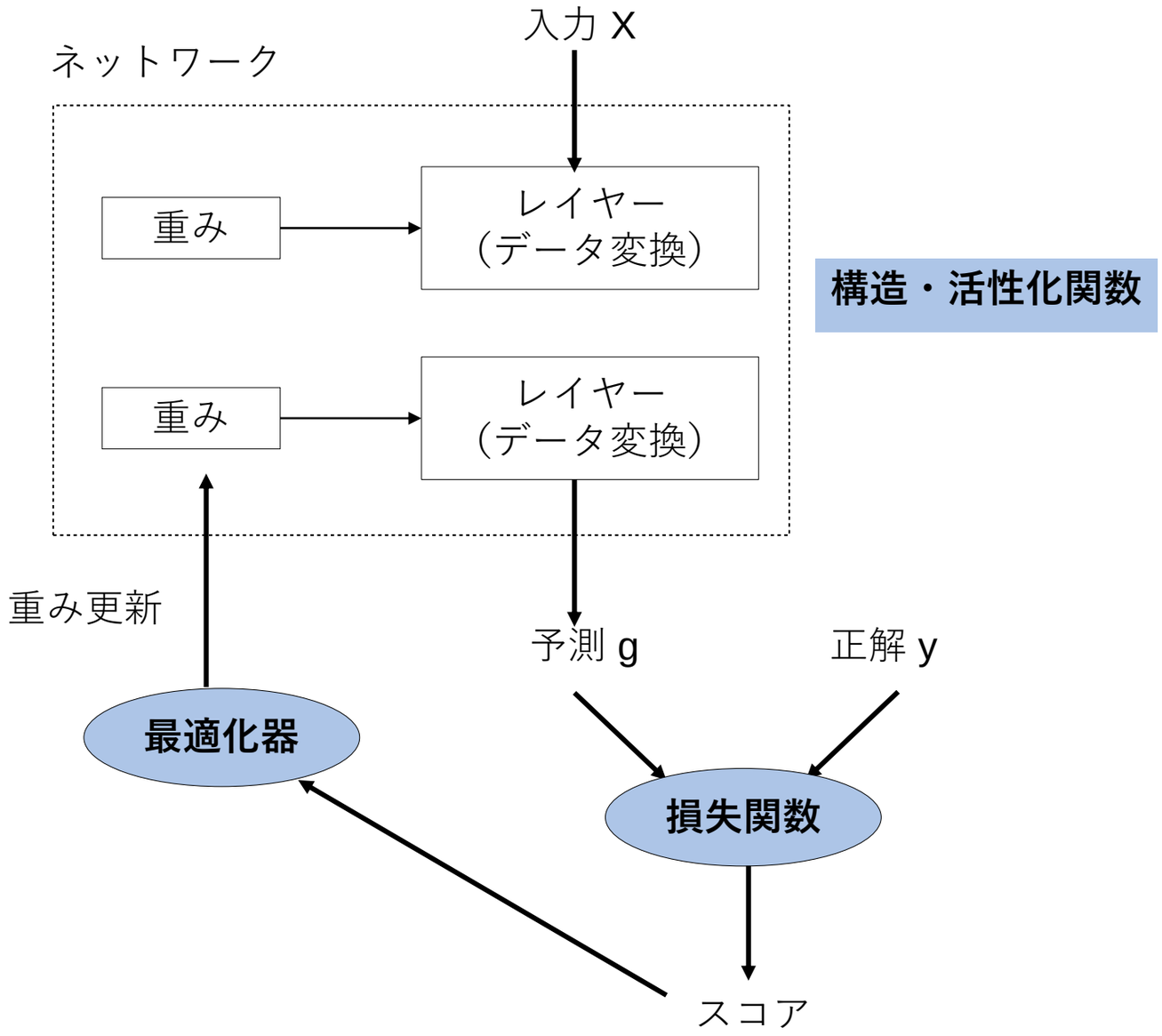


深層学習



9.2 深層学習の枠組み

復習



9.3 多階層ニューラルネットワーク
9.3.1 多階層ニューラルネットワークの学習

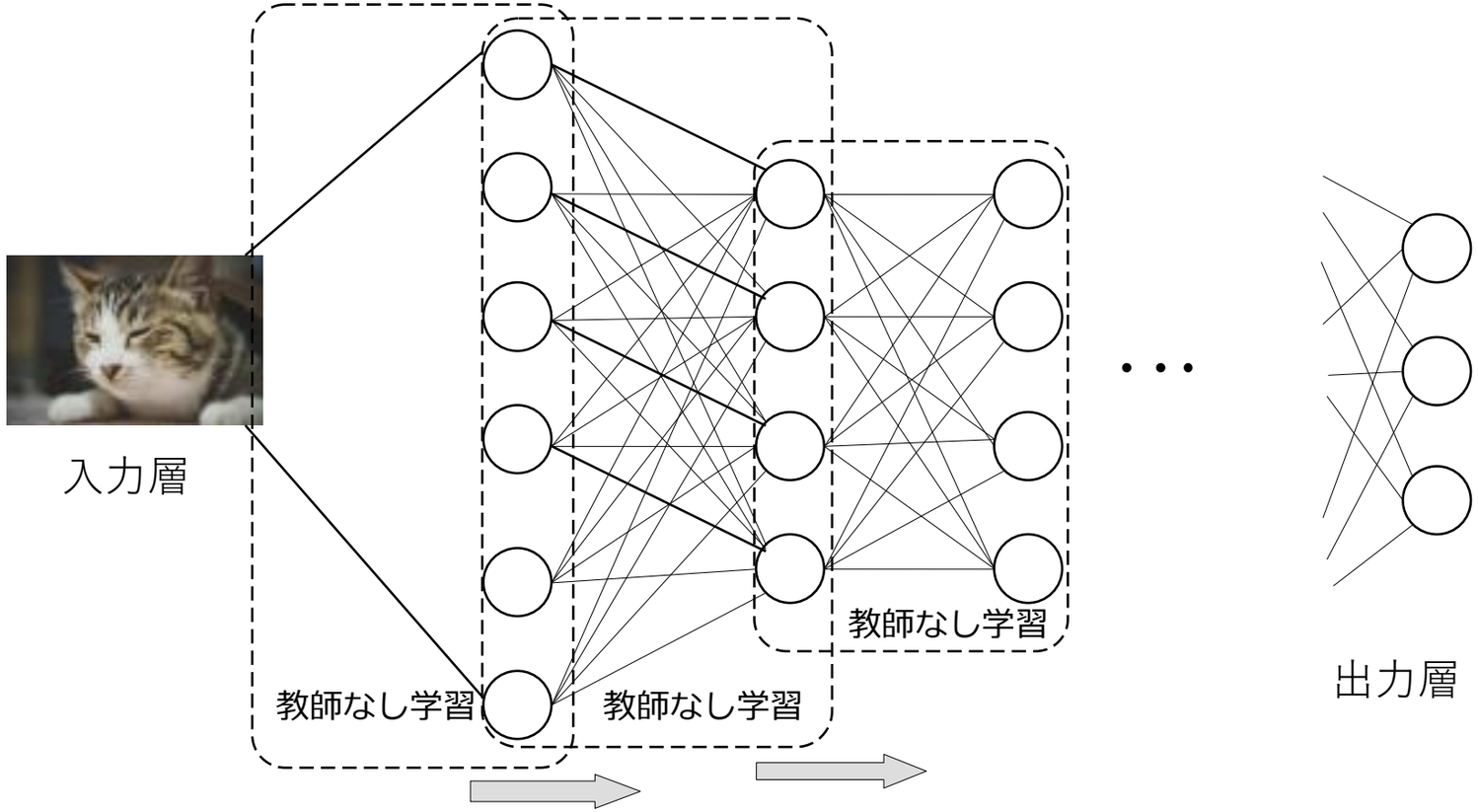
- 多階層学習の問題点
  - 勾配消失問題　第8章で説明済み
  - 多数のパラメータ（＝過学習になりやすい）
- 解決法
  - 事前学習（現在はあまり使われていない）
    - ただし、オートエンコーダの概念は有用
  - 活性化関数、最適化器の工夫
    - 活性化関数：ReLU, tanh, LeakyReLU, ELU
    - 最適化器: Adam, AdaGrad, RMSProp

第8章で説明済み

  - ドロップアウト
    - 正則化と同様の機能を持つ、学習上の工夫

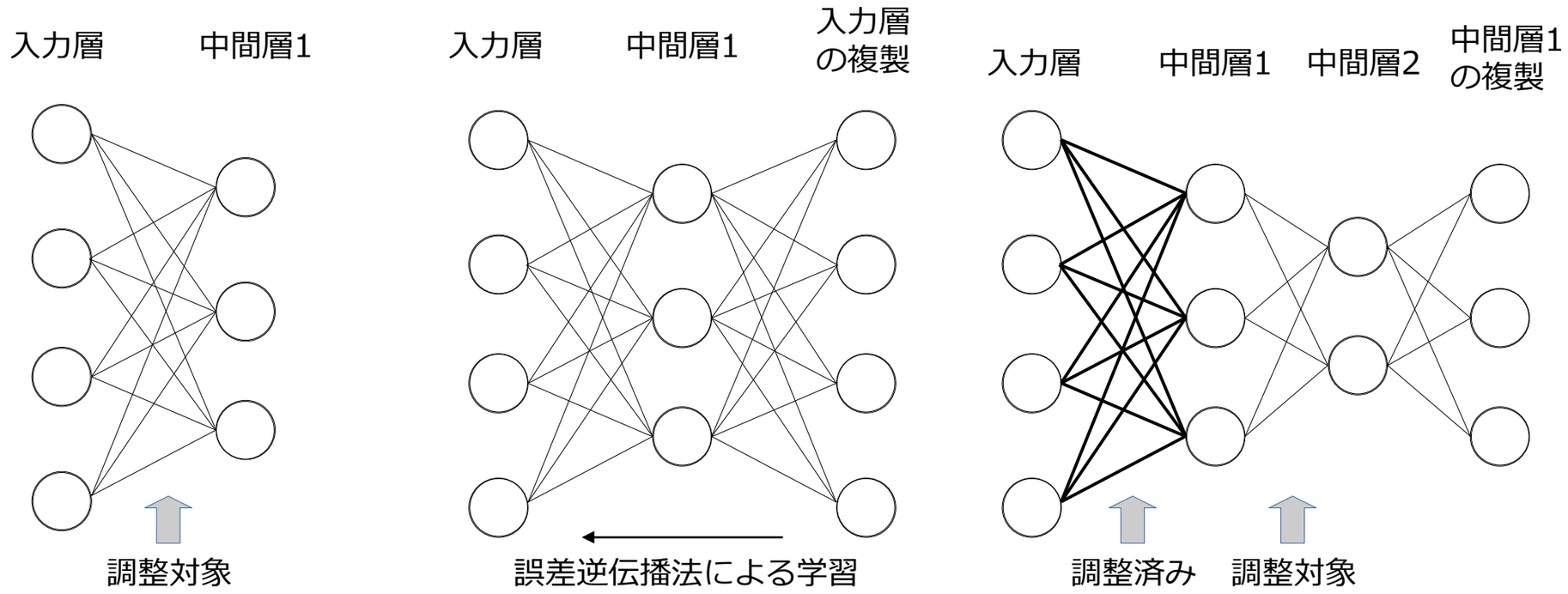
9.3.1 多階層ニューラルネットワークの学習

- 勾配消失に対処する事前学習法
  - 重みの初期値を入力層から順に自己写像で学習 (pre-training)
  - その後、ネットワーク全体を教師あり学習(fine-tuning)



9.3.2 オートエンコーダ

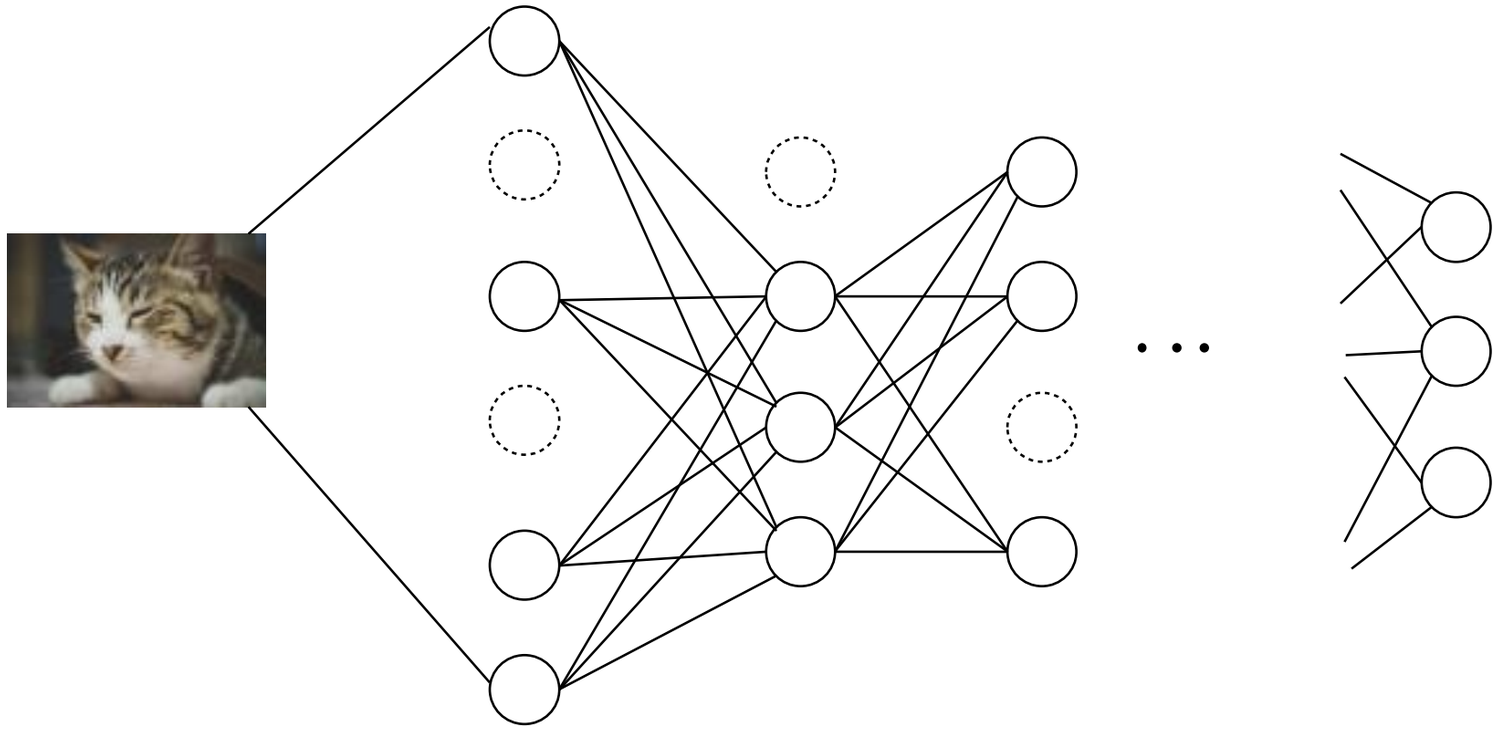
- アイディア:自己写像の学習で情報圧縮を実現
  - 少ない数の中間層で元の情報を保持させる



(a) 事前調整対象の重み　(b) オートエンコーダによる復元学習　(c) 1階層上の事前調整

9.3.3 多階層学習における工夫

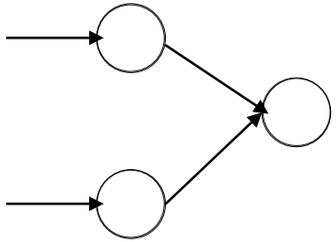
- 過学習の回避:ドロップアウト
  - 学習時に一定割合のユニットをランダムに消す
  - 認識時には学習後の重みに消去割合を掛ける



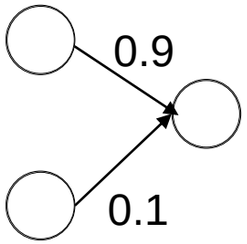
9.3.3 多階層学習における工夫

- ドロップアウトの効果
  - 正則化のような役割

下位二つのユニットが活性化（出力=1）したときのみ、上位のユニットも活性化させたい



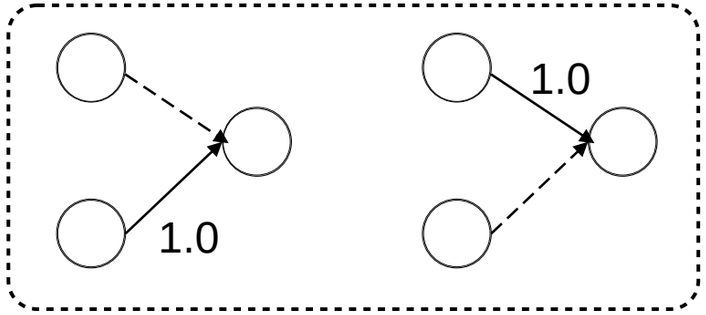
通常の学習



重みが偏る可能性
＝汎用性の低下

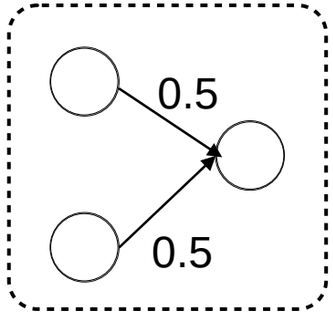
ドロップアウト
$p = 0.5$

学習時



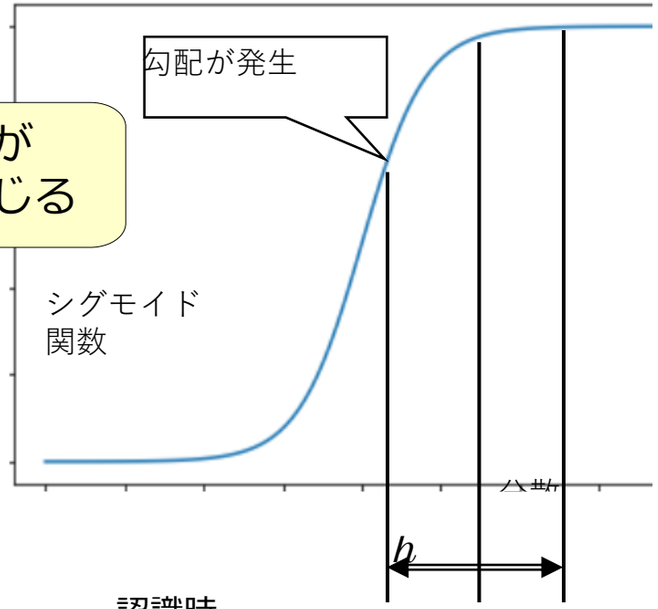
片方だけでもなるべく正解に
近づこうとする＝汎用性の向上

認識時



学習した重みを p 倍

ユニットへの入力の分散が
大きくなるので勾配が生じる



深層学習における現実と理論のギャップ

- なぜ多層にすると性能が向上する？
  - 3層NN中間層のユニット数が十分に多ければ、複雑な非線形関数を任意の精度で近似できるはず
- なぜ過学習しない？
  - パラメータが多い＝バイアスが少ないモデルは過学習しやすいはず
- なぜ最適に近いパラメータが見つかる？
  - 損失関数が複雑な形をしていると、局所最適解で学習が止まることが多いはず

深層学習における現実と理論のギャップ

- 多層で性能が向上する理由
  - 近似誤差レート（パラメータ数と関数近似誤差の関係）
    - 3層NNもDNNもパラメータ数は有限なのでこのレートで評価する
    - 関数が滑らかな場合、3層NNの近似誤差レートを、多層で改善することはできない
  - DNNによる改善
    - 関数がジャンプを持つ場合
      - ✓ 滑らかな関数を表現する層の間に階段関数の層を挟むことで近似
    - 関数が非均一的な滑らかさを持つ場合
      - ✓ 層毎に異なる幅を持つ短冊状の関数を表現することで近似

これらにより、データのより複雑な特徴が抽出でき、特徴抽出と識別の役割分担も可能になっている

深層学習における現実と理論のギャップ

- 過学習しにくい理由
  - 仮説1：暗黙的正則化が行われている
    - 宝くじ仮説：巨大なDNNは小さなNNの集合体であり、学習によって当たりのNNを引き当てている
    - 根拠：学習後に枝刈りを行っても、ほぼ同じ性能が維持できる
  - 仮説2：損失平坦性
    - 学習後の最適パラメータの近くで損失関数が平坦なら過学習は起こりにくい
  - 仮説3：二重降下
    - パラメータ数の増加によって、一旦下降した汎化誤差は上昇に転じるが、パラメータ数がデータ数より多くなると再度下降する

深層学習における現実と理論のギャップ

- 局所最適解で学習が止まらない理由
  - 過剰パラメータ化
    - NNのある層のパラメータを一定以上過剰にすると、最急勾配法は損失関数の値をほぼ0にするパラメータに到達する
    - 過剰パラメータを持つ層が損失関数全体を押し下げる
      - ✓ 損失関数は負にならないため、0に近い多くの値が最適値となる
  - 確率的最急勾配法の効果の解明
    - 更新されたパラメータの散らばりは、損失関数の値が小さくなる点に集中する分布に従う

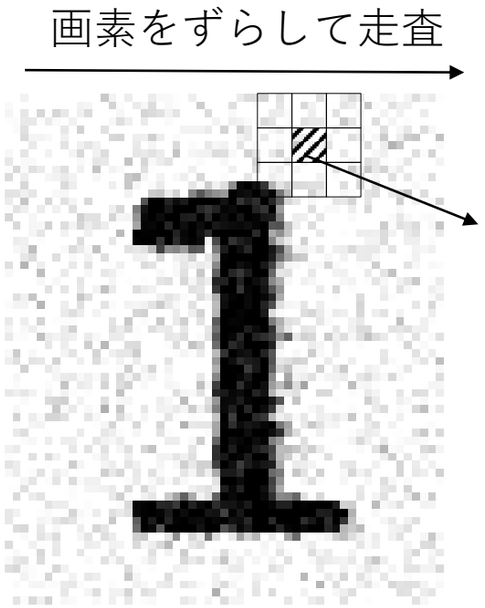
9.4 畳み込みネットワーク

- 畳み込みネットワーク(CNN)とは
  - 画像処理に適したネットワーク
  - 畳み込み層とプーリング層を交互に重ねて特徴抽出
    - 畳み込み層：フィルタを使って画像のパターンを見つける
    - プーリング層：位置の変動を吸収するダウンサンプリング
    - これらを交互に重ねることで、複雑な特徴を表現可能
  - バッチ標準化層で入力値を調整することもある
  - 最後は数段の密結合層（ReLU+Softmax）

9.4 畳み込みネットワーク

- 畳み込み層
  - 画像フィルタの適用と同じ

特定の画像入力に反応する
脳の視覚野領域の処理に対応



この画素の値を

$$\sum_{p=0}^{2}\sum_{q=0}^{2} x_{i+p,j+q} w_{pq}$$

と置き換える

$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$
$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$
$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$

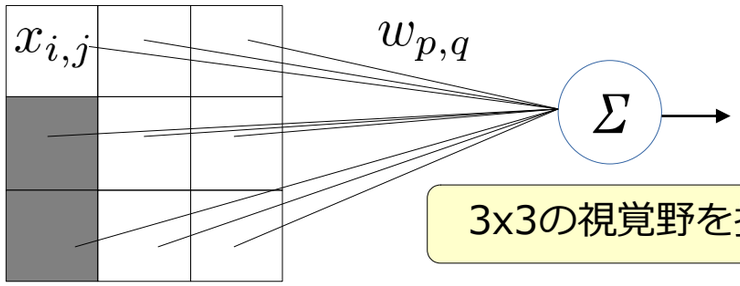
平均値フィルタ

ノイズを軽減する

-1	0	1
-1	0	1
-1	0	1

（縦）エッジフィルタ

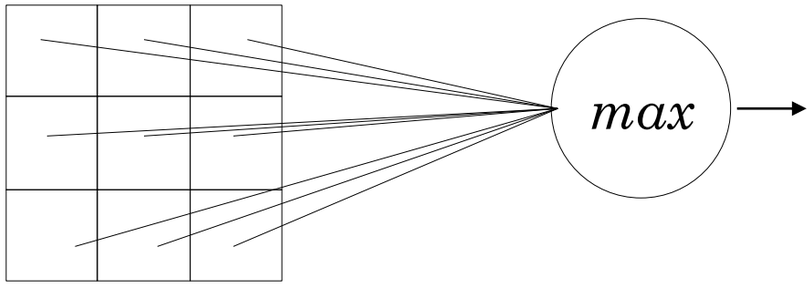
縦線を検出する



3x3の視覚野を持つニューロン

9.4 畳み込みネットワーク

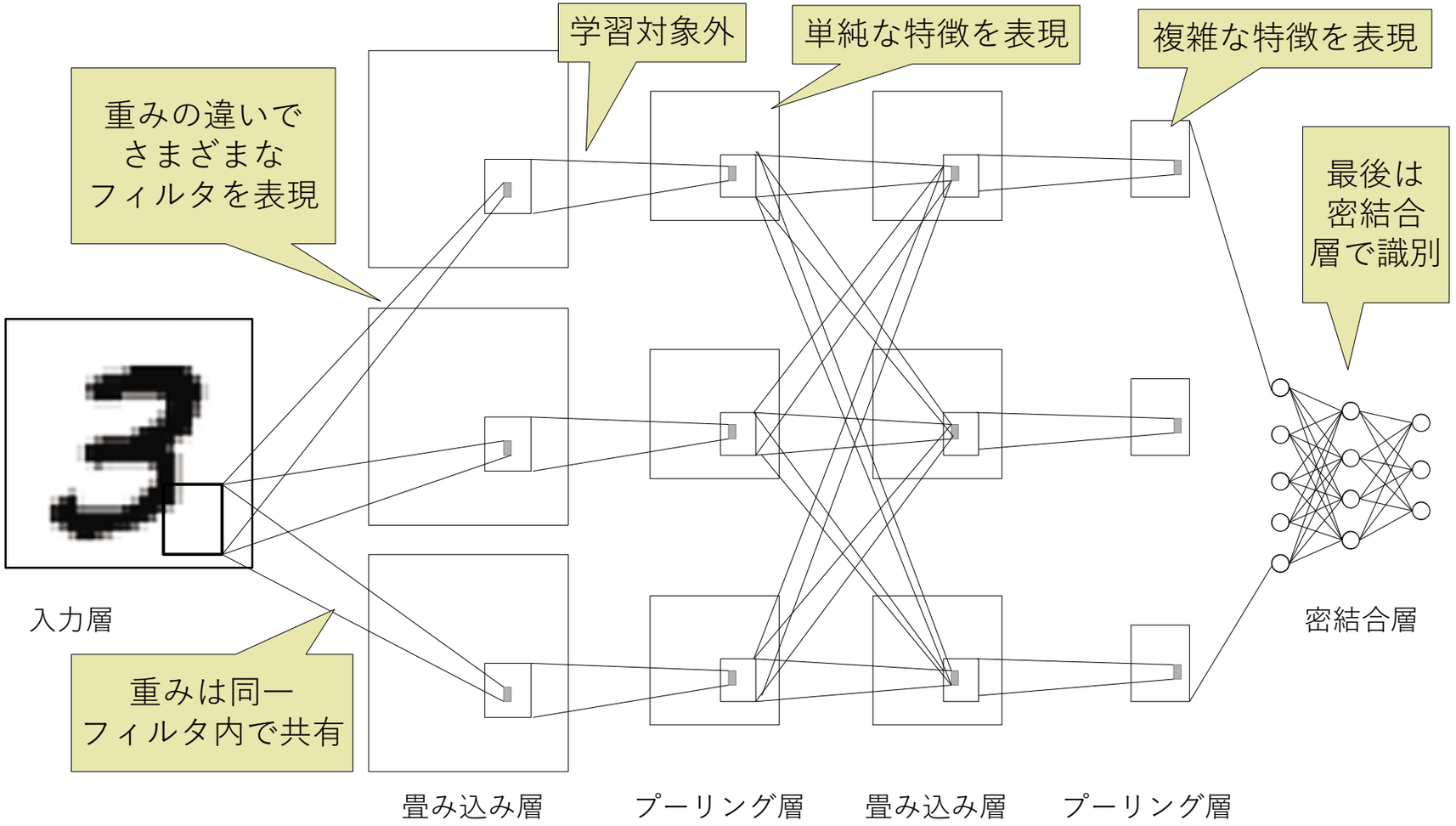
- プーリング層
  - 一定範囲の最大値あるいは平均値を計算
  - 畳み込み層によって発見した特徴の位置をぼかす役割



- バッチ標準化層
  - 入力値からミニバッチ毎の平均値を引いて標準偏差で割る
  - 多階層の演算による値の大きな変動やミニバッチ毎の分布の違いを吸収

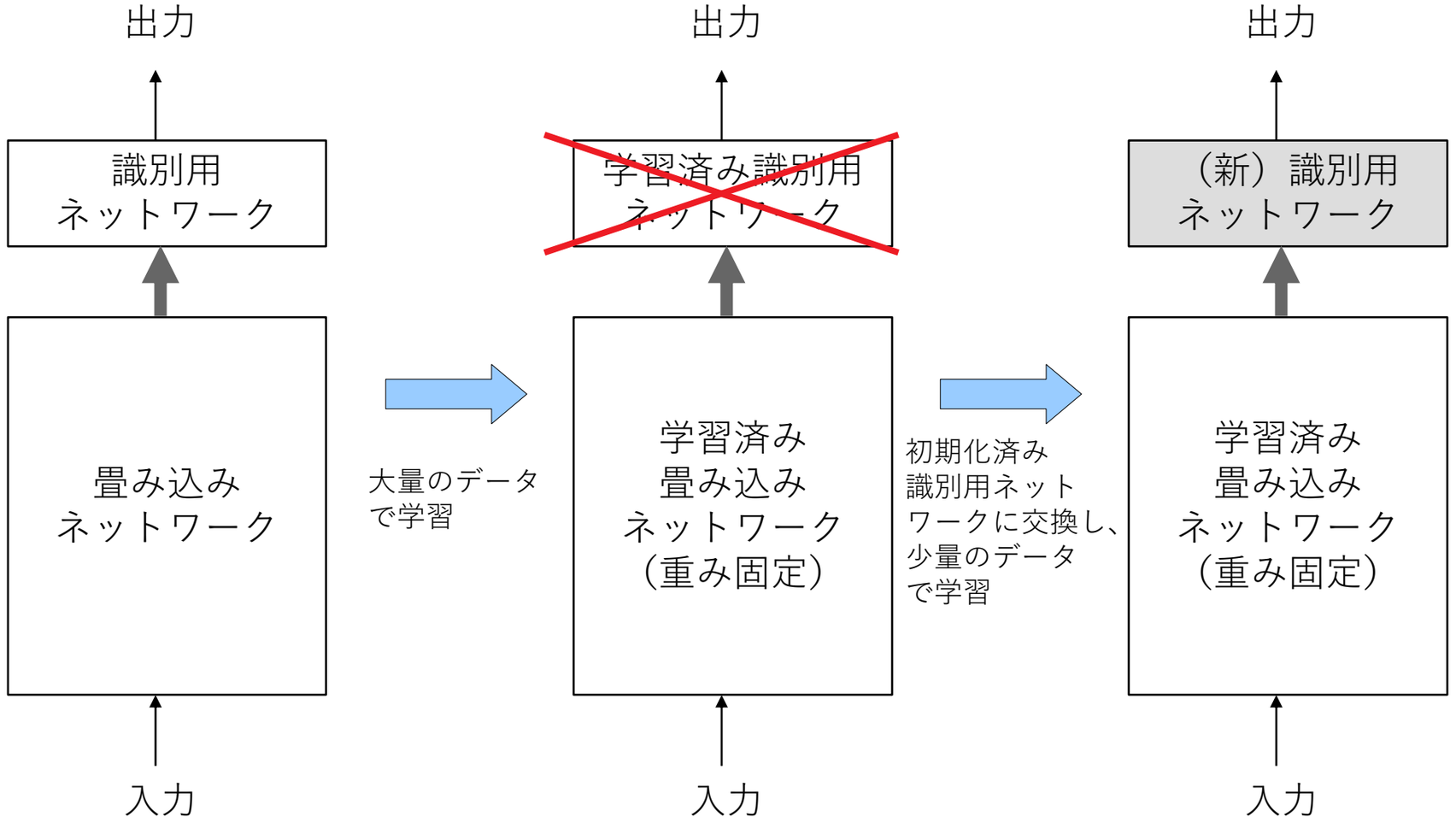
9.4 畳み込みネットワーク

- 畳み込みネットワークの構造



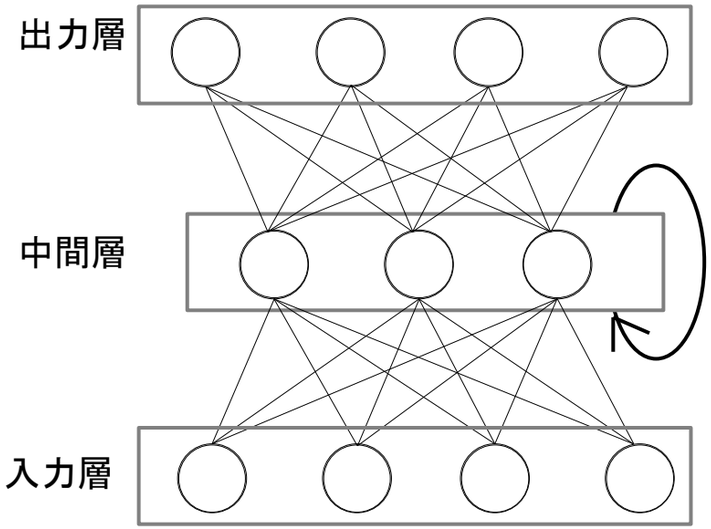
転移学習

- 大量のデータで学習済みのDNNがあるとき
- 対象タスクのデータが少量でも転移学習が可能

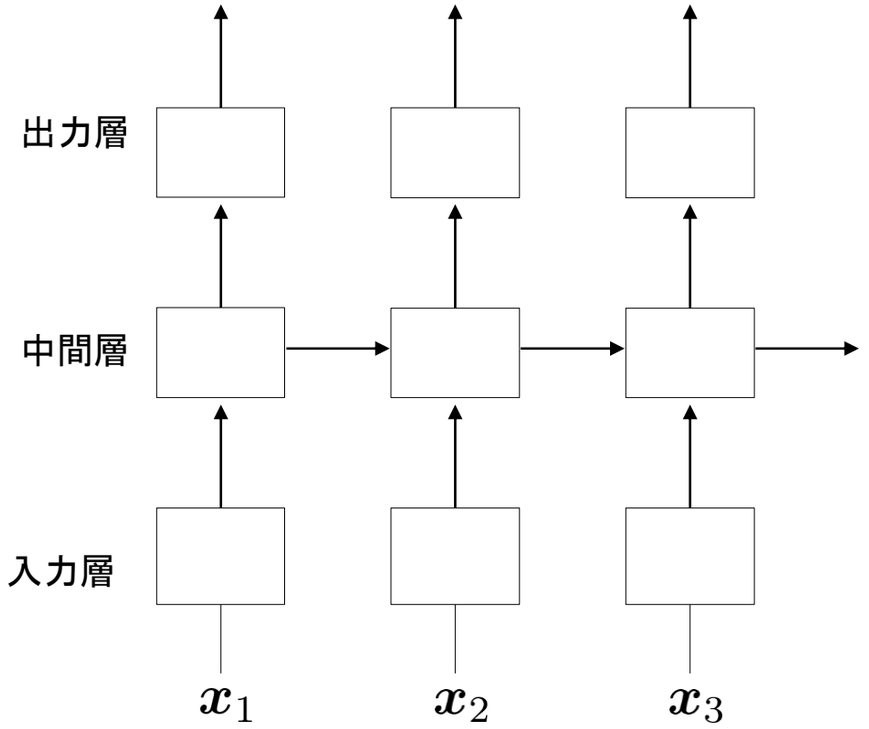


9.5 リカレントニューラルネットワーク

- リカレントニューラルネットワークとは
  - 時系列信号の認識や自然言語処理に適する
  - 一つ前の中間層の出力を、次の入力と結合



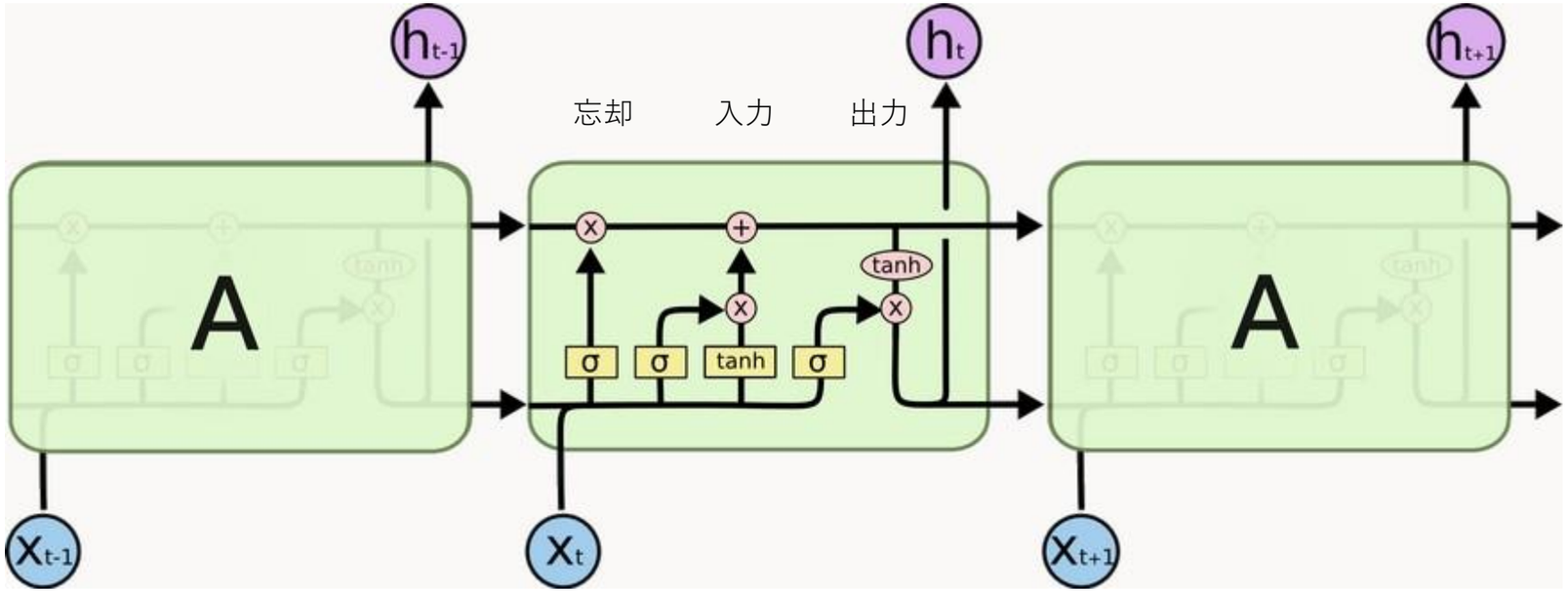
(a) リカレントニューラルネットワーク



(b) 帰還路を時間方向に展開

9.5 リカレントニューラルネットワーク

- LSTM (long short-term memory)
  - いくつかのゲートからなる内部構造をもつユニット
  - ゲート：選択的に情報を通すメカニズム



X：全要素に対する掛け算（**ゲート**）

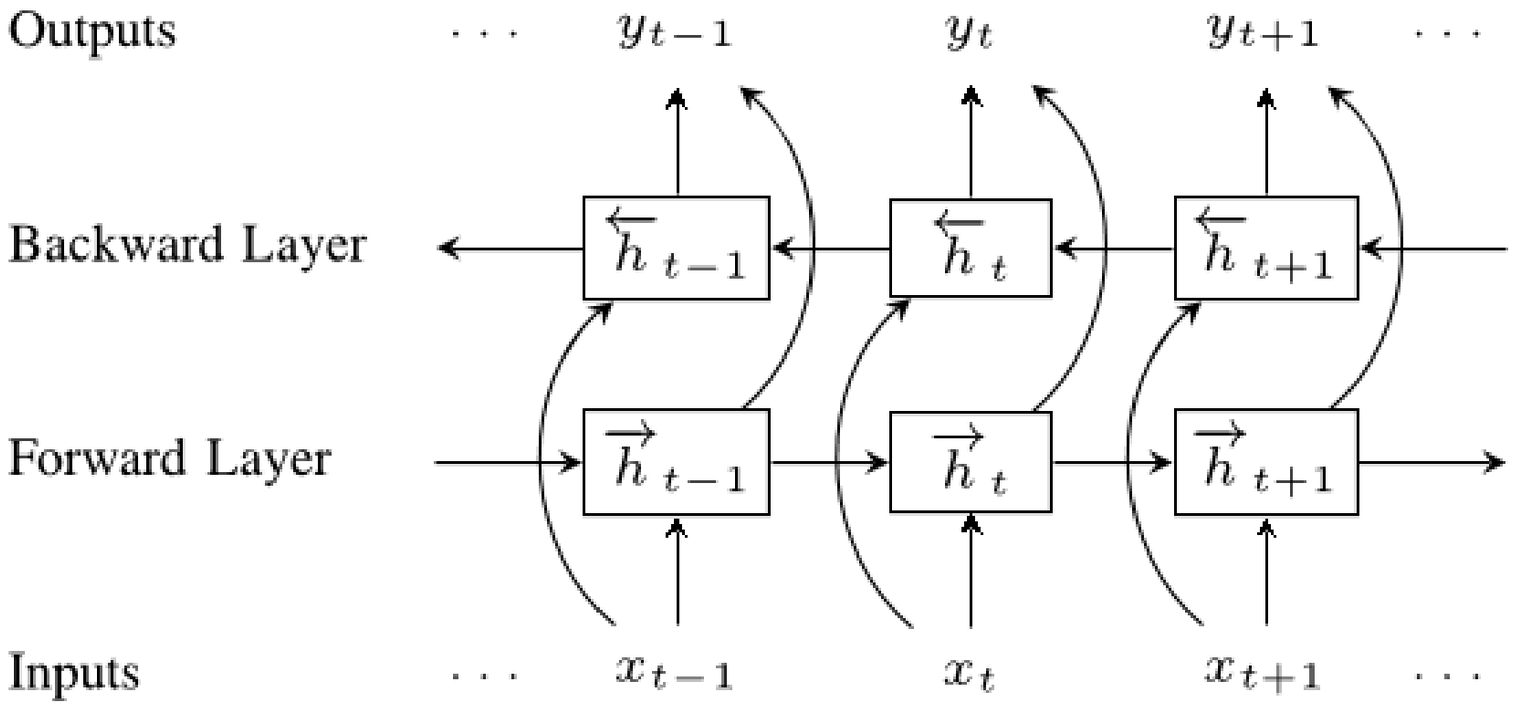
+：要素同士の足し算

σ：シグモイドを活性化関数とするNN

http://colah.github.io/posts/2015-08-Understanding-LSTMs/

9.5 リカレントニューラルネットワーク

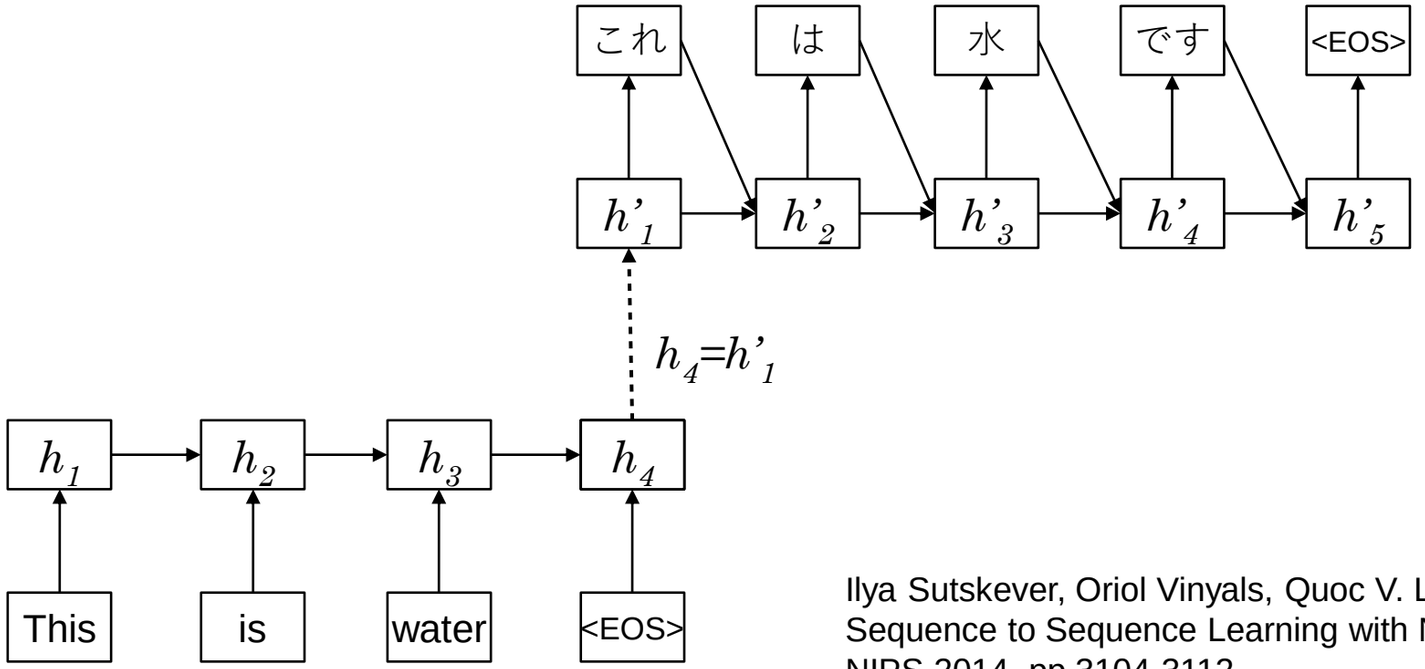
- 双方向RNN
  - 過去だけでなく、未来の情報も用いて出力を計算



He, L., Qian, Y., Soong, F.K., Wang, P., & Zhao, H. (2015). A Unified Tagging Solution: Bidirectional LSTM Recurrent Neural Network with Word Embedding. CoRR, abs/1511.00215.

9.5 リカレントニューラルネットワーク

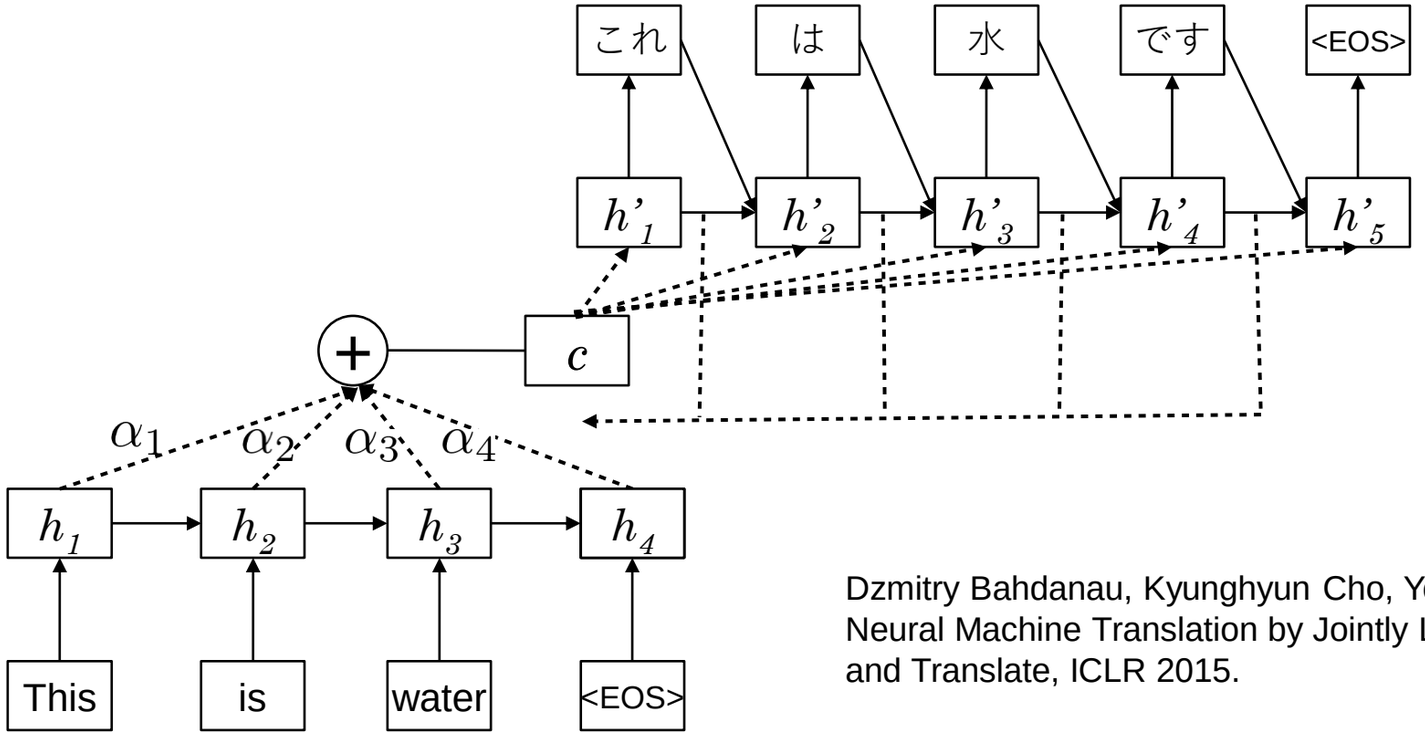
- Encoder-Decoder
  - 入力の内容をひとつの表現にまとめて、そこから出力を生成



Ilya Sutskever, Oriol Vinyals, Quoc V. Le :
Sequence to Sequence Learning with Neural Networks,
NIPS 2014, pp.3104-3112.

9.5 リカレントニューラルネットワーク

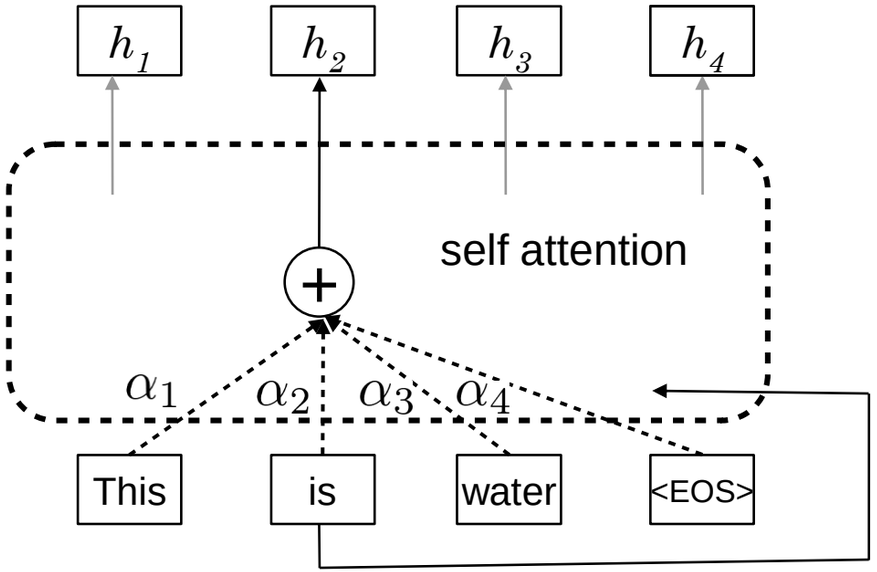
- Encoder-Decoder + Attention
  - アテンション：Encoder内のすべての中間層出力の重み付き和（入力のどの部分を見るかという情報）

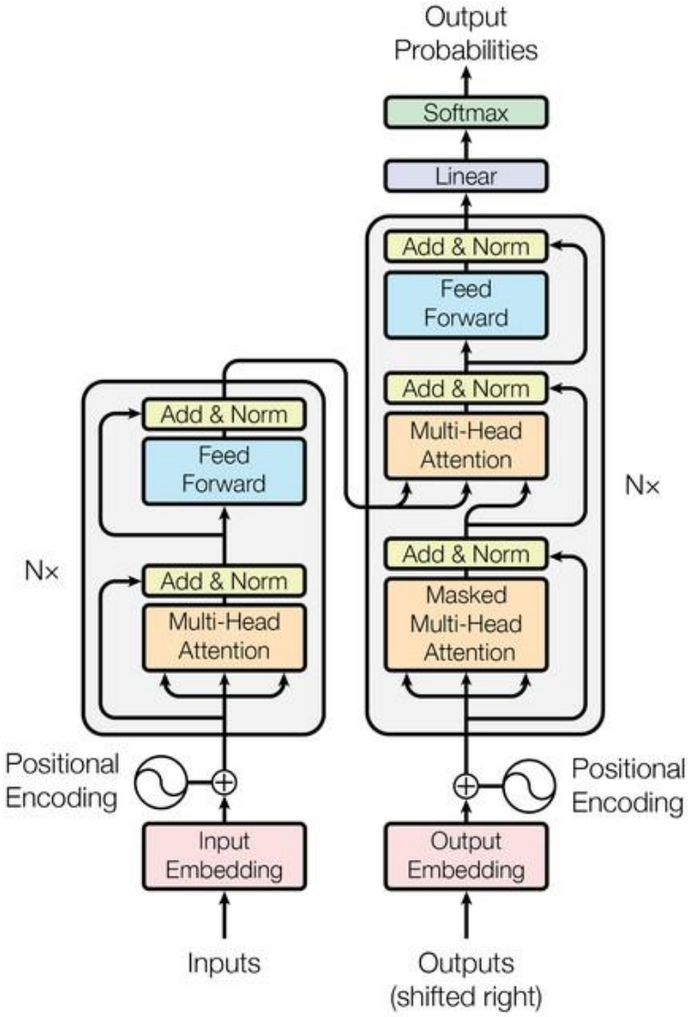


Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio : Neural Machine Translation by Jointly Learning to Align and Translate, ICLR 2015.

Transformer

- Self-attention + フィードフォワードNN
  - 自分の中間表現を作るときに、入力の他の部分との関係を計算
  - BERTなどの事前学習モデルに使われる





Ashish Vaswani et.al. : Attention Is All You Need, arXiv:1706.03762

まとめ

- 深層学習は多階層のニューラルネットワーク
  - モデルの構造・損失関数・最適化器を指定して学習を行うという点では通常のニューラルネットワークと同じ
- 多階層学習における特異な振る舞い
  - 膨大なパラメータを持つ深層学習では、通常の機械学習の常識に当てはまらないことが起こっている可能性
- 画像認識・自然言語処理などのタスクに適したネットワーク構造が提案されている
- 事前学習モデルの活用が有望