iTHEMS RIKEN interdisciplinary Theoretical & Mathematical Sciences

RIKEN

Residual Networkと深層化について

Masato Taki RIKEN, iTHEMS

2017.9/25

@Astro-AI workshop

深層学習の急速な進展

ここ数年の典型的な達成の例を挙げると

深層学習の急速な進展

ここ数年の典型的な達成の例を挙げると

- top5エラー2.25%の達成とImageNetコンペの終了

深層学習の急速な進展

ここ数年の典型的な達成の例を挙げると

- top5エラー2.25%の達成とImageNetコンペの終了
- アテンション機構の導入による自然言語処理への応用

深層学習の急速な進展

ここ数年の典型的な達成の例を挙げると

- top5エラー2.25%の達成とImageNetコンペの終了
- アテンション機構の導入による自然言語処理への応用
- 強化学習への波及

深層学習の急速な進展

ここ数年の典型的な達成の例を挙げると

- top5エラー2.25%の達成とImageNetコンペの終了
- アテンション機構の導入による自然言語処理への応用
- 強化学習への波及
- 携帯端末など身近なデバイスへの積極的導入(e.g. neural engine & CoreML for iPhoneX/iOS)

深層学習の急速な進展

ここ数年の典型的な達成の例を挙げると

- top5エラー2.25%の達成とImageNetコンペの終了
- アテンション機構の導入による自然言語処理への応用
- 強化学習への波及
- 携帯端末など身近なデバイスへの積極的導入(e.g. neural engine & CoreML for iPhoneX/iOS)
- 深層学習をうたう数多くのスタートアップ企業

深層学習の急速な進展

ここ数年の典型的な達成の例を挙げると

- top5エラー2.25%の達成とImageNetコンペの終了
- アテンション機構の導入による自然言語処理への応用
- 強化学習への波及
- 携帯端末など身近なデバイスへの積極的導入(e.g. neural engine & CoreML for iPhoneX/iOS)
- 深層学習をうたう数多くのスタートアップ企業

もはや深層学習は特別な手法ではなく通常技術化し始めている。

深層(ディープ)の意味

深層学習： ニュアンスは状況・時期で若干変容

共通点は多層ニューラルネットをもとに組み立てられたモデル/アーキテクチャでの階層的な情報処理

深層(ディープ)の意味

深層学習： ニュアンスは状況・時期で若干変容

共通点は**多層ニューラルネット**をもとに組み立てられたモデル/アーキテクチャでの階層的な情報処理

多層＝**深層**(+ 正則化による大きなモデル自由度の制御)＝情報表現の階層性が高い性能の秘密だと信じられている

深層(ディープ)の意味

深層学習： ニュアンスは状況・時期で若干変容

共通点は**多層ニューラルネット**をもとに組み立てられたモデル/アーキテクチャでの階層的な情報処理

多層＝**深層**(+ 正則化による大きなモデル自由度の制御)＝情報表現の階層性が高い性能の秘密だと信じられている

表現学習のコンセプトを極めてよく実現する

深層(ディープ)の意味

深層学習： ニュアンスは状況・時期で若干変容

共通点は**多層ニューラルネット**をもとに組み立てられたモデル/アーキテクチャでの階層的な情報処理

多層＝**深層**(+ 正則化による大きなモデル自由度の制御)＝情報表現の階層性が高い性能の秘密だと信じられている

表現学習のコンセプトを極めてよく実現する

深層にすることによって引き起こされかねない問題を解決する技術群が深層学習

本質的にはモデル選択(正則化)によって性能がほとんど決まる
[Zhang et.al, 2016]

どれほど深層か？- ILSVRC (classification)

2012年　AlexNet (16%)　8層

2013年　ZFNet (Clarifai) (11%)　8層

2014年　GoogLeNet (6.6%)　22層

2015年　Microsoft Asia (3.57%)　152層

2016年　Trimps-Soushen (2.99%)　model ensemble

2017年　SENet (2.25%)　層

どれほど深層か？- ILSVRC (classification)

2012年　AlexNet (16%)　8層

2013年　ZFNet (Clarifai) (11%)　8層

2014年　GoogLeNet (6.6%)　22層

2015年　Microsoft Asia (3.57%)　152層　◀　ResNet

2016年　Trimps-Soushen (2.99%)　model ensemble

2017年　SENet (2.25%)　層

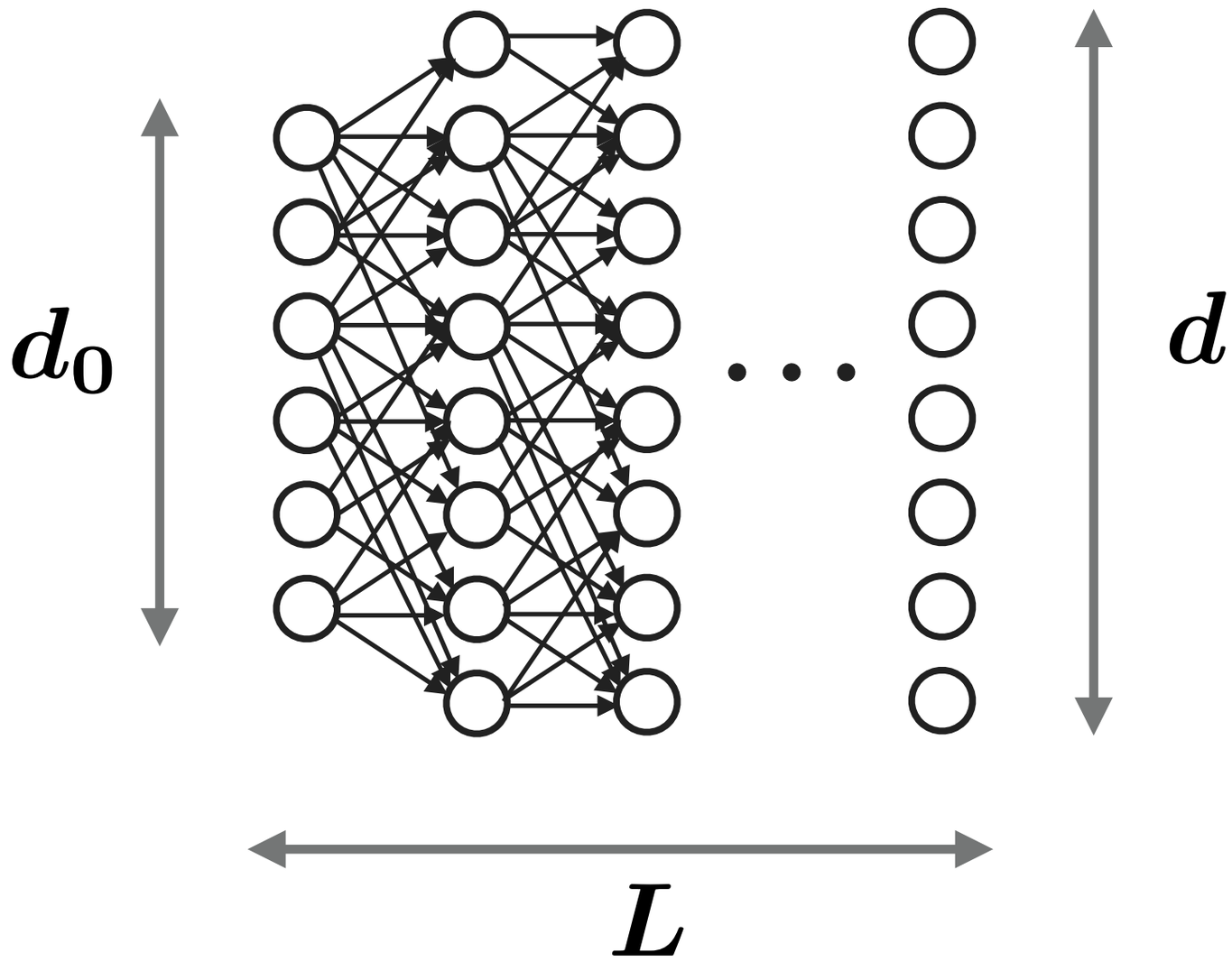
1. Depth and Width

表現能力と深さ

ニューラルネットの深さは、表現の複雑さとどのような関係にあるか

表現能力と深さ

ニューラルネットの深さは、表現の複雑さとどのような関係にあるか



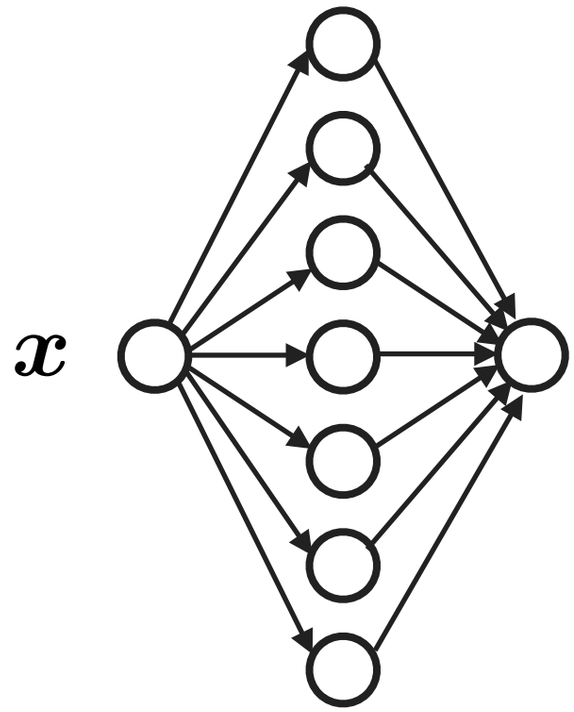
表現能力と深さ

ニューラルネットの深さは、表現の複雑さとどのような関係にあるか

$$\sim \left(\frac{d}{d_0} \right)^{L d_0} d^{d_0}$$

表現能力と深さ

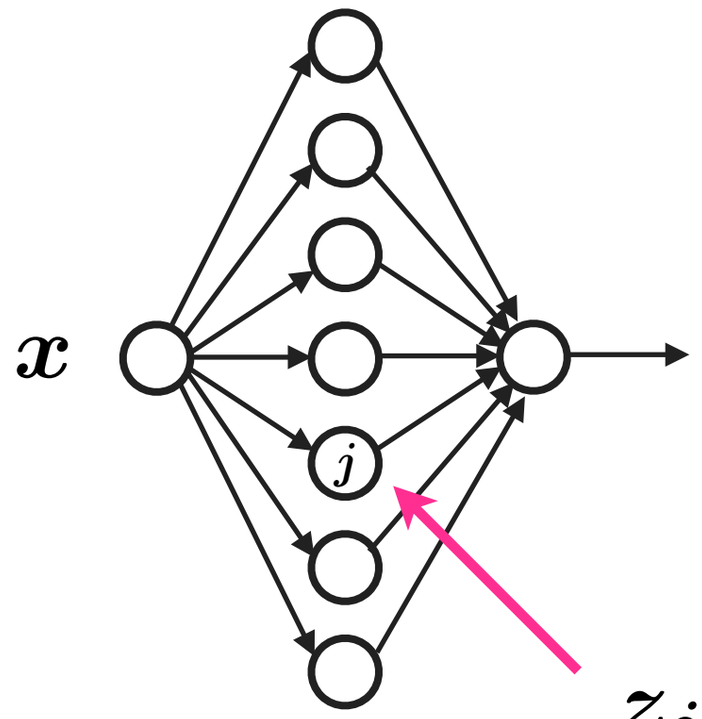
ニューラルネットの深さは、表現の複雑さとどのような関係にあるか



$$j = 1, 2, \cdots, d/d_0$$

表現能力と深さ

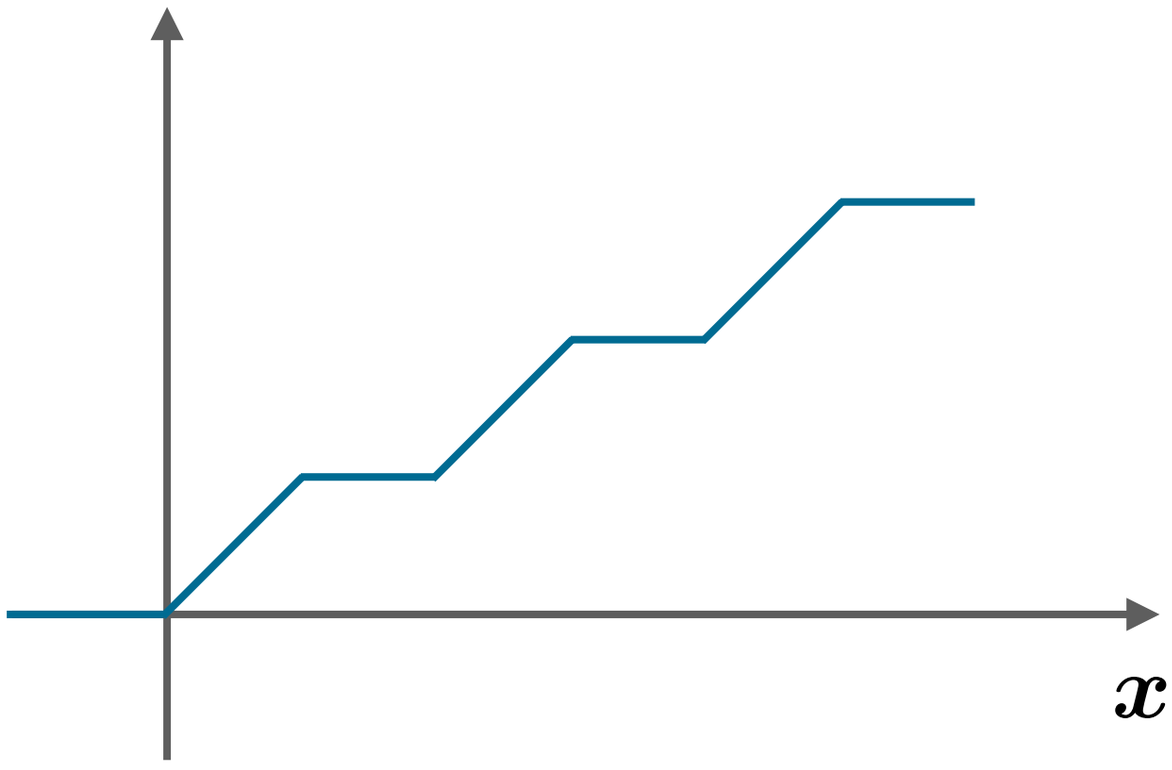
ニューラルネットの深さは、表現の複雑さとどのような関係にあるか



$j = 1, 2, \cdots, d/d_0$

表現能力と深さ

ニューラルネットの深さは、表現の複雑さとどのような関係にあるか

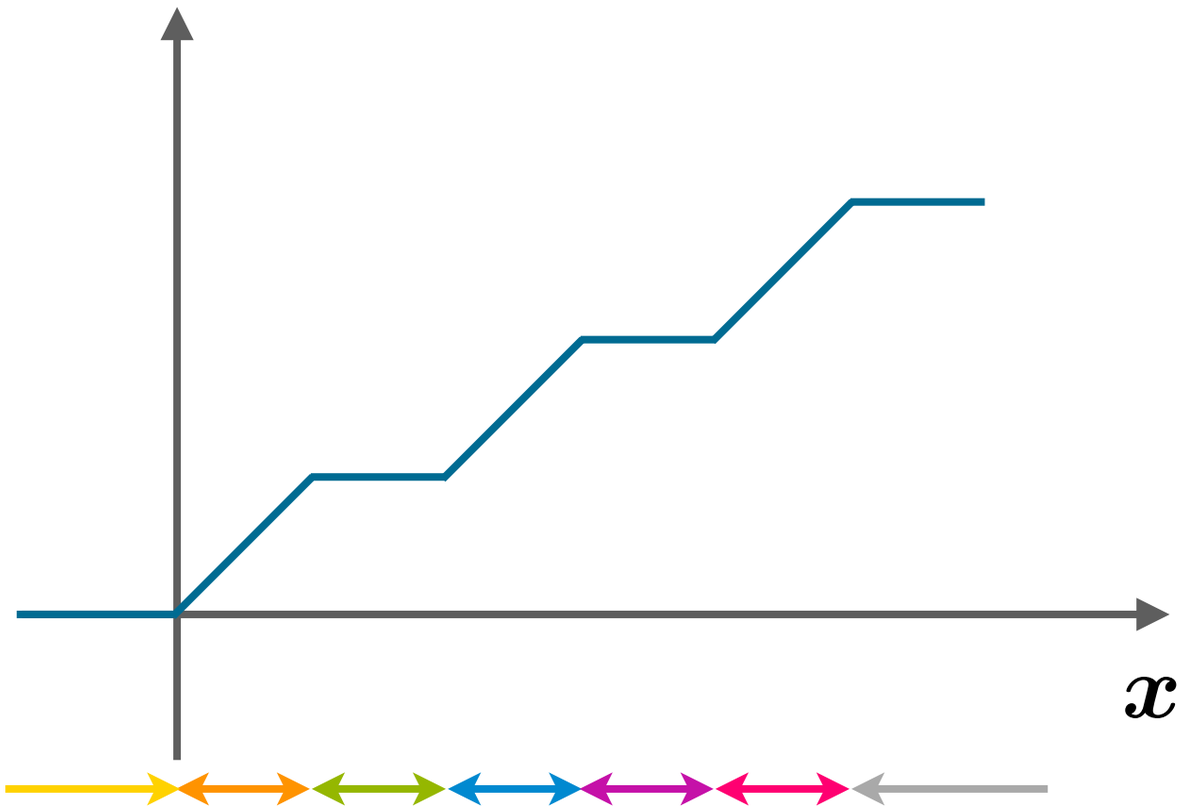


$$\sum_j (-1)^{j-1} z_j$$

$$z_j = \max(0, jx - j)$$

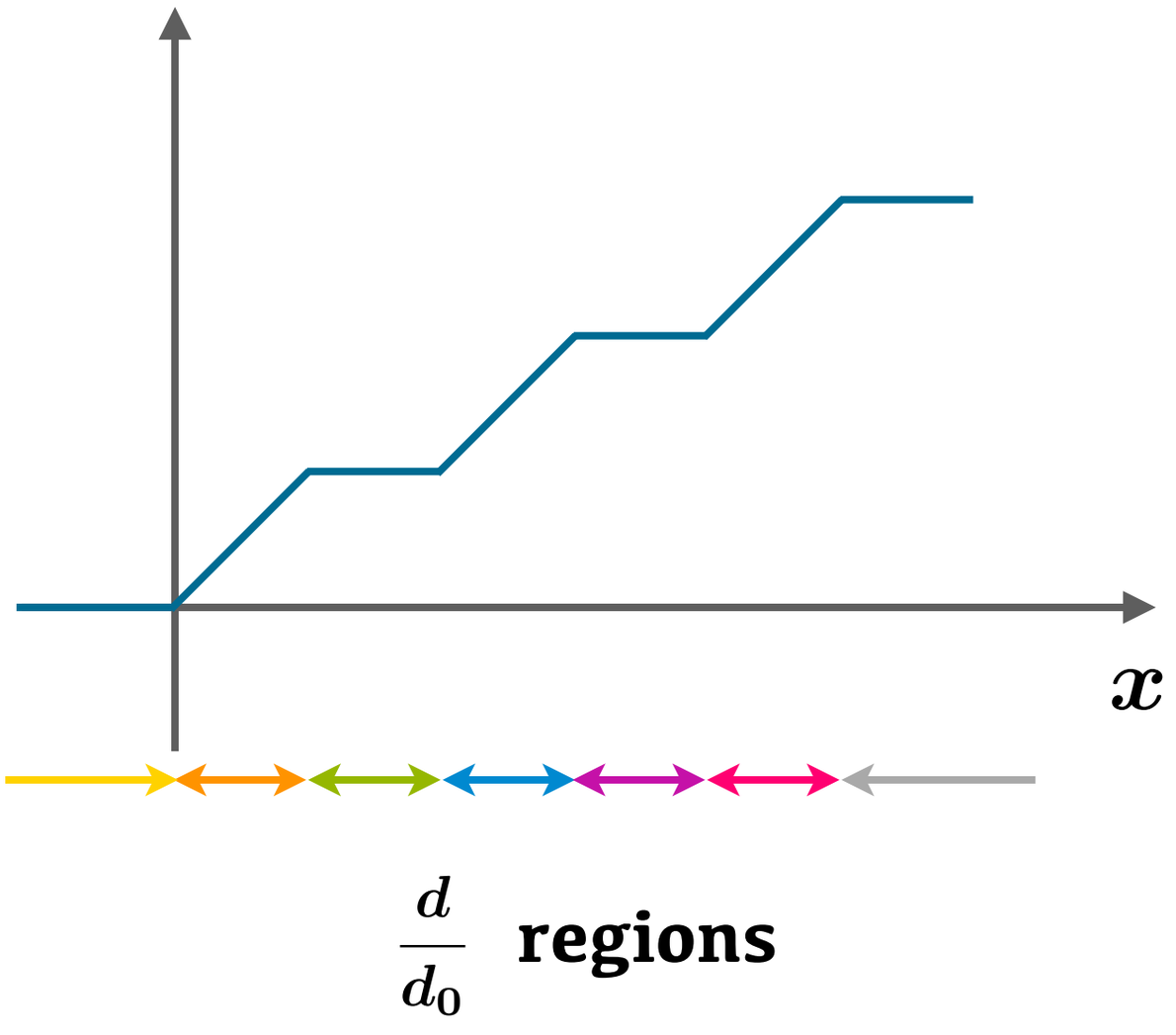
表現能力と深さ

ニューラルネットの深さは、表現の複雑さとどのような関係にあるか

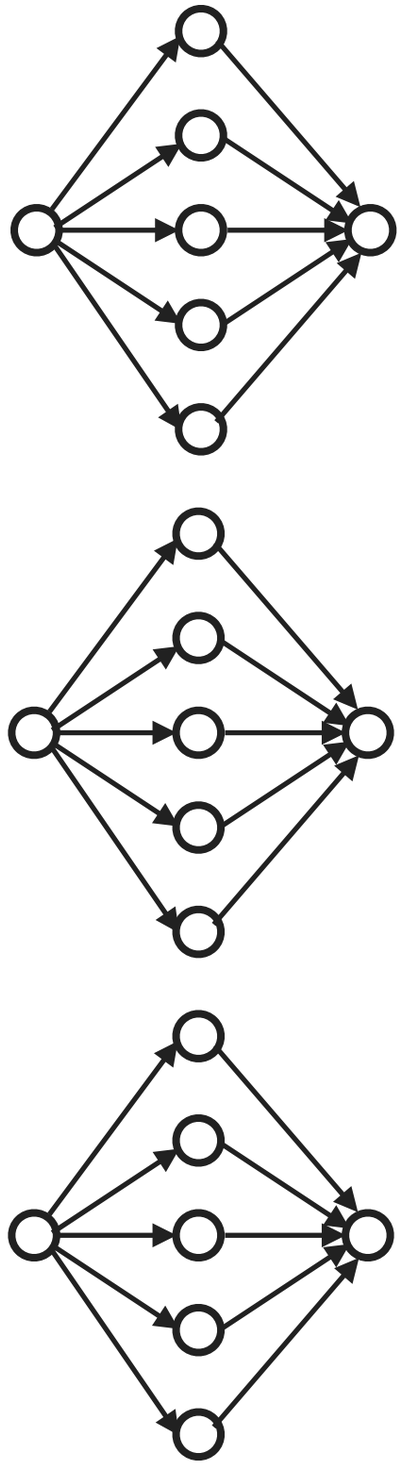


$\frac{d}{d_0}$ **regions**

表現能力と深さ

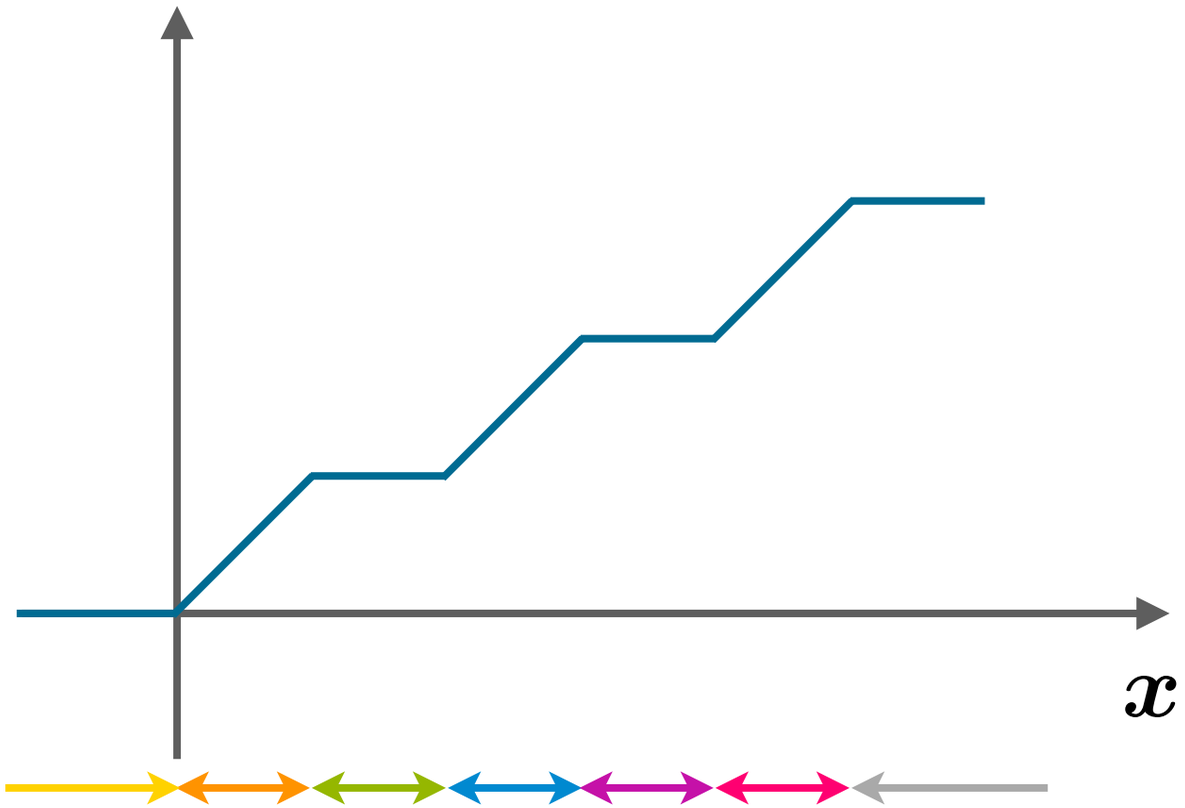


$\frac{d}{d_0}$ **regions**



d_0 個

表現能力と深さ



$\frac{d}{d_0}$ **regions**

d_0 **個** → $\left(\frac{d}{d_0}\right)^{d_0}$ **個の超立方体領域へ分割**

勾配消失・爆発問題

単純に多層にすると、学習に問題が生じることが古くから知られていた [S. HochreiterのD論, 1991]

勾配消失・爆発問題

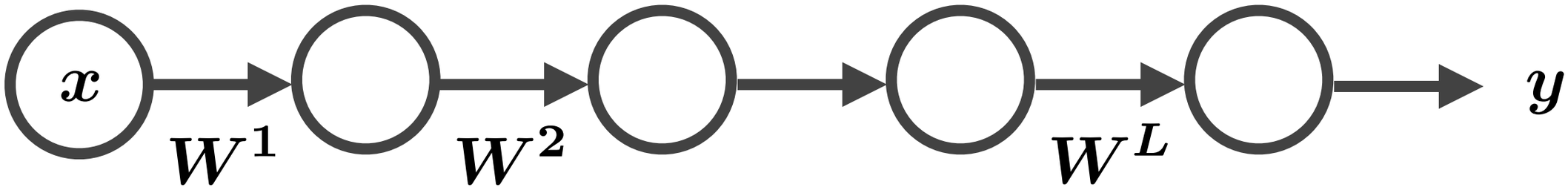
単純に多層にすると、学習に問題が生じることが古くから知られていた [S. HochreiterのD論, 1991]

勾配降下法

$$W^{\ell} \leftarrow W^{\ell} - \eta \frac{\partial E}{\partial W^{\ell}}$$

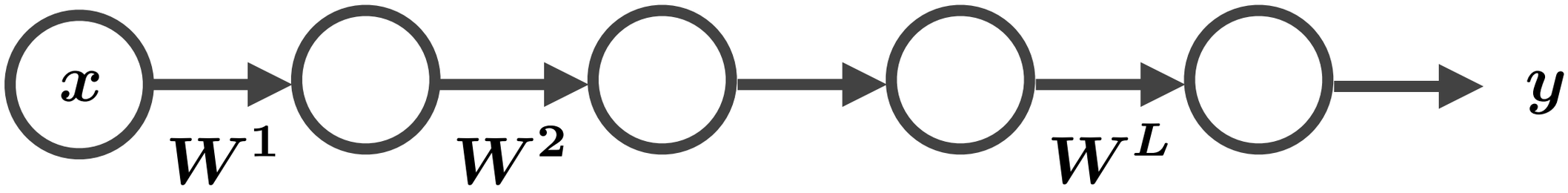
消失・爆発

勾配消失・爆発問題 - 誤差逆伝播法



$$y = f(W^L f(W^{L-1} f(\cdots W^2 f(W^1 x))))$$

勾配消失・爆発問題 - 誤差逆伝播法

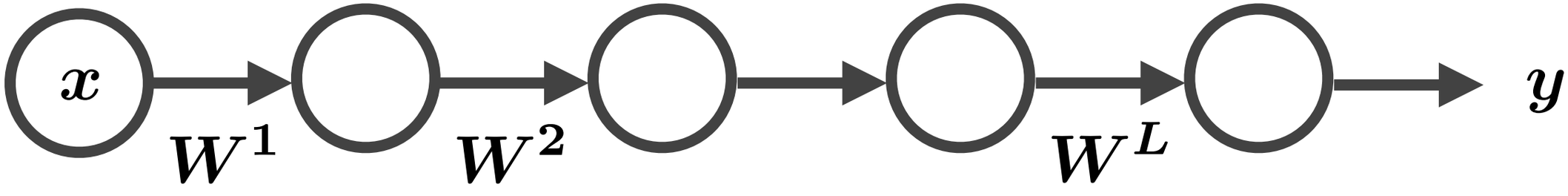


$$y = f(W^L f(W^{L-1} f(\cdots W^2 f(W^1 x))))$$

勾配降下法

$$W^\ell \leftarrow W^\ell - \eta \frac{\partial E}{\partial W^\ell}$$

勾配消失・爆発問題 - 誤差逆伝播法

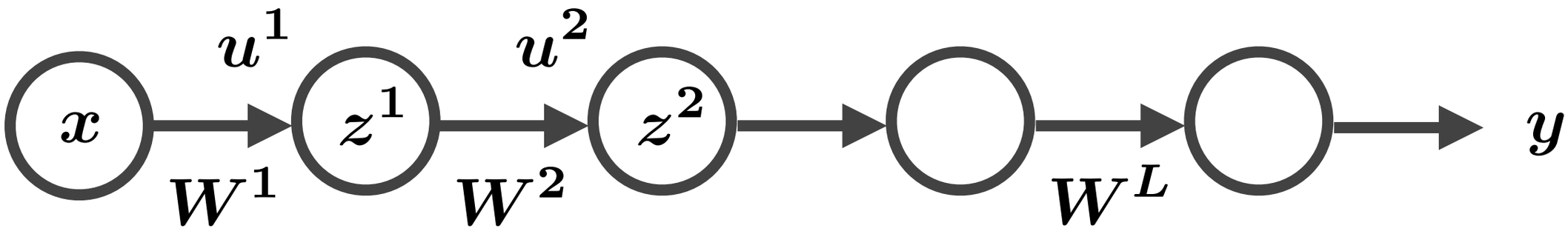


$$y = f(W^L f(W^{L-1} f(\cdots W^2 f(W^1 x))))$$

$$E = E(y(W's))$$

$$\frac{\partial E(W)}{\partial W^2}$$

勾配消失・爆発問題 - 誤差逆伝播法



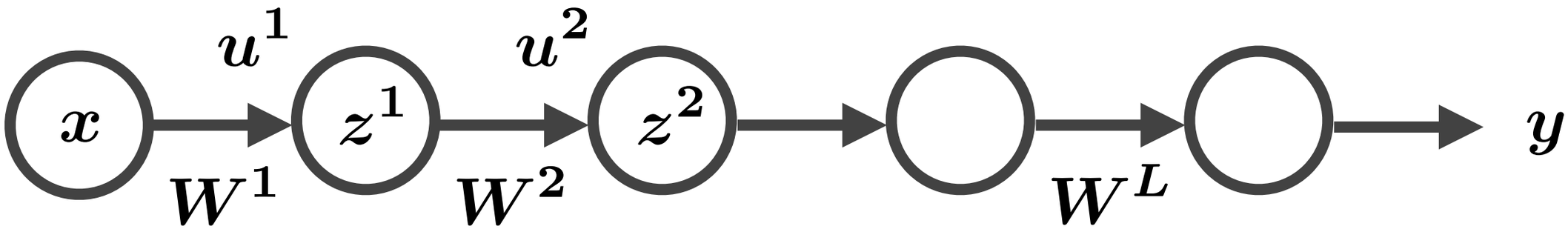
$$y = f(W^L f(W^{L-1} f(\cdots W^2 f(W^1 x))))$$

$$u^\ell = W^\ell z^{\ell-1}$$

$$z^\ell = f(u^\ell)$$

$$x = z^0, \quad y = z^L$$

勾配消失・爆発問題 - 誤差逆伝播法



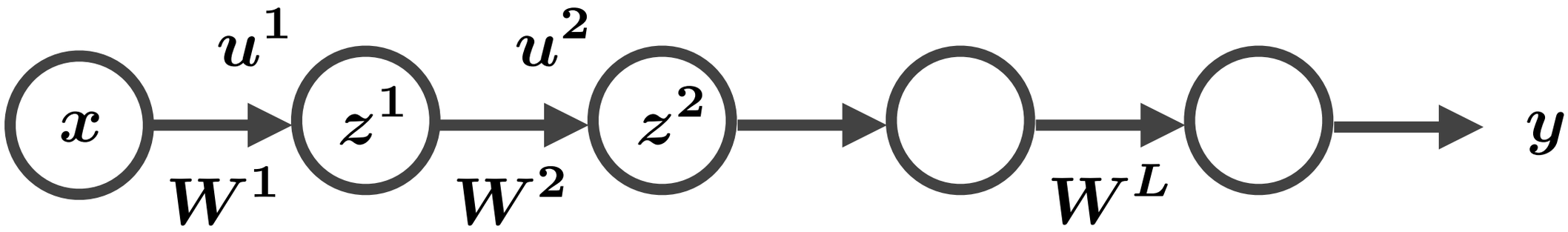
$$u^\ell = W^\ell z^{\ell-1}$$

$$z^\ell = f\left(u^\ell\right)$$

$$x = z^0, \quad y = z^L$$

$$\frac{\partial E}{\partial W^\ell} = \frac{\partial E}{\partial u^\ell}\frac{\partial u^\ell}{\partial W^\ell} = \frac{\partial E}{\partial u^\ell} z^{\ell-1}$$

勾配消失・爆発問題 - 誤差逆伝播法



$$u^\ell = W^\ell z^{\ell-1}$$

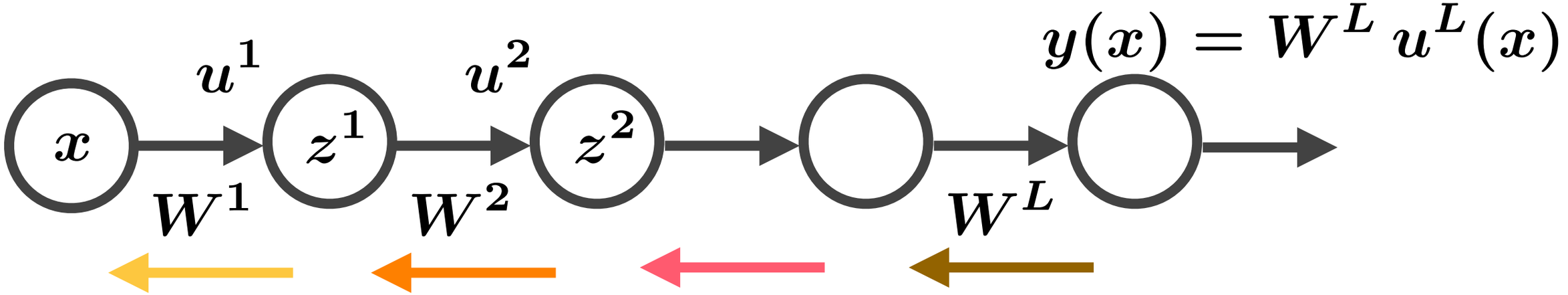
$$z^\ell = f\left(u^\ell\right)$$

$$x = z^0, \quad y = z^L$$

$$\frac{\partial E}{\partial W^\ell} = \frac{\partial E}{\partial u^\ell}\frac{\partial u^\ell}{\partial W^\ell} = \frac{\partial E}{\partial u^\ell} z^{\ell-1}$$

$$\frac{\partial E}{\partial u^\ell} = \frac{\partial E}{\partial u^{\ell+1}}\frac{\partial u^{\ell+1}}{\partial u^\ell} = \frac{\partial E}{\partial u^{\ell+1}} W^{\ell+1} f'\left(u^\ell\right)$$

勾配消失・爆発問題 - 誤差逆伝播法

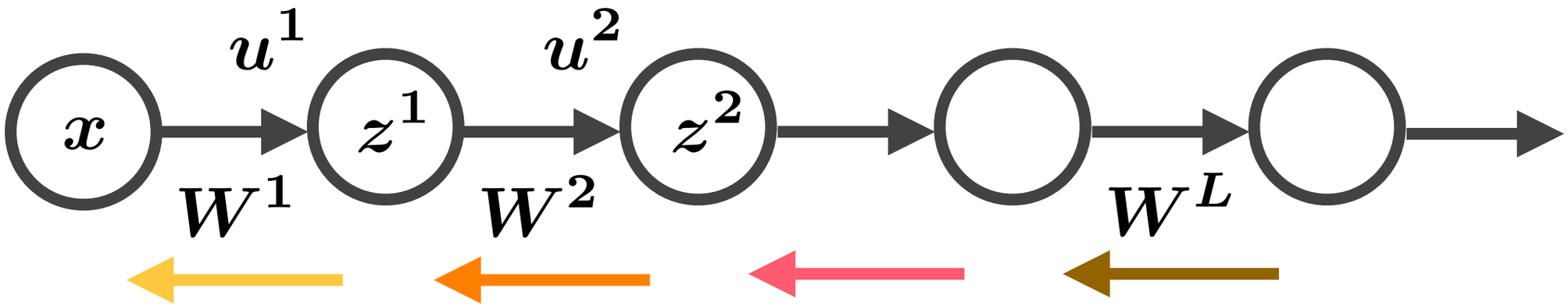


$$\frac{\partial E}{\partial u^L} = \sum_n \left(y(x^{(n)}) - y^{(n)}\right)$$

$$\frac{\partial E}{\partial W^\ell} = \frac{\partial E}{\partial u^\ell}\frac{\partial u^\ell}{\partial W^\ell} = \frac{\partial E}{\partial u^\ell} z^{\ell-1}$$

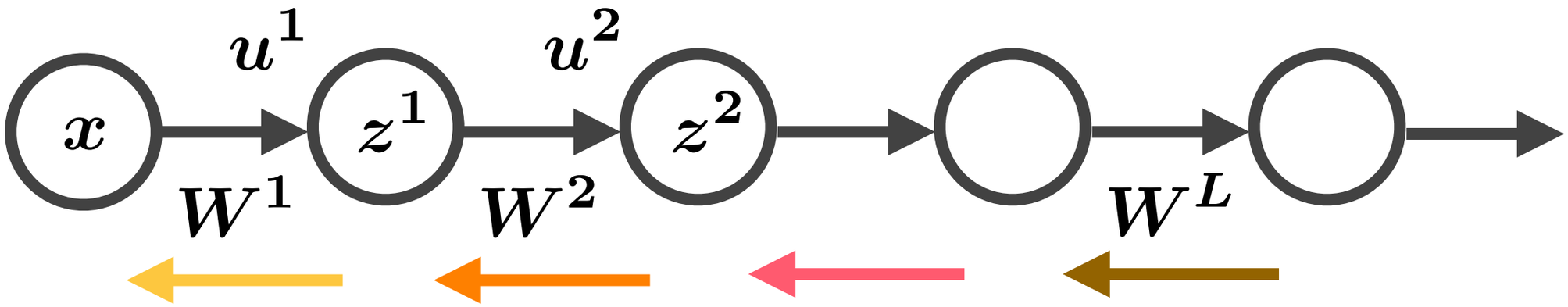
$$\frac{\partial E}{\partial u^\ell} = \frac{\partial E}{\partial u^{\ell+1}}\frac{\partial u^{\ell+1}}{\partial u^\ell} = \frac{\partial E}{\partial u^{\ell+1}} W^{\ell+1} f'(u^\ell)$$

勾配消失・爆発問題 - 誤差逆伝播法



$$\frac{\partial E}{\partial u^{\ell}} = \frac{\partial E}{\partial u^{\ell+1}} W^{\ell+1} f'\left(u^{\ell}\right)$$

勾配消失・爆発問題 - 誤差逆伝播法



$$\frac{\partial E}{\partial u^{\ell}} = \frac{\partial E}{\partial u^{\ell+1}} W^{\ell+1} f'\left(u^{\ell}\right)$$

初期値を工夫

区分線形関数

$$f(u) = \max(0, u)$$

これでもう安心か？

あとは深さを追求するだけか？

「落ちぶれ」問題　degradation problem

[K.He & J.Sun, 2014]

ネットワークの深さを増やすに連れ、やがて性能が頭打ちになり、
やがては性能が下がってゆく。

過学習とは別の問題（訓練誤差も上昇）

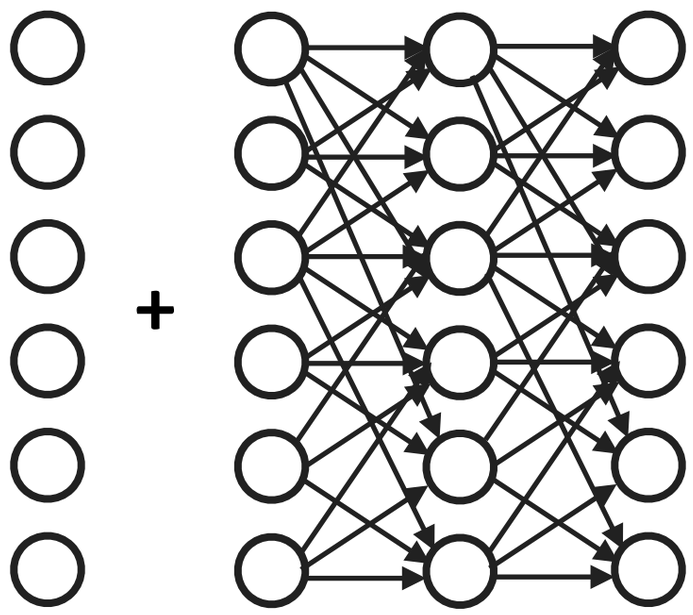
「落ちぶれ」問題　degradation problem

[K.He & J.Sun, 2014]

ネットワークの深さを増やすに連れ、やがて性能が頭打ちになり、
やがては性能が下がってゆく。

過学習とは別の問題（訓練誤差も上昇）

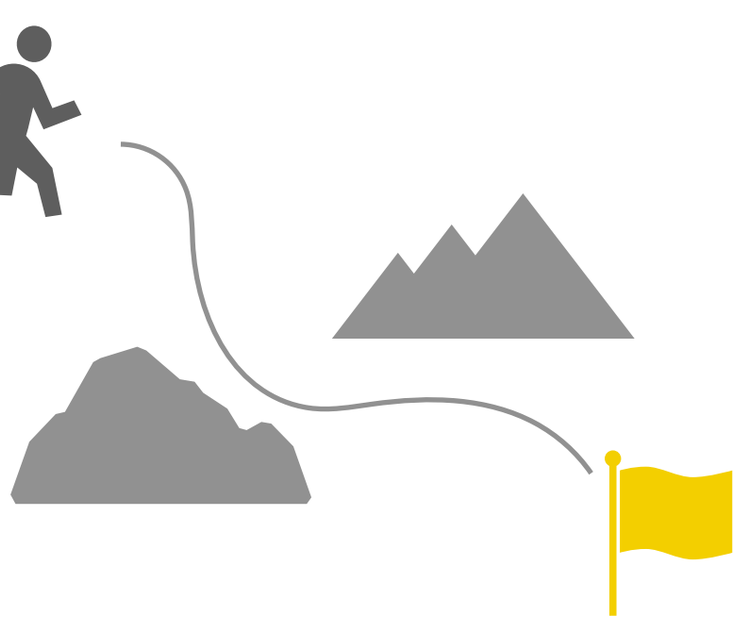
性能が下がるのは不思議

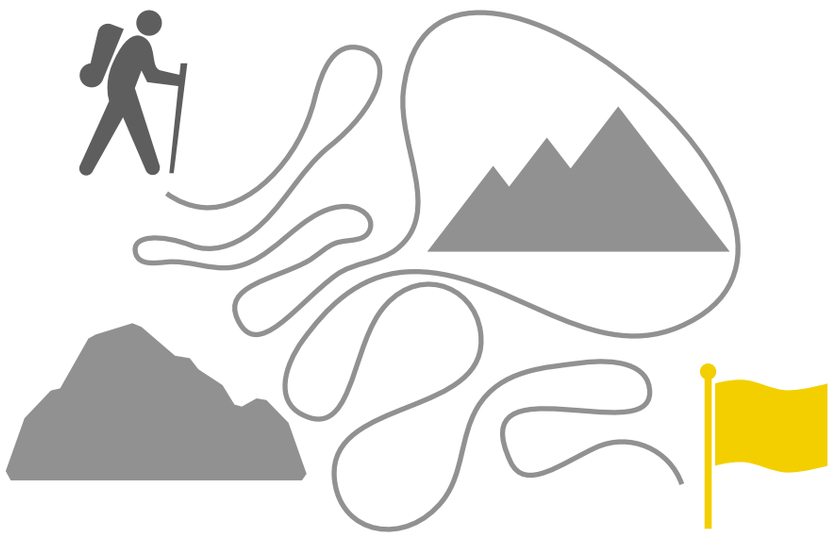


「落ちぶれ」問題　degradation problem

[K.He & J.Sun, 2014]

収束に要する時間が指数関数的に増大？





恒等写像を学ぶのは困難？　勾配消失？？

「落ちぶれ」問題　degradation problem

[K.He & J.Sun, 2014]

いずれにせよ深層ネットワークとすることは、それほど簡単なことではない

→ **residual neural network (ResNet)**

2. ResNet

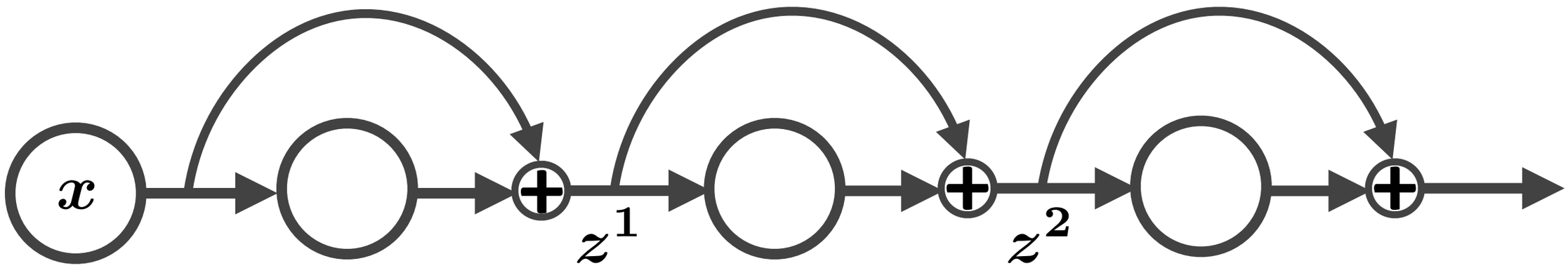
Residual Network [K.He, X.Zhang, S.Ren & J.Sun, 2015]

低層からの情報が高層で欠落することがさまざまな問題に関係？

Residual Network [K.He, X.Zhang, S.Ren & J.Sun, 2015]

低層からの情報が高層で欠落することがさまざまな問題に関係？

ショートカット結合を導入する

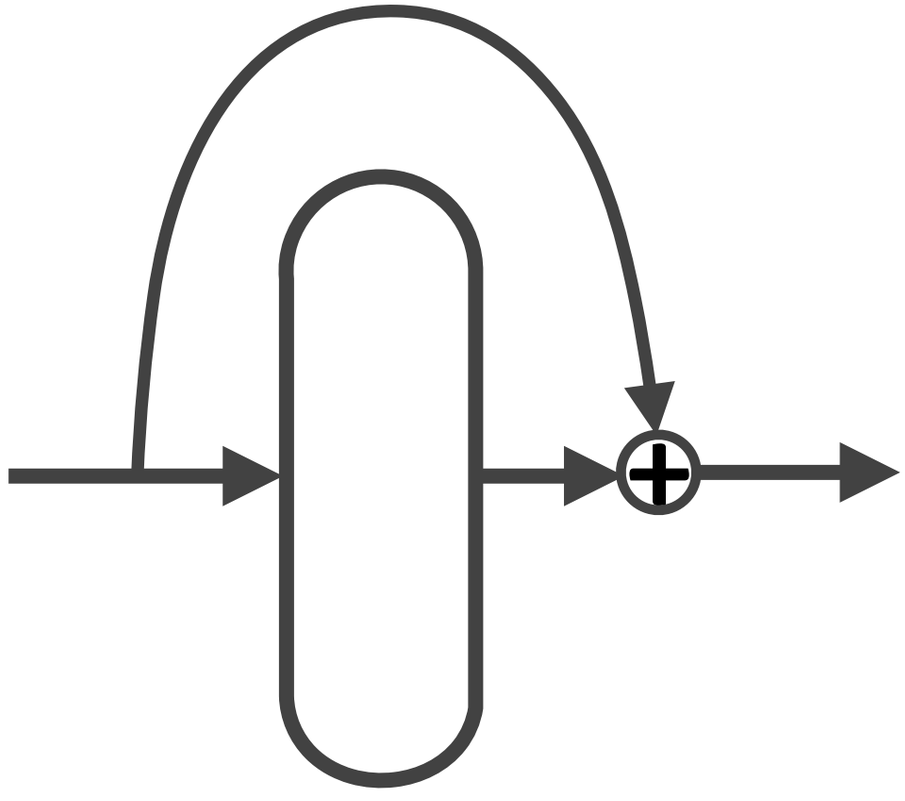


$$z^{\ell} = F(z^{\ell-1}) + z^{\ell-1}$$

Residual Network [K.He, X.Zhang, S.Ren & J.Sun, 2015]

低層からの情報が高層で欠落することがさまざまな問題に関係？

ショートカット結合を導入する



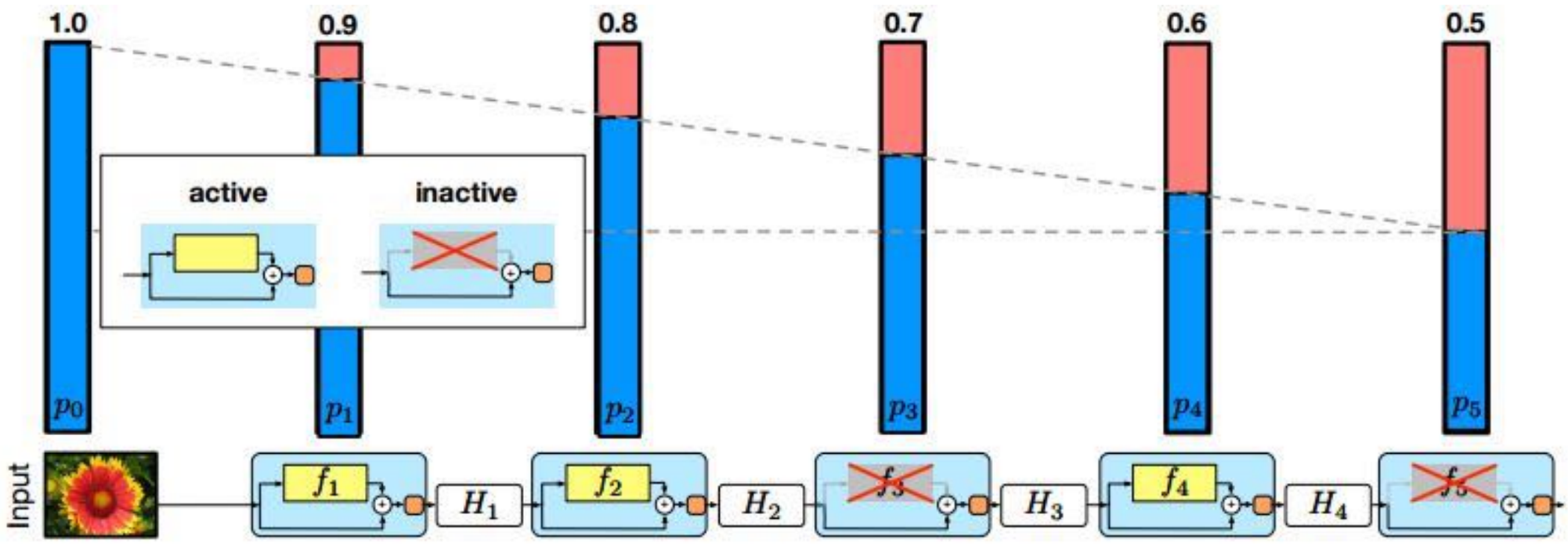
residual block

Residual Network [K.He, X.Zhang, S.Ren & J.Sun, 2015]

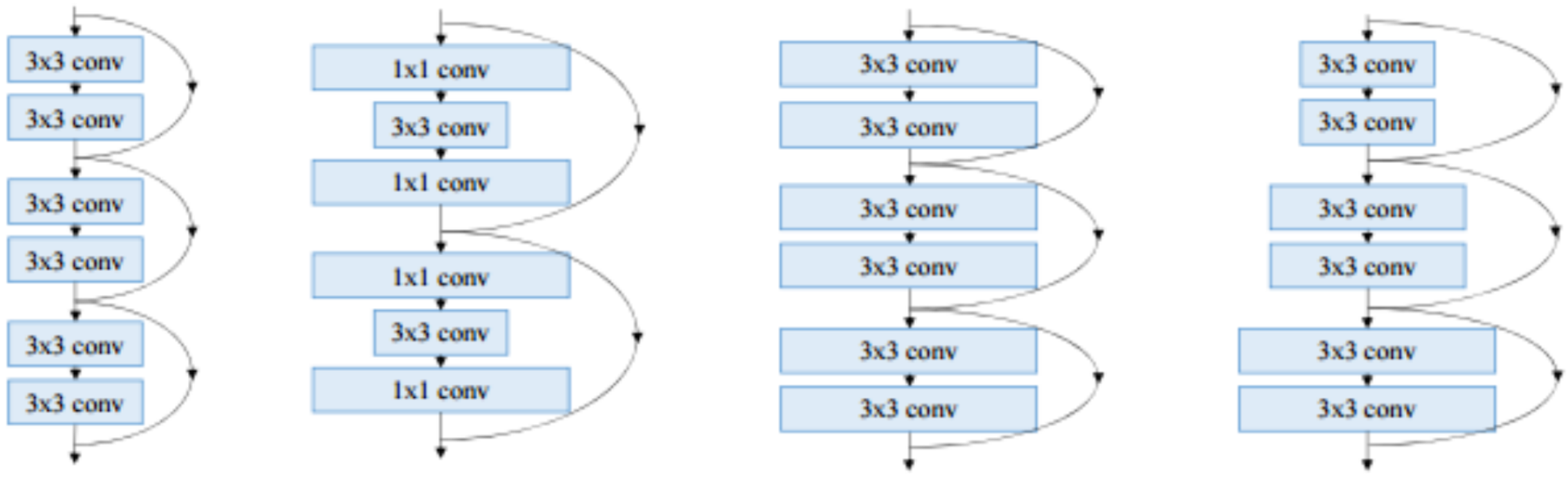
2015年時点での画像分類タスクにおけるSOTA

昨今のSOTA級モデルも、大抵はresidual構造を持つ

Stochastic Depth [G. Huang et al., 2016]



Pyramidal Residual Network [D.Han et al., 2016]

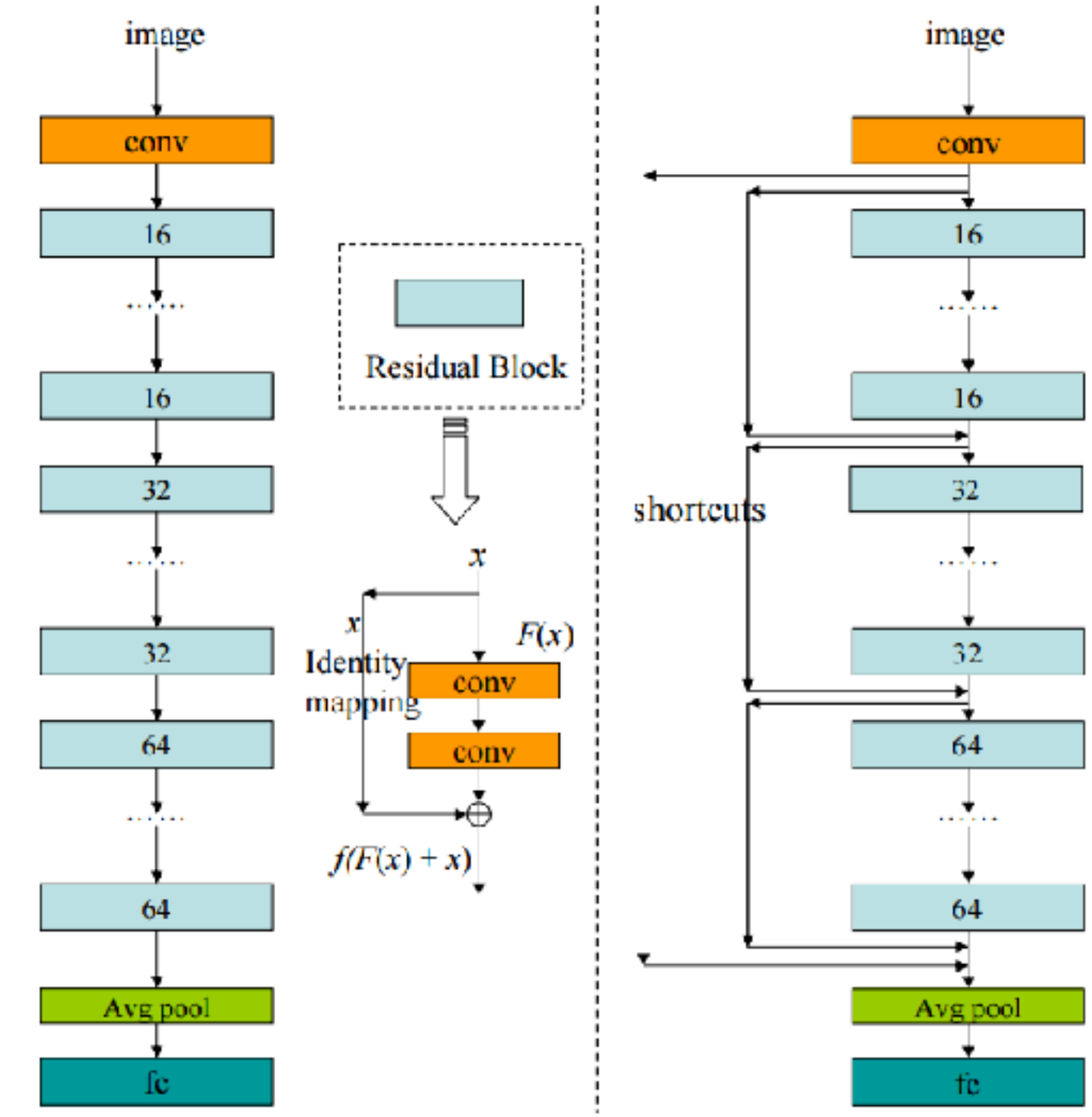


(a) basic (b) bottleneck (c) wide (d) pyramidal

Figure 1. Schematic illustration of (a) basic residual units [7], (b) bottlenecks [7], (c) wide residual units [34], and (d) our pyramidal residual units.

Residual Network of Residual Network

[Zhang et al., 2016]



Residual Networks **RoR**

ResNeXt [S.Xie et al., 2016]

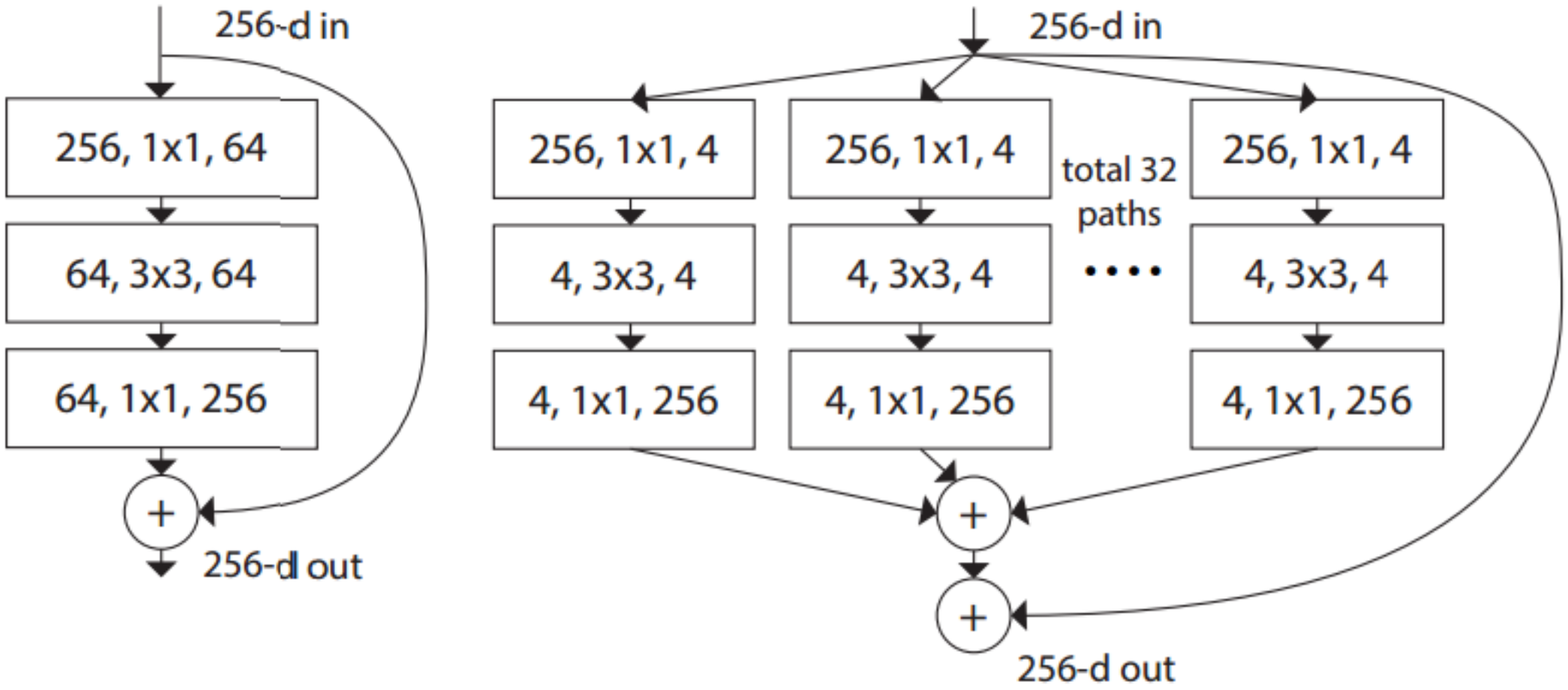
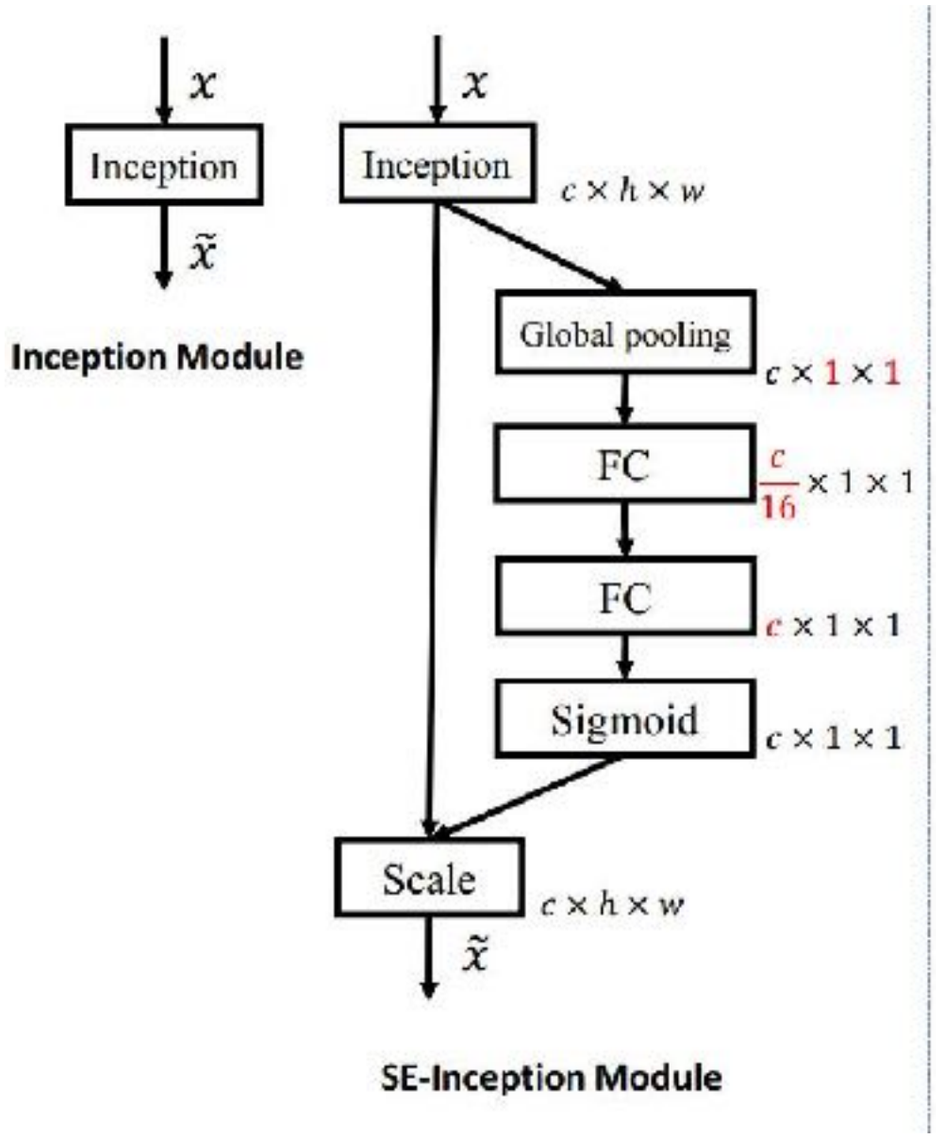
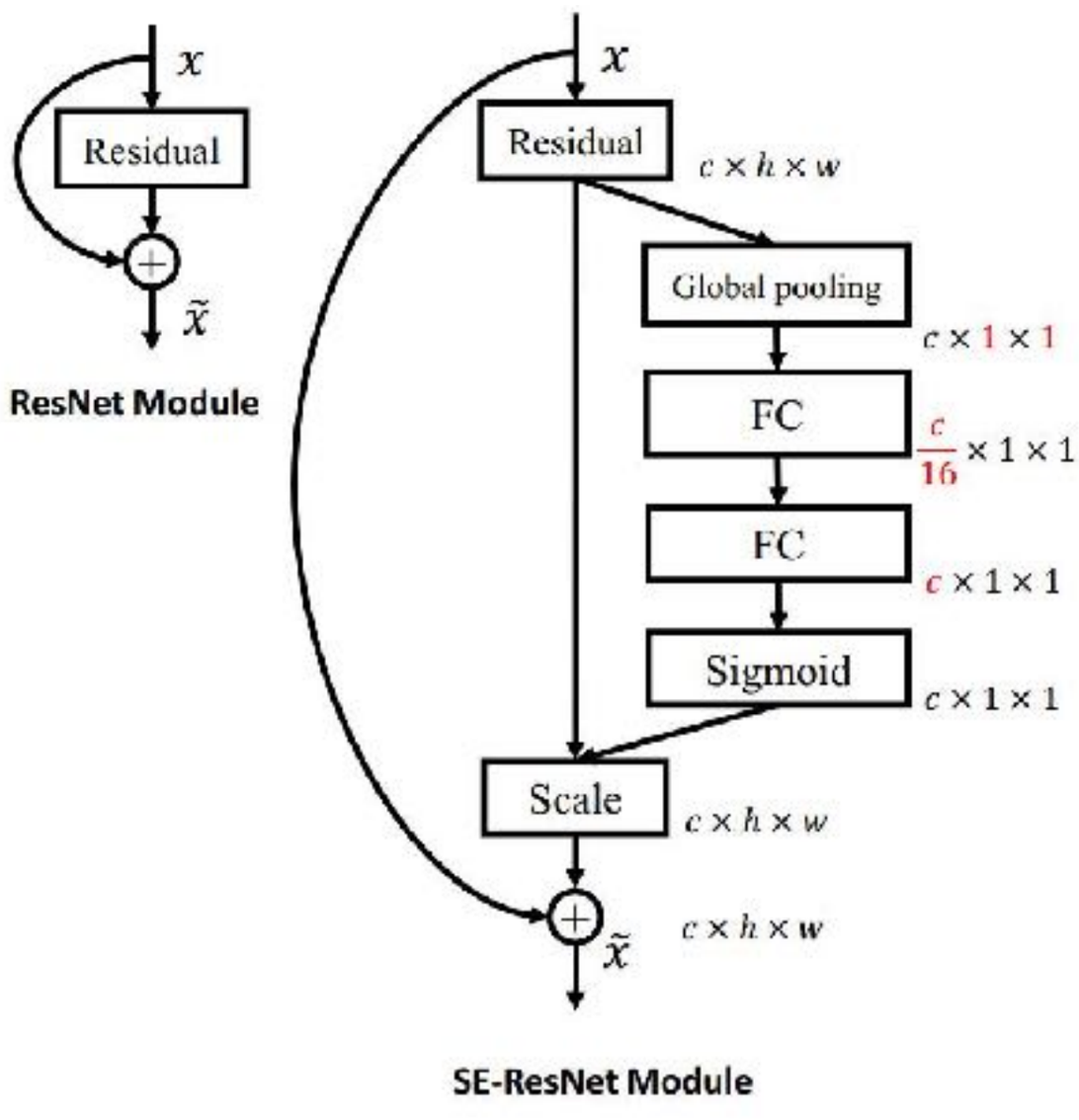


Figure 1. **Left**: A block of ResNet [13]. **Right**: A block of ResNeXt with cardinality = 32, with roughly the same complexity. A layer is shown as (# in channels, filter size, # out channels).

SENet [J.Hu et al., 2017]





これで深層化は問題なくうまくいく

しかし一見単純に見えるこのResNetというモデルが意外とまた曲者

階層的表現？ [Srivastava et al., 2015]

深層学習Folklore

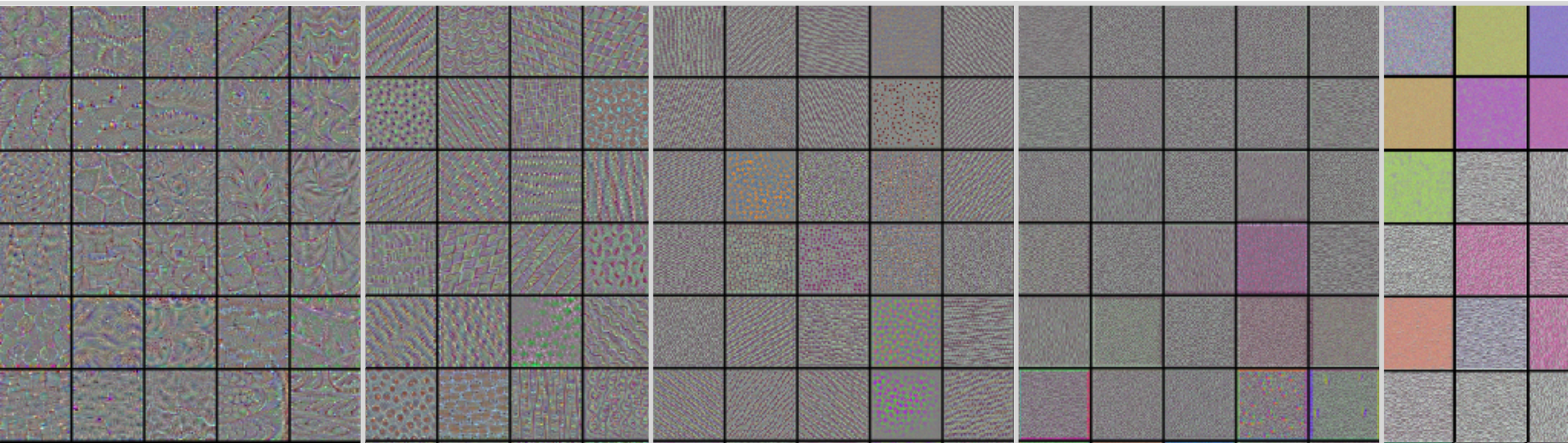
多層ニューラルネットは階層的な表現を学習することで高い認識性能を発揮

階層的表現？ [Srivastava et al., 2015]

深層学習Folklore

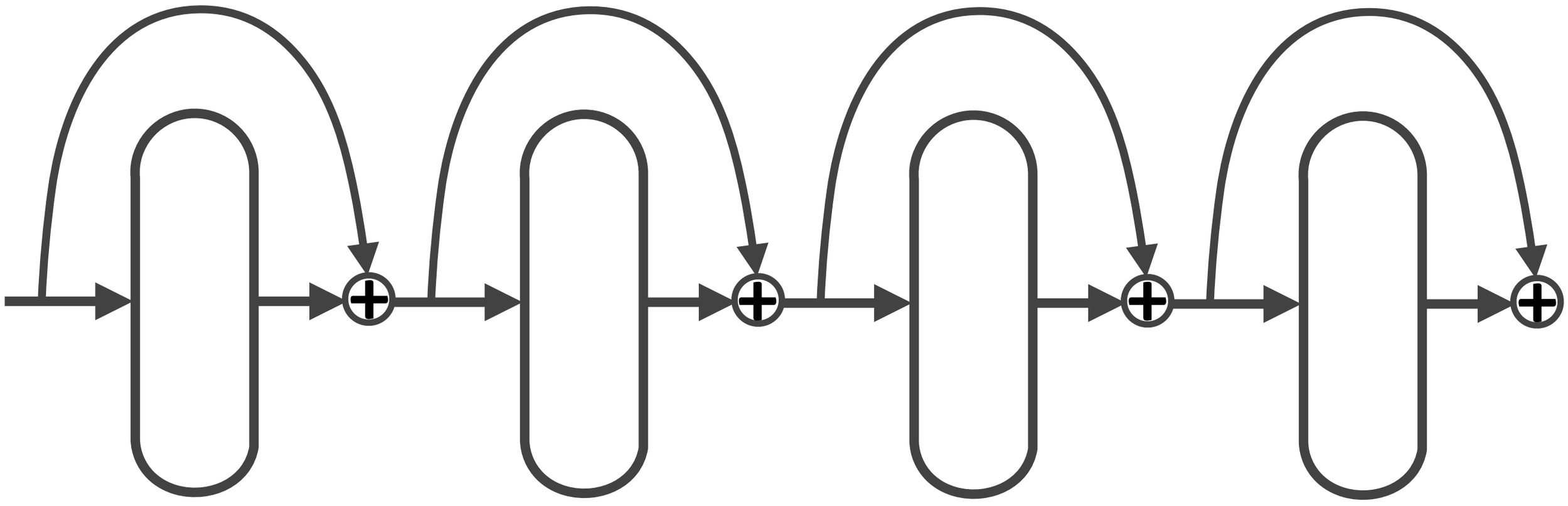
多層ニューラルネットは階層的な表現を学習することで高い認識性能を発揮

VGGの場合



階層的表現？ [Srivastava et al., 2015]

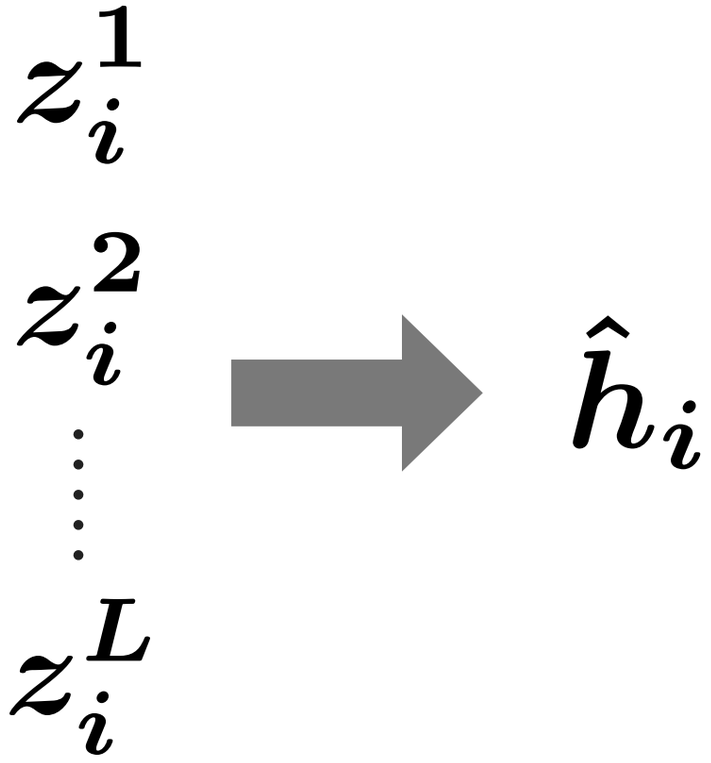
ResNetの場合



層をシャッフルしてもほとんと性能は下がらない

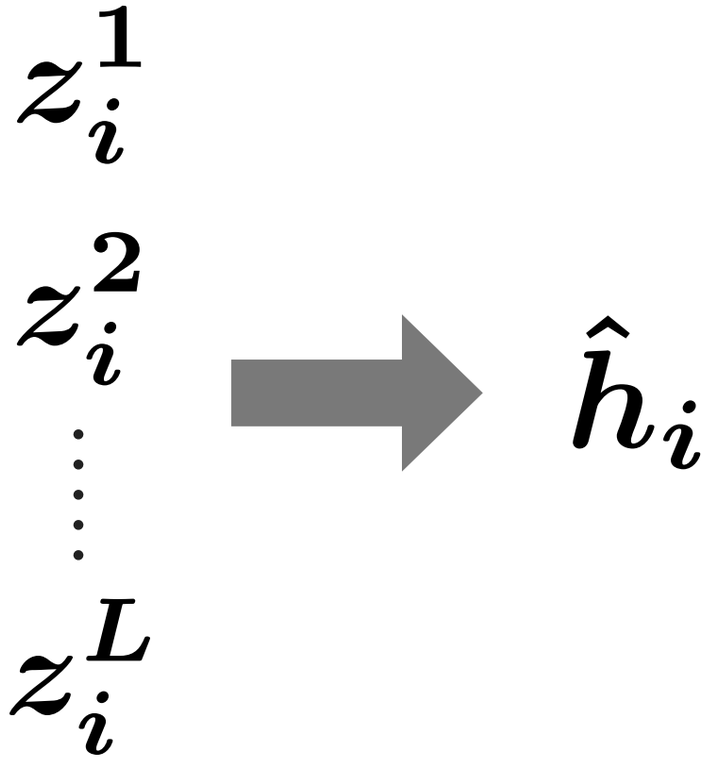
Iterative Estimator [Greff et al., 2017]

各層の出力が、同じ特徴量の不偏推定量となっている



Iterative Estimator [Greff et al., 2017]

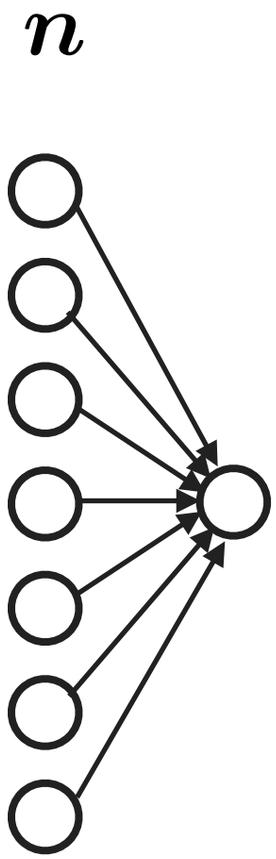
各層の出力が、同じ特徴量の不偏推定量となっている



（実はちょっと解析して見ると正しくない...）

3. Initialization

LeCun初期化

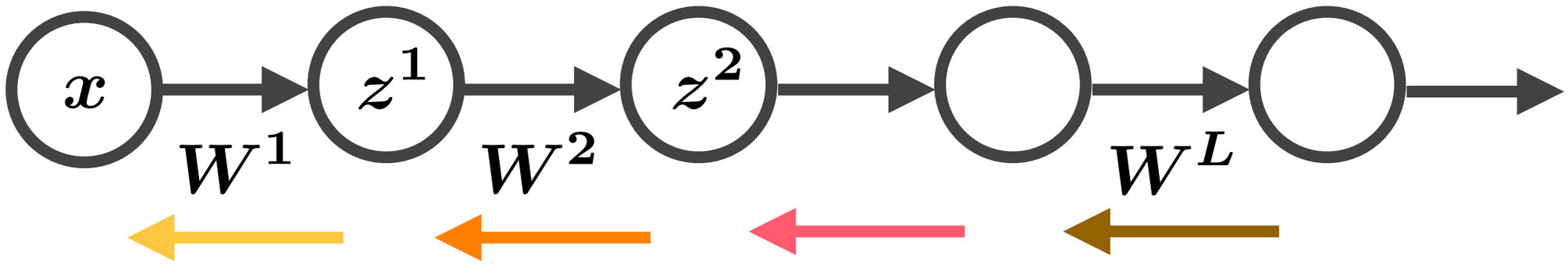


fan-inが大きすぎると信号がきすぎて出力が大きくなる。そのぶん重みを小さくしてスケール感を整える

$$\mathbf{Var}[w] = \frac{1}{n}$$

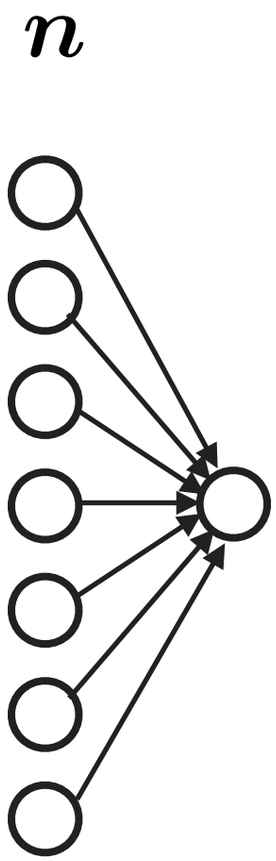
なる分布から初期値をランダムにサンプリング

勾配はどのように消失・発散するか



$$\mathbf{Var}\left[\frac{\partial E}{\partial w^{\ell}}\right] \propto \prod_{\ell} \frac{n}{2} \mathbf{Var}[w^{\ell}]$$

He初期化



勾配の値のばらつきのスケールが発散しないように、
重みのばらつきを抑える

$$\mathbf{Var}[w] = \frac{\mathbf{2}}{n}$$

なる分布から初期値をランダムにサンプリング

ResNetと初期化 [MT, 2017]

ショートカット結合はこの分散にどれほど影響するのか？

$$\mathbf{Var}\left[\frac{\partial E}{\partial w^{\ell}}\right] \propto \prod_{\ell}\left(1 + a\, n \mathbf{Var}[w^{\ell}]\right)$$

ResNetと初期化 [MT, 2017]

ショートカット結合はこの分散にどれほど影響するのか？

$$\mathbf{Var}\left[\frac{\partial E}{\partial w^{\ell}}\right] \propto \prod_{\ell}\left(1 + a\, n \mathbf{Var}[w^{\ell}]\right)$$

$$\rightarrow \left(1 + a\, n \mathbf{Var}[w]\right)^{L}$$

ResNetと初期化 [MT, 2017]

ショートカット結合はこの分散にどれほど影響するのか？

$$\mathbf{Var}\left[\frac{\partial E}{\partial w^{\ell}}\right] \propto \prod_{\ell}\left(1 + a\, n \mathbf{Var}[w^{\ell}]\right)$$

$$\rightarrow \left(1 + a\, n \mathbf{Var}[w]\right)^{L}$$

$$\approx e^{anL\mathbf{Var}[w]}$$

ResNetと初期化 [MT, 2017]

ショートカット結合はこの分散にどれほど影響するのか？

$$\mathbf{Var}[w] = \frac{c}{nL} \longrightarrow \mathbf{Var}\left[\frac{\partial E}{\partial w^{\ell}}\right] \propto e^{ac}$$

ResNetと初期化 [MT, 2017]

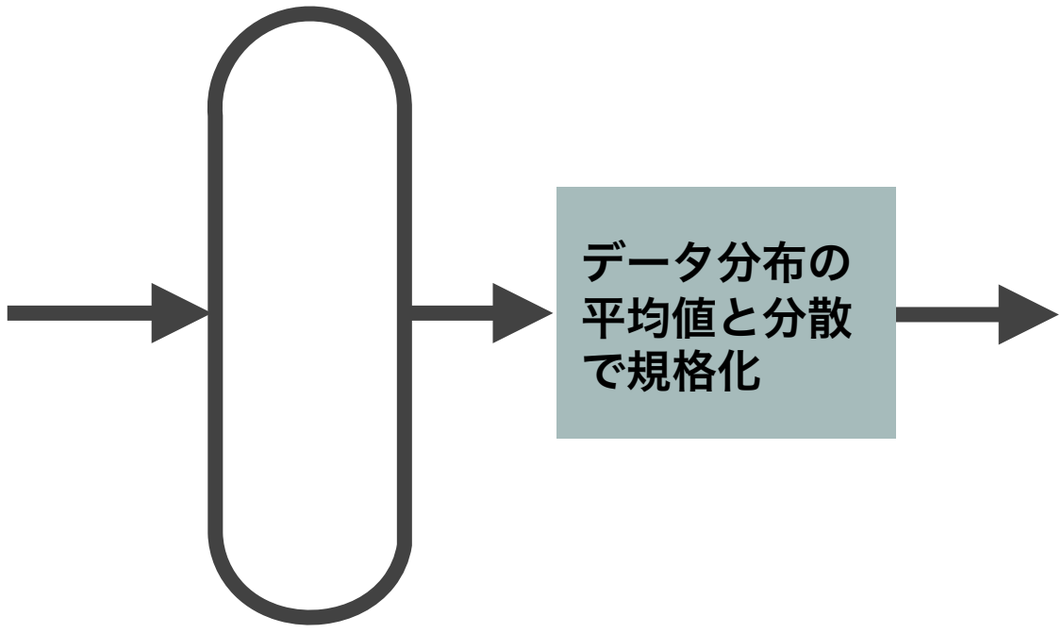
ショートカット結合はこの分散にどれほど影響するのか？

$$\mathbf{Var}[w] = \frac{c}{nL} \longrightarrow \mathbf{Var}\left[\frac{\partial E}{\partial w^{\ell}}\right] \propto e^{ac}$$

プレーンなニューラルネット

$$\mathbf{Var}[w] = \frac{b}{n} \longrightarrow \mathbf{Var}\left[\frac{\partial E}{\partial w^{\ell}}\right] \propto b^{L}$$

Batch規格化　[S.Ioffe & C.Szegedy, 2015]



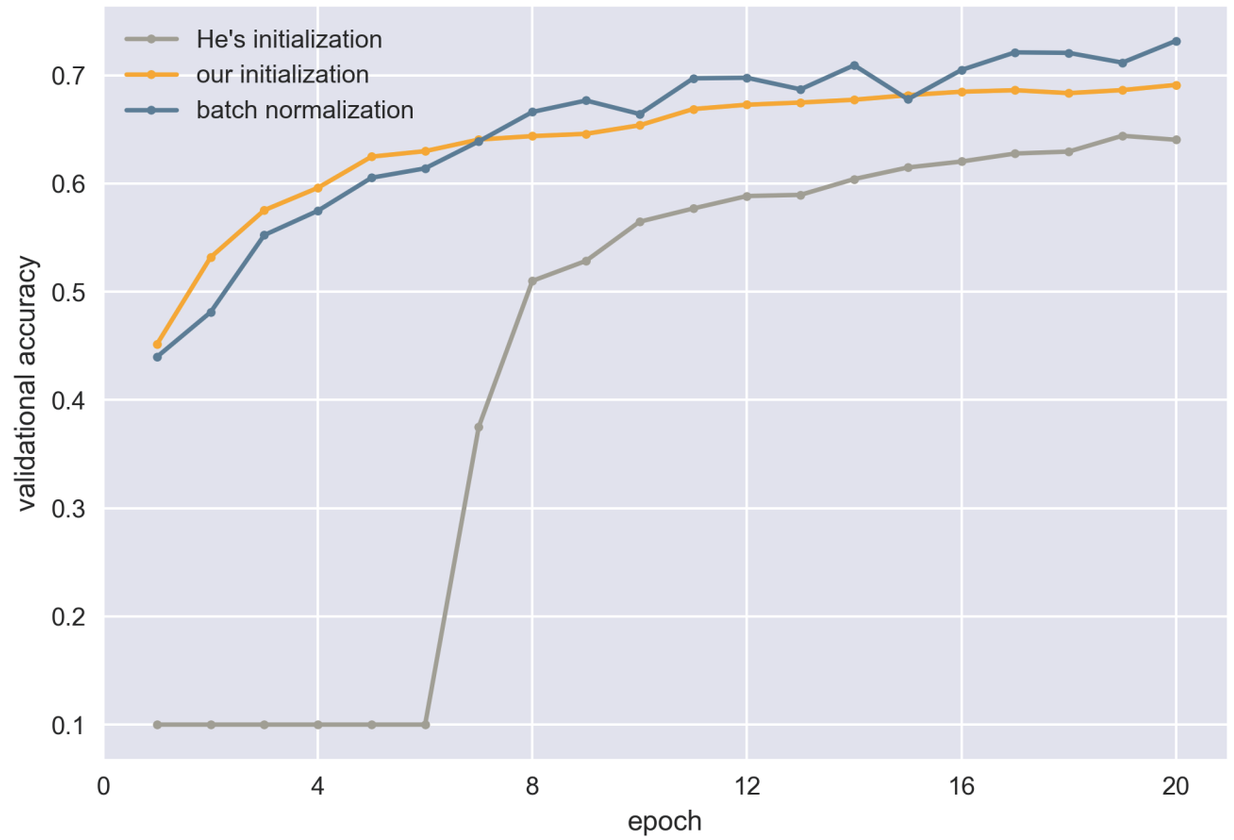
＊ 平均値と分散は、ミニバッチに関する統計量で代用する

Batch規格化の効果 [MT, 2017]

$$\mathbf{Var}\left[\frac{\partial E}{\partial w^{\ell}}\right] = \frac{L}{\ell}\mathbf{Var}[\delta_y]$$

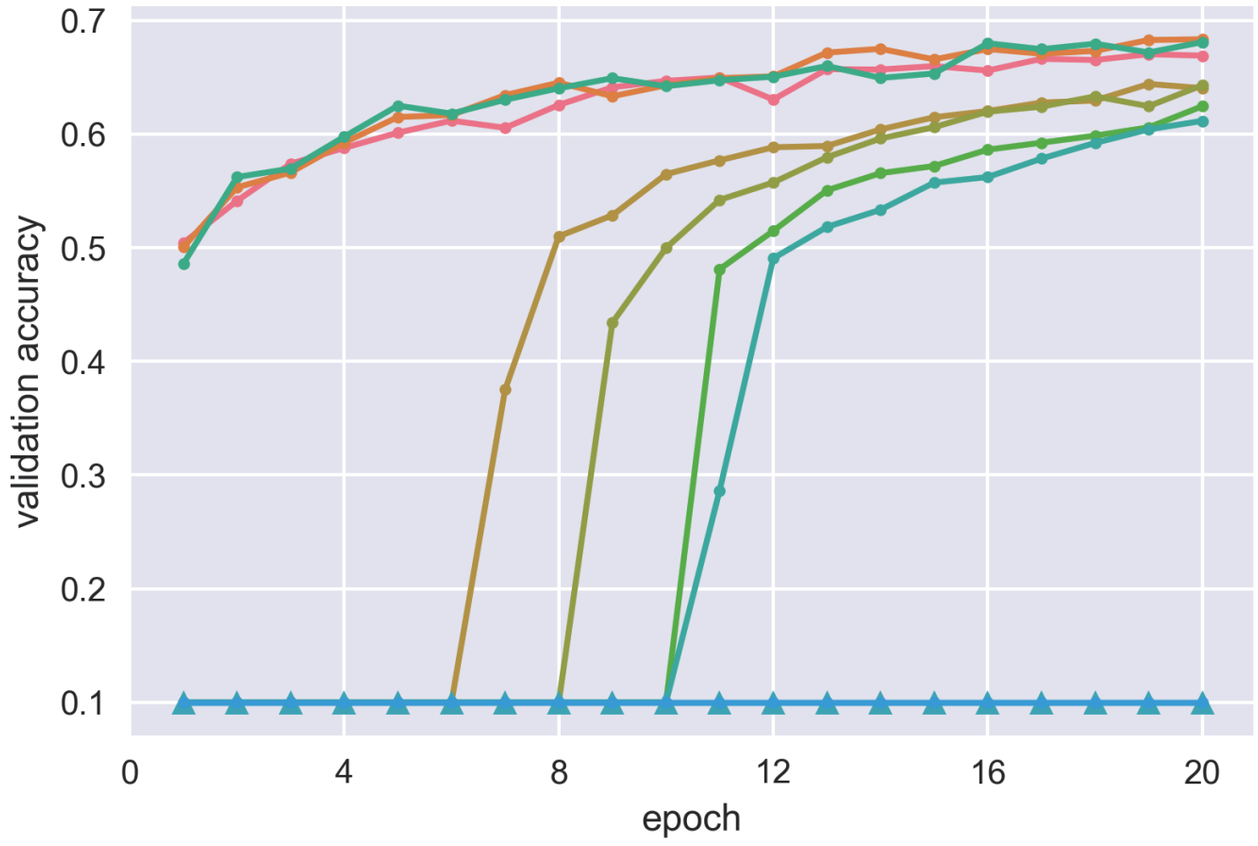
簡単な実験での検証 [MT, 2017]

かなり単純な100ブロックResNetで実験



簡単な実験での検証 [MT, 2017]

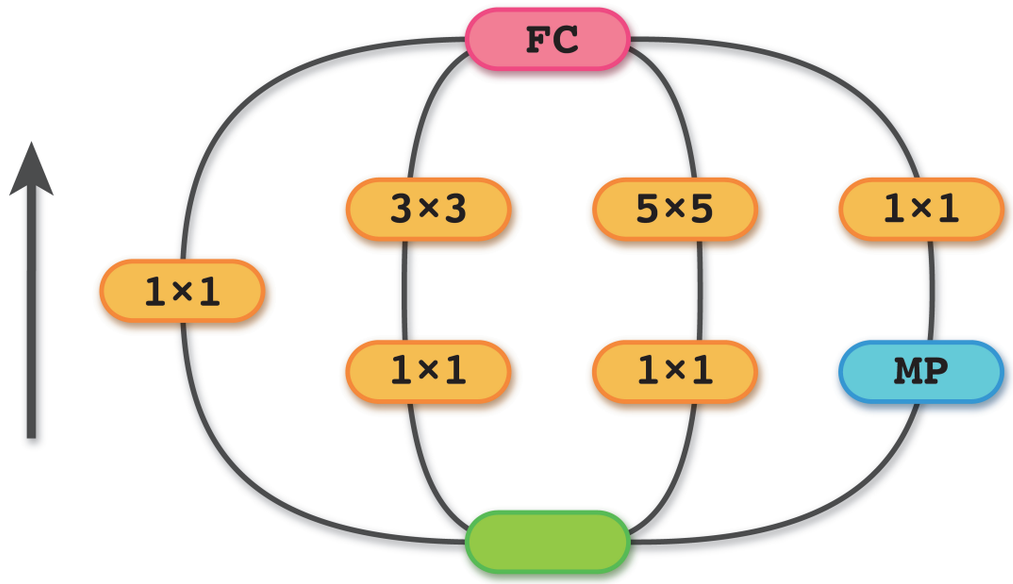
batch規格化なしでHe初期化を使うと、学習が極めて不安定



4. Further Problems

さまざまなモデルデザインにおける深さ？

e.g. Inceptionモジュール



深さは本当に表現能力に効いている？

[C.Zhang et al., 2016]

ドメイン全体の分割ではなく、有限データセットのフィッティングを考えると

サイズNのd次元データセットを学習できるL層ReLUニューラルネットで、幅$\mathcal{O}(N/L)$、重み$\mathcal{O}(N+d)$個のモデルが存在する。

多くの未解決問題は深さがやはり鍵

訓練可能性（局所最適解の問題）

表現能力の高さ

汎化可能性

モデルを「深くする」と性能が上がる、という時の「深さ」
の意味がもっとわかればいいな、と思う。

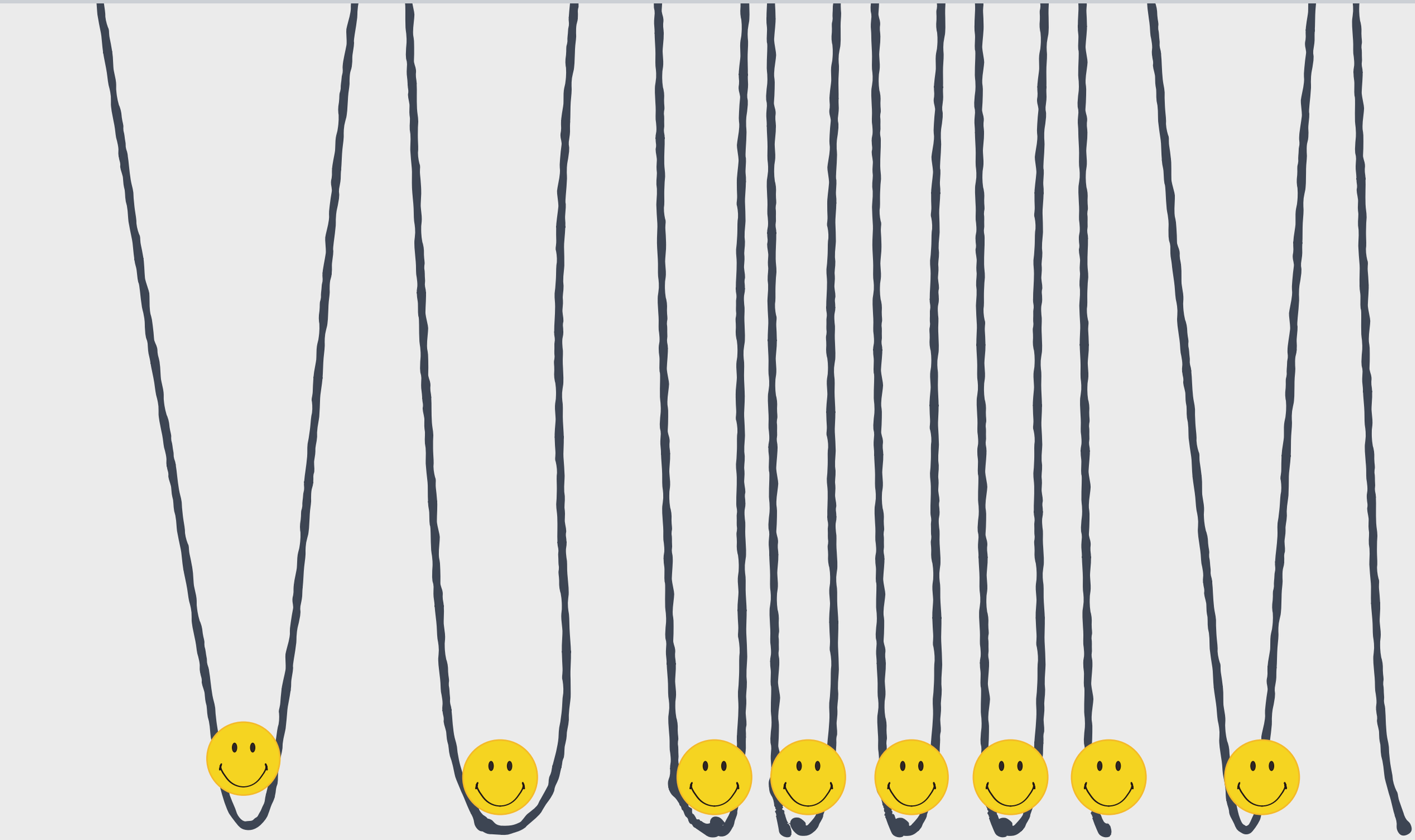
Fin

局所最適解と深層化


エラー
POISON
POISON
POISON
POISON
POISON
POISON
POISON
POISON
POISON
POISON

観察事実：深くすれば良い
なぜか？

局所最適解と深層化



[Choromanska-Henaff-Mathieu-Arous-LeCun, 2015]