視覚情報処理モデルとしての深層学習

浅川伸一 asakawa@ieee.org

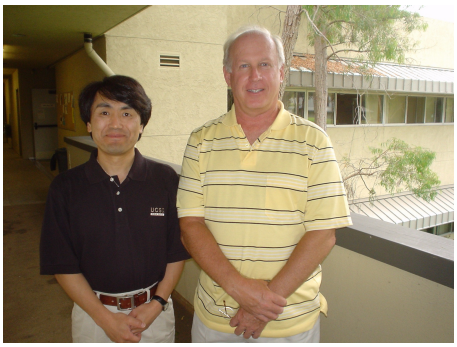
2020-03-20

- 1 謝辞
- 2 自己紹介
- 3 時間配分
- 4 要旨
- 5 Takeaways
- 6 本日お話しないこと
- 7 キーコンセプト
- 8 Van Essen らの視覚野地図
- 9 Hubel and Wiesel
- 10 Blackmore and Cooper (1970)
- 11 Livingstone and Hubel(1987)
- 12 CNN の生理学的対応物
- 13 Yamins ら(2014,2016) の CNN と脳部位との対応
- 14 機械学習的概観
- 15 機械学習的概観 (2)
- 16 機械学習的概観 (3)
- 17 本質的困難さ
- 18 画像認識の進歩
- 19 勾配降下法
- 20 項目反応理論のニューラルネットワークアナロジー Neural network analogy of IRT
- 21 畳み込み演算
  - 21.1 定義
- 22 ディープラーニングの特徴
- 23 人工ニューロン
- 24 CNN 視覚化
- 25 福島のネオコグニトロン
- 26 LeNet
- 27 AlexNet
- 28 GoogLeNet インセプション
- 29 インセプション (2)
- 30 ResNet
- 31 ResNet のスキップ結合
- 32 R-CNN
- 33 Linsker の最大情報量保存原理
- 34 ガウシアンピラミッド
- 35 標準正則化理論とマックスプーリング
- 36 正則化項 ラグランジェ乗数
- 37 オイラー=ラグランジェ Euler Lagrange 方程式と正則化項
- 38 Poggio の標準正則化理論
- 39 Softmax 関数
- 40 腹側経路と背側経路
- 41 セマンティックセグメンテーションとインスタンスセグメンテーション
- 42 U-Net と afferent/efferent connections
- 43 拡張畳み込み Dialated Convolutions
- 44 クロネッカー積 Kronecker Product
- 45 情報量最大化原理に基づく正規分布
- 46 ResNet と Van Essen
- 47 畳み込み層
- 48 データ拡張
- 49 SGD
- 50 転移学習と蒸留
- Bibliography

1 謝辞

このような機会を与えていただきました 緑川 晶 先生，開催につきまして種々の便宜をはかっていただきました 重宗 弥生 先生に感謝申し上げます。

2 自己紹介



浅川伸一:博士(文学) 東京女子大学情報処理センター勤務。 早稲田大学在学時はピアジェの発生論的認識論に心酔する。 卒業後エルマンネットの考案者 ジェフ・エルマン に師事，薫陶を受ける。 以来人間の高次認知機能をシミュレートする ことを通して 知的であるとはどういうことかを考えていると思っていた。 著書に
「AI白書 2019, 2018」 (2019年, アスキー出版, 共著)， 「深層学習教科書 ディープラーニング G検定（ジェネラリスト）公式テキスト」 (2018年，翔泳社，共著)，「Python で体験する深層学習」 (コロナ社, 2016)，「ディープラーニング，ビッグデータ，機械学習あるいはその心理学」 (新曜社, 2015)， 「ニューラルネットワークの数理的基礎」「脳損傷とニューラルネットワークモデル，神経心理学への適用例」 いずれも守一雄他編「コネクショニストモデルと心理学」(2001)北大路書房など

3 時間配分

1. 自己紹介 2-5 分
2. 生理学的事実の確認 0 分 （スキップ）
3. 畳み込みニューラルネットワークの基礎概念。畳み込み，プーリング，ストライド，データ拡張，ソフトマックス関数，20 分
4. 畳み込みニューラルネットワークモデル AlexNet, Inception, ResNet, R-CNN, YOLO, SSD, U-Net, MobileNet, EfficientNet, MNet 30 分
5. 発展的話題 VAE, GAN 0 分 （時間切れ）
6. まとめ 5 分

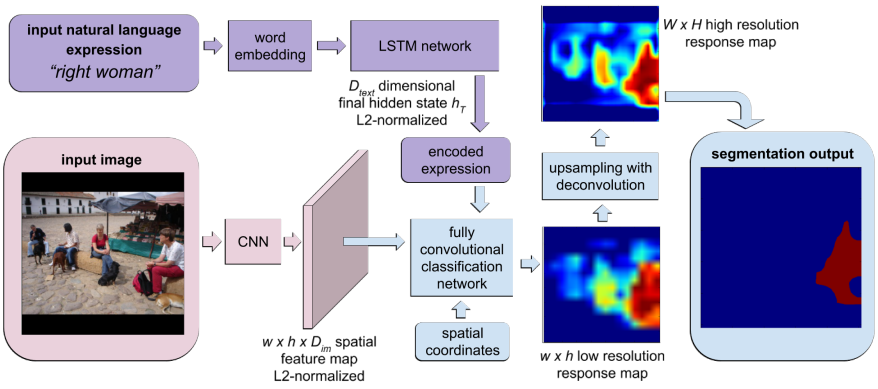
4 要旨

- 40 年前，福島は Hubel と Wiesel の生理学的知見を元にネオコグニトロンを提案した (Fukushima 1980) 。
- ルカンの提案した LeNet は福島のモデルの現代化である(LeCun et al. 1998)。
- 以来，畳み込みニューラルネットワークは SVM など従来モデルを席巻している。
- ここでは，それらの鍵となる概念を解説し，福島の目指した生理学的知見との対応を整理する
- ここに示す対応関係は，Hassabis ら [2017HassabisNeuron] の言う，人工知能研究と神経科学との良質な相互関係を構築する礎となるだろう

5 Takeaways

1. 福島のネオコグニトロンは Hubel and Wiesel の実装である。ネオコグニトロンの後継モデルが近年の LeNet(LeCun et al. 1998), AlexNet(Krizhevsky, Sutskever, and Hinton 2012), Inception(Szegedy et al. 2015), VGG(Simonyan and Zisserman 2015), などである。畳み込み，プーリング，ストライドなどの諸概念が実装されている
2. Yamins ら(Yamins et al. 2014), (Yamins and DiCarlo 2016) は複数の CNN 層をまとめて一つの皮質領野と見なしうる。これはインセプションモジュールの 1x1 畳み込みと併せて，側抑制を実現していると考えれば納得がいく
3. データ拡張 data augmentation, データ膨張 data dilation は眼球運動のトレモロ，ドリフトと関連し，微小な位置ずれや拡大，縮小，歪み，回転，射影変換などに頑健な特徴抽出器の創出に寄与していると考えられる。
4. 確率的勾配降下法(Bottou and Bousquet 2008) SGD (Stochastic Gradient Decent methods) と，モメンタム (momentum) は短期記憶，あるいはプライム刺激と解釈できるかも知れない。この効果はバッチ正則化 (Ioffe and Szegedy 2015) でも見られる。
5. 整流線型ユニット ReLU (Recutified Liner Units) は興奮，抑制を表す神経伝達物質を反映していると考えられる
6. ドロップアウト dropout は神経回路が決定論的に振る舞うのではなく確率的に動作することと対応している。
7. 正則化項 (L1, L2 normalization, weight decay, pruning) は，Marr & Poggio の原理を反映している。条件付き最適化であり広義のオイラー=ラグランジェ方程式と見なしうる
8. ソフトマックス関数は記号接地問題の解放と考えられる
9. セマンティックセグメンテーション，インスタンスセグメンテーションは，図地分離 figure/ground segmentations を実現している。腹側経路と背側経路との実装である R-CNN(Girshick et al. 2014), YOLO(Santosh Divvala, Girshick, and Farhadi 2016),(Redmon and Farhadi 2016), SSD(Liu et al. 2016) で実装で実装されている。
10. Fellman & Van Essen の皮質地図は ResNet(He et al. 2015) のスキップ結合で実装された。
11. 白質内の求心性，遠心性結合は U-Net(Ronneberger, Fischer, and Brox 2015) で実現され，図地分離の性能向上に寄与した
12. 転移学習で線画を認識できれば神経心理学検査をシミュレートできるだろう(Hinton, Vinyals, and Dean 2015)

6 本日お話しないこと



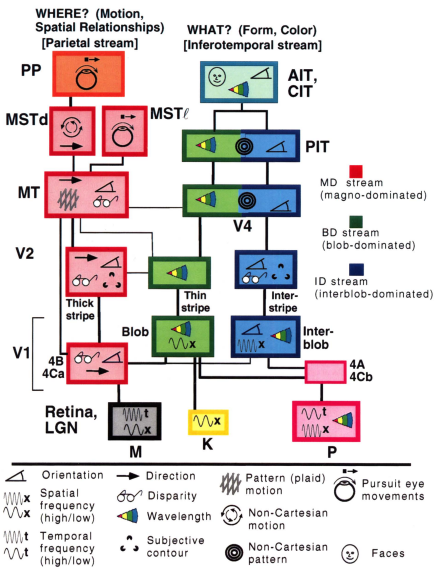
From (Hu et al. 2017)

7 キーコンセプト

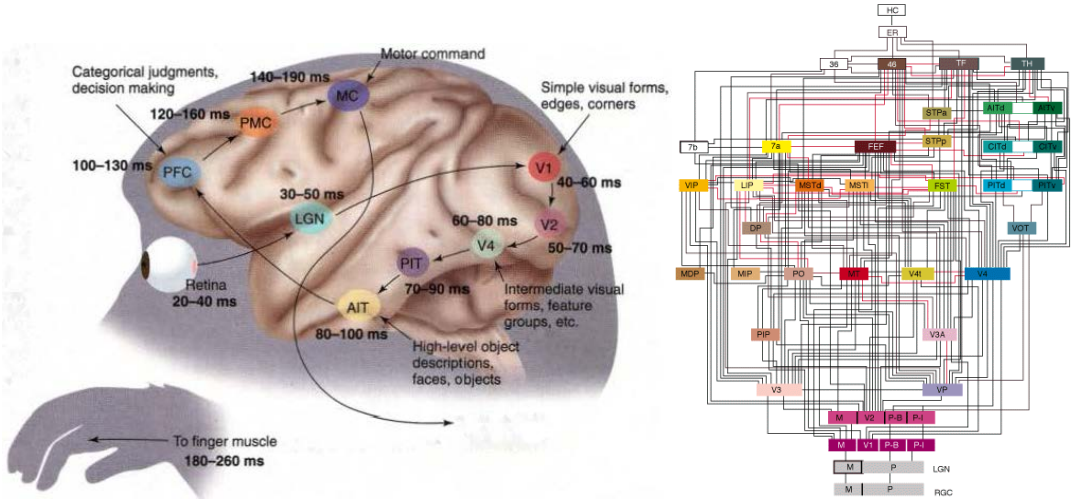
1. 畳み込みネットワークと初期視覚領野との対応関係，およびその特徴抽出や表現学習，記号接地問題，(ソフトマックス関数と penultimate 層)，Inception, VGG, ResNet

2. 畳み込み演算，福島のネオコグニトロン，C 層と S 層，マックスプーリング，
3. スキップコネクト(ResNet) と Van Essen ら視覚野結合地図との対応
4. データ拡張 data augmentation データ拡張は，そもそも視覚体験は Mountcastle のネコの新生児の研究から分かるとおり，毎日毎日大量の視覚刺激に晒されていると考えられる。
5. 膨張畳み込み dialted convolution
6. ReLU, tanh: tanh に関しては，純粋に計算論的な効率の側面がある。すなわち勾配消失問題 the vanishing gradient problems
7. Dropout
8. 確率的勾配降下法 (SGD: Stochastic Gradient Descent)，オンライン，バッチ，ミニバッチと短期記憶
9. 転移学習，一撃学習，小数例学習
10. 腹側=背側視覚経路と位置=対象の処理および領域切り出し，対象切り出し
11. 正確さと反応速度のトレードオフ(speed-accuracy tradeoff)，MnasNet(Tan et al. 2019)
12. エネルギー関数，最適化問題，制約付き正則化，オイラーラグランジェ方程式, InfoMax principle (Linsker, 1989), Entropy maximization with constraint is the unit Gaussian distribution.
13. 変分ベイズ，敵対生成ネットワーク（Nash 均衡），解絡表現 (disentangled representation) 。ヘルムホルツの自由エネルギー。

8 Van Essen らの視覚野地図

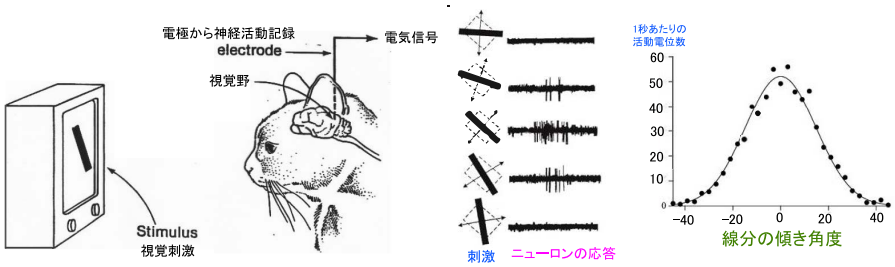


From (Essen and Gallant 1994) Fig. 2

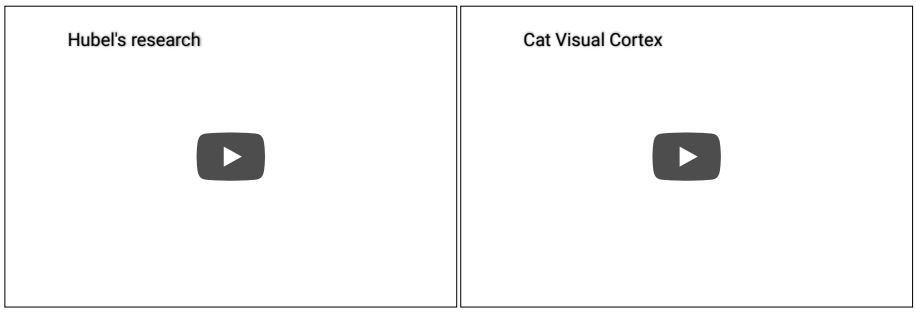


Left: (Thorpe and Fabre-Thorpe 2001), Right: (Felleman and Essen 1991)

9 Hubel and Wiesel



From (Hubel and Wiesel 1968)



10 Blackmore and Cooper (1970)

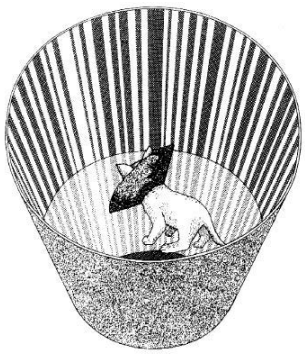


Fig. 1. The visual display consisted of an upright plastic tube, about 2 m high, with an internal diameter of 46 cm. The kitten, wearing a black ruff to mask its body from its eyes, stood on a glass plate supported in the middle of the cylinder. The stripes on the wall were illuminated from above by a spotlight. The luminance of the dark bars was about 10 cd. m^{-2} and of the bright stripes about 130 cd. m^{-2}; they were of several different widths. For this diagram the top cover and the spotlight have been removed from the tube.

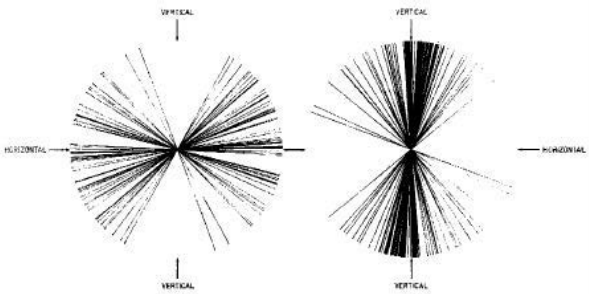
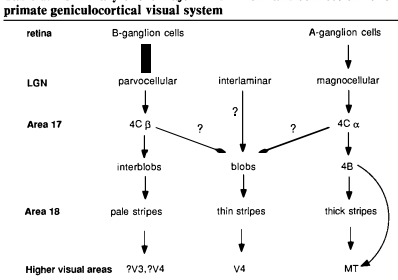


Fig. 2. These polar histograms show the distributio s of optimal orientatio s for fifty-two neurones from a horizontally experienced cat on the left, and seventy-two from a vertically experienced cat on the right. The slight torsion of the eyes, caused by the relaxant drug, was assessed by photographing the pupils before and after anaesthesia and paralysis. A correction has been applied for torsion, so the polar plots are properly orientated for the cats' visual fields. Each line shows the optimum orientation for a single neurone. For each binocular cell the line is drawn at the mean of the estimates of optimal orientation in the two eyes. No cells have been disregarded except for one with a concentric receptive field and hence no orientational selectivity.

11 Livingstone and Hubel(1987)

Table 2. Summary of the major subdivisions and connections of the primate geniculocortical visual system



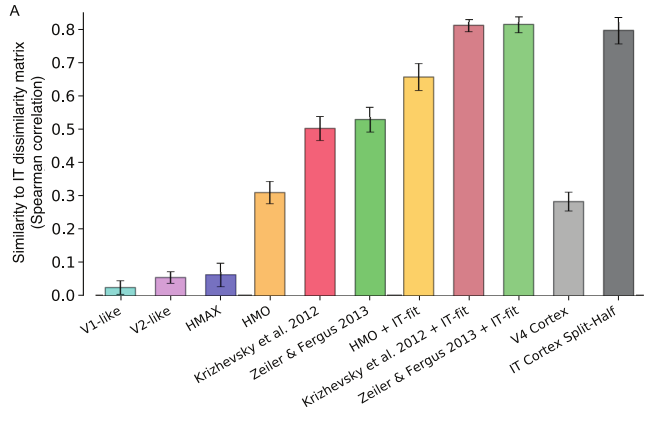
Property			
Color	Yes/no[a]	Yes	No
Contrast sensitivity	Low	High[d]	High
Spatial resolution	High	Low	Low
Orientation selectivity	Yes	No	Yes
Movement sensitivity	Yes	No	Yes
Directionality	No[b]	No	Yes[e]
Stereopsis	No[c]	No	Yes

[a] Cells beyond 4Cβ do respond to color-contrast borders but are not overtly color-coded.
[b] At least it is not prominent.
[c] In anesthetized animals, we have seen only a few stereotuned cells in upper-layer area 17 (Livingstone and Hubel, 1984b; see also Hubel and Wiesel, 1970). In attentive animals, cells coded for stereoscopic depth have been reported both above and below layer 4C of area 17, but are especially concentrated in layer 4B

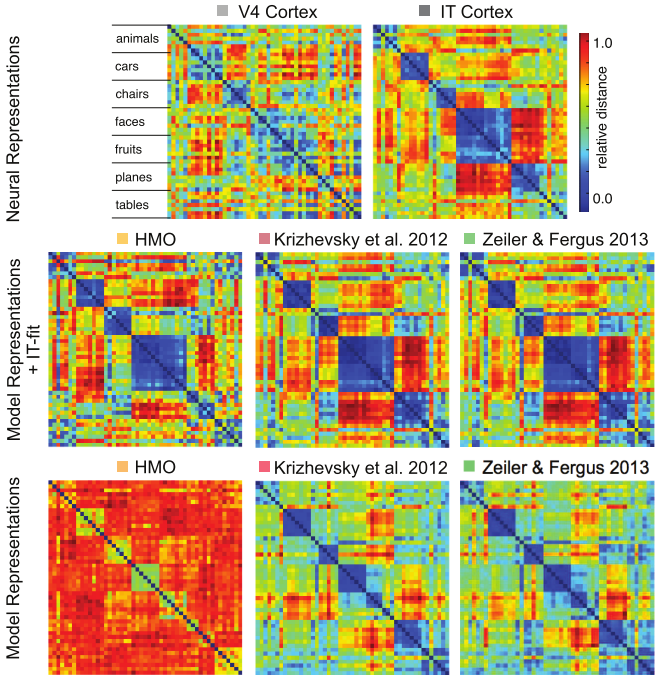
(Poggio and Fischer, 1977; Poggio et al., 1985; G. F. Poggio, personal communication). We do not understand these differences in results, but one possibility is that the stereo mechanisms are built up in 18, and the stereotuning in 17 is the result of a back projection that is suppressed by anesthesia.
[d] By deoxyglucose (Tootell et al., 1985).
[e] Rare in thick stripes in area 18 but very common in layer 4B of area 17 and in MT.

12 CNN の生理学的対応物

(Cadieu et al. 2014),(Cichy et al. 2016),(Marblestone, Wayne, and Kording 2016),(Cadieu et al. 2014) らは，物体認識時におけるサルの下側頭葉のニューロンの活動と CNN の活動とを比較した。 サルの下側頭葉で記録された数千ニューロンで表象されるカテゴリー空間は， 位置，ポーズ，縮尺，背景などにかかわらず，CNN と類似した。 下側頭葉ニューロン集団とのカテゴリー毎の混同行列 (Cadieu et al. 2014)



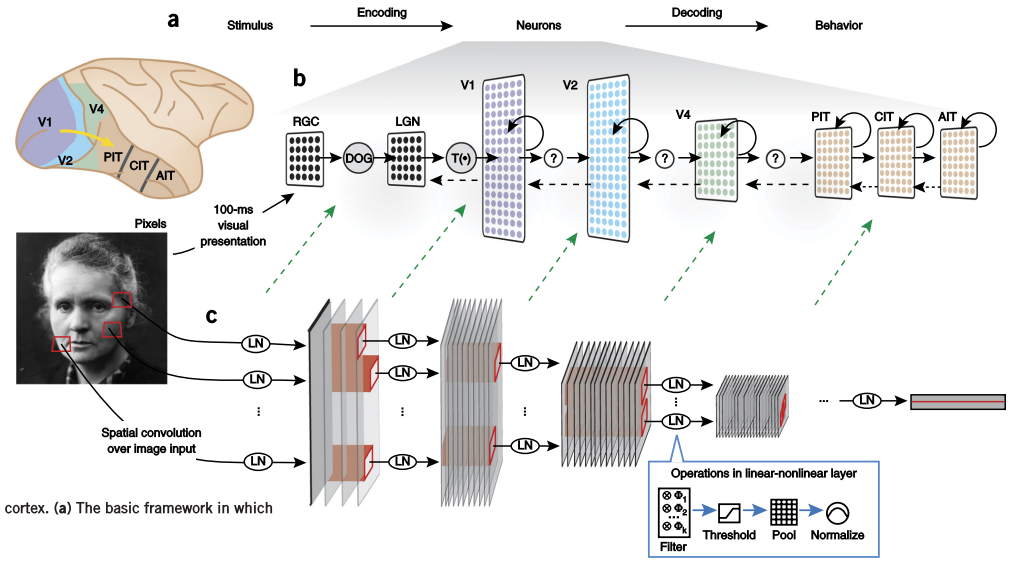
(Cadieu et al. 2014) Fig. 7A



(Cadieu et al. 2014) Fig. 7B

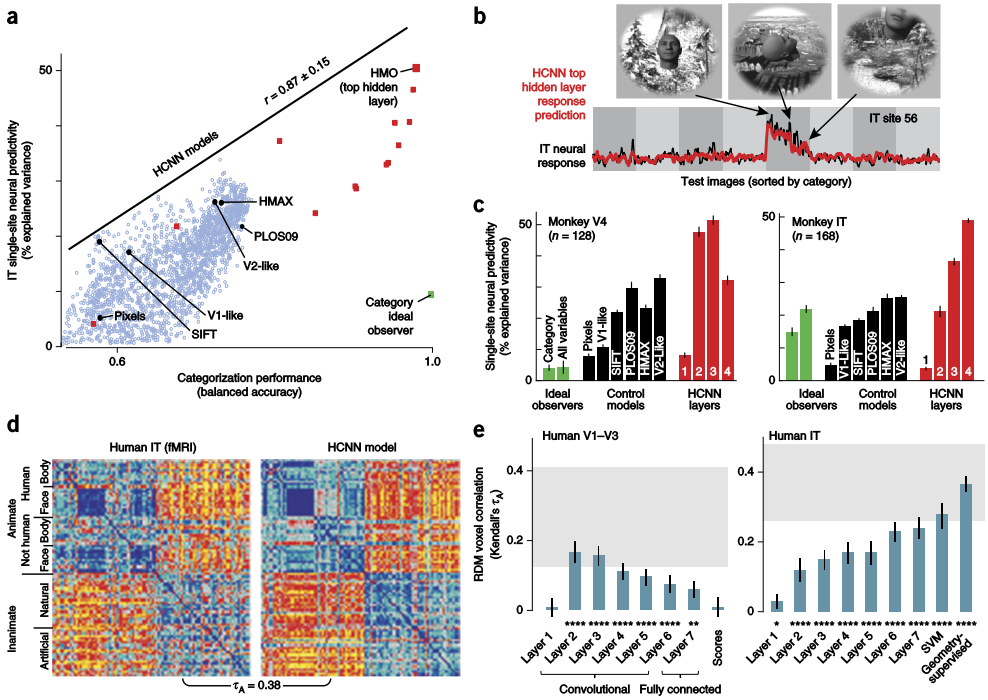
13 Yamins ら(2014,2016) の CNN と脳部位との対応

- Yamins ら (2014, 2016) によれば下図のような対応を考えることが可能です。 この主張に従えば CNN の各層を破壊もしくは機能不全にすれば視覚障害をシミュレートすることが可能になります。



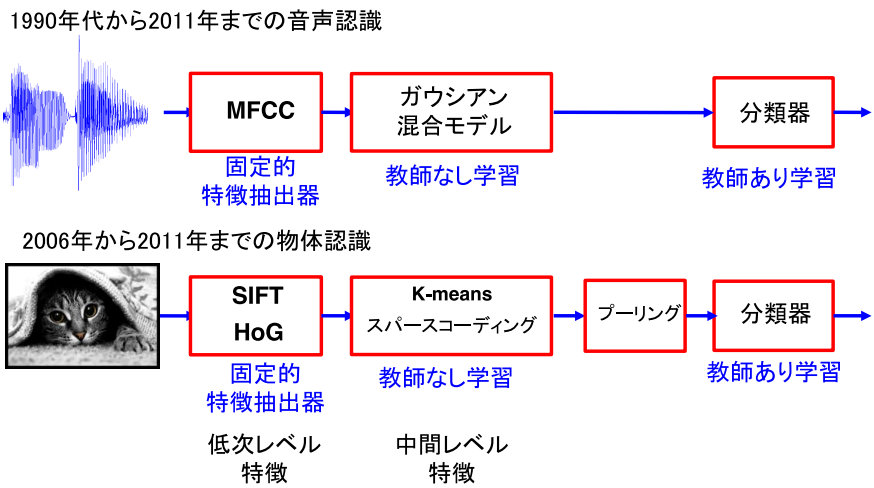
cortex. (a) The basic framework in which

Yamins (2016) Fig. 1 より



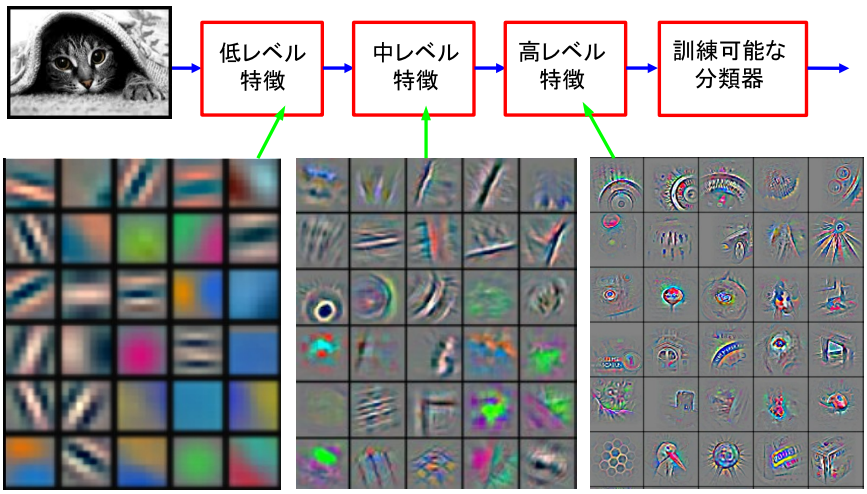
Ymains (2016) Fig. 2 より

14 機械学習的概観



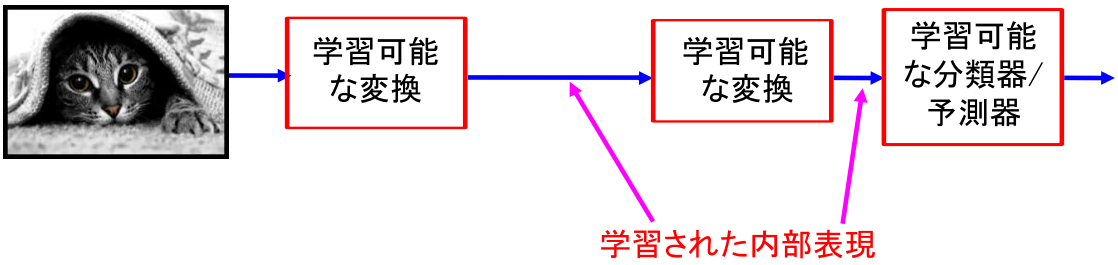
LeCun (2013) より

15 機械学習的概観 (2)



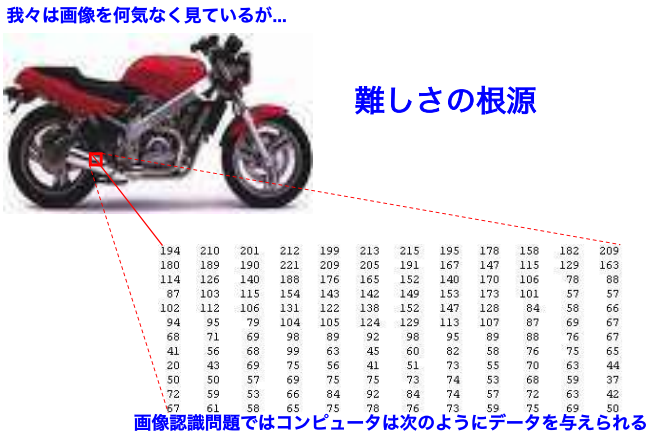
LeCun (2013) より

16 機械学習的概観 (3)



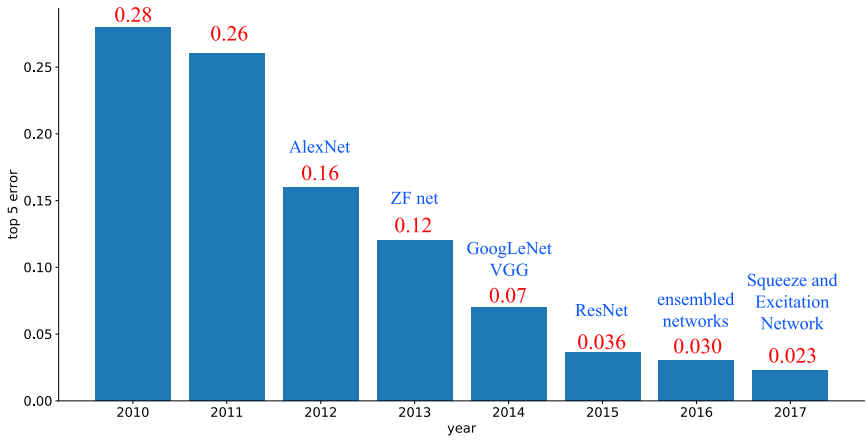
LeCun (2013) より

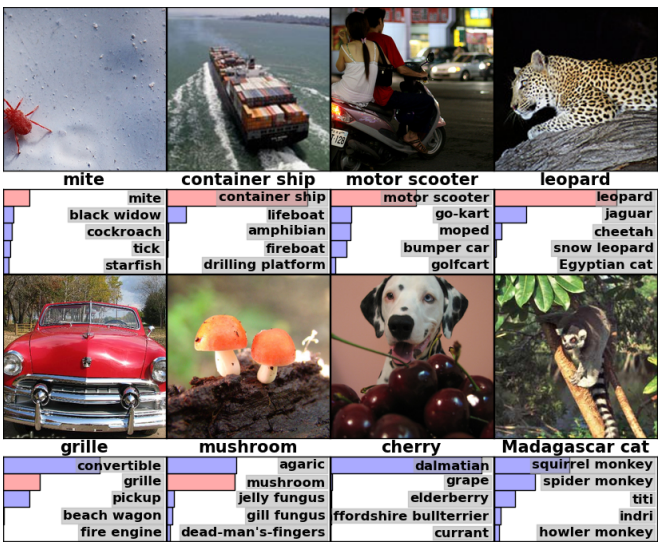
17 本質的困難さ

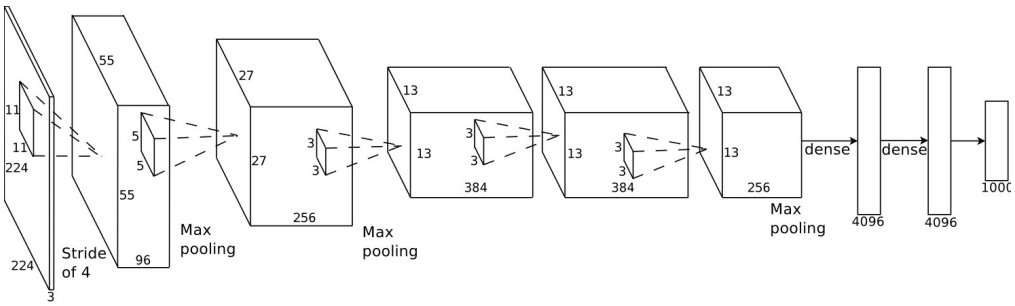


Ng (2017) を改変

18 画像認識の進歩







19 勾配降下法

- ニューラルネットワークにおける学習は，誤差，または損失関数を最小にするようなパラメータを更新することである。

$$\nabla\theta = \epsilon\frac{\partial\ell(\mathbf{x},\theta)}{\partial\theta} \tag{1}$$

を勾配降下法という。一次微分だけしか用いないで，二次微分を用いるニュートン法を用いないのは，計算コストがかかるからである。ℓ の定義には，交差エントロピー，最小自乗誤差，尤度などが用いられる。

このとき合成関数の微分公式を用いて，微分情報を下位層へ伝播させることを誤差逆伝播法という。

20 項目反応理論のニューラルネットワークアナロジー Neural network analogy of IRT

$$p(\theta_i, a_j, b_j, c_j, x_{ij}) = \frac{1+c_j}{1+\exp\left(a_j\theta_i + b_j\right)} \tag{2}$$

ニューラルネットワークと項目反応理論の相違は，推定すべきパラメータの相違である。

21 畳み込み演算

Convolution 日本語では 畳込み と表記されることもある。ここでは **畳み込み** と表記する。A Japanese floor mat の名詞の意味で用いられるときに **畳** と表記し， 動詞の **畳む** に当たる単語としては送り仮名を付して表記する。

畳み込み（操作，演算）は 2 つの関数 f と g との間で定義される数学的演算である。一方の関数が他方によってどのように変更されるかを表す関数を生成する演算である。 畳み込みという用語が用いられた場合，その演算の結果生成された関数を指す場合もあれば，その演算を指す場合もある。

畳み込み演算は以下のような特徴を持つ。 連続変数または離散変数の実数値関数の場合，$f(x)$ または $g(x)$ のいずれかが y 軸について反映されるという点でのみ相互相関と異なる。 したがって $f(x)$ と $g(-x)$ の相互相関 または $f(-x)$ と $g(x)$ である。 連続関数の場合 相互相関演算子は畳み込み演算子の随伴である。

畳み込み演算は 確率，統計，コンピュータービジョン，自然言語処理，画像処理，信号処理，微分方程式などに用いられている。

21.1 定義

f と g との畳み込み演算（操作）は $f * g$ と表記される。 一方の関数が反転されてシフトされた後，2 つの関数の積の積分として定義される。 このため畳み込み演算とは積分変換の一種である。

$$(f * g)(t) \triangleq \int_{-\infty}^{\infty} f(\tau)g(t-\tau)\,d\tau. \tag{3}$$

以下の定義も同等の結果を得る:

$$(f * g)(t) \triangleq \int_{-\infty}^{\infty} f(t-\tau)g(\tau)\,d\tau. \tag{4}$$

記号 t は時間領域だけではなく，空間領域に対しても用いられる。画像処理における畳み込みは 2 次元の画像を入力空間として扱う。 畳み込み演算は 関数 $f(\tau)$ の重み付き平均とも見なしうる。 t が変化すると，重み関数は入力関数のさまざまな部分を強調あるいは変調する。

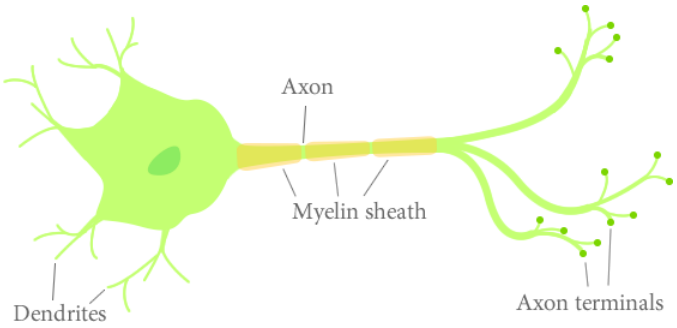
22 ディープラーニングの特徴

- データハングリー data hungry
- 計算資源ハングリー resource hungry
- 理論欠如 theory lagging
- 不透明 opacity

- ニューラルネットワークは素人の統計学である, Anderson et. al (1993)

… But Neural networks are not alchemy.

23 人工ニューロン



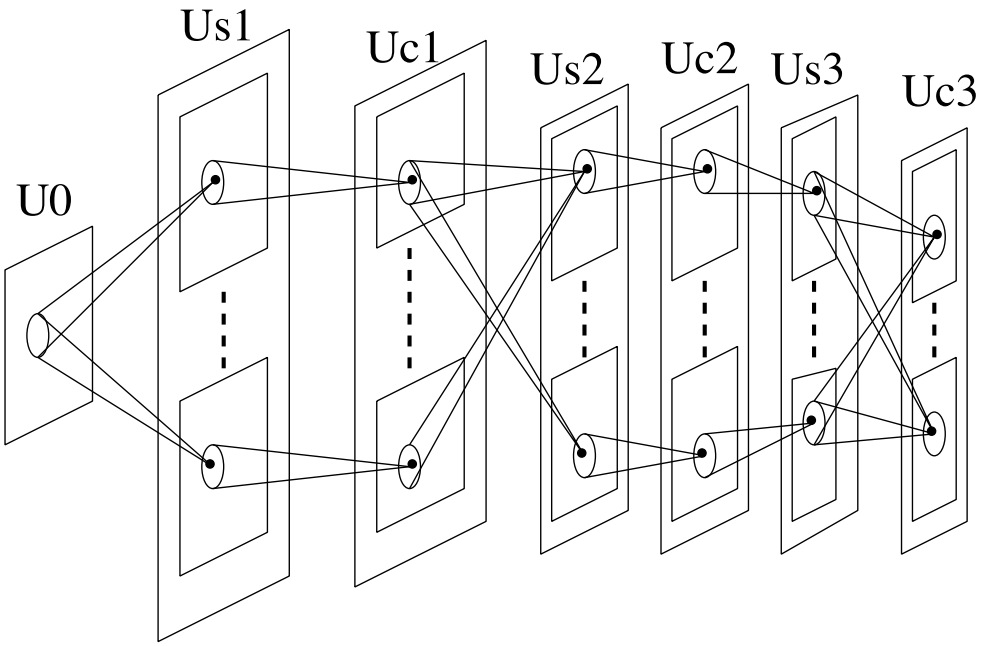
ニューロンの模式図

人口ニューロンモデル

24 CNN 視覚化

- http://storage.googleapis.com/deepdream/visualz/tensorflow_inception/index.html
- http://storage.googleapis.com/deepdream/visualz/vgg16/index.html
- インセプション

25 福島のネオコグニトロン



ネオコグニトロン は Hubel と Wiesel の生理学実験 (Hubel and Wiesel 1959),(Hubel and Wiesel 1962),(Hubel and Wiesel 1968) から得られた事実に基づき，視覚認識を行うモデルである。 S 層と C 層とはそれぞれ単純細胞と複雑細胞から構成される層を示している。 生理学的事実に基づき，ネオコグニトロンのニューロンの受容野は層を登るに従って大きくなる。 同時に，受容野内に照射された刺激は位置不変性を持っている。すなわち回転，拡大縮小，移動などアフィン変換 (affine transform) に対して頑健である。

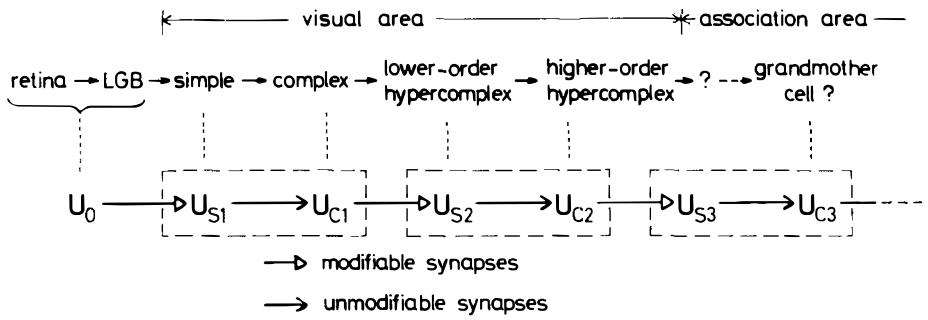


Fig. 1. Correspondence between the hierarchy model by Hubel and Wiesel, and the neural network of the neocognitron

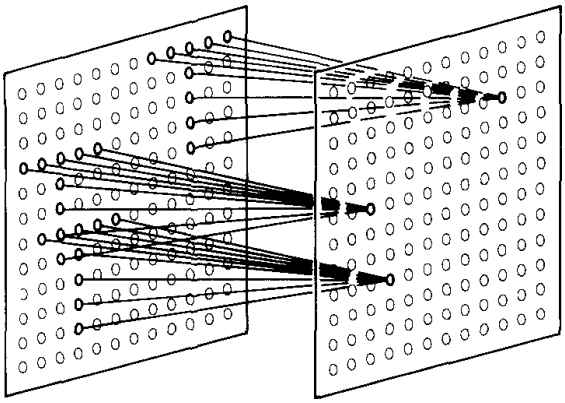
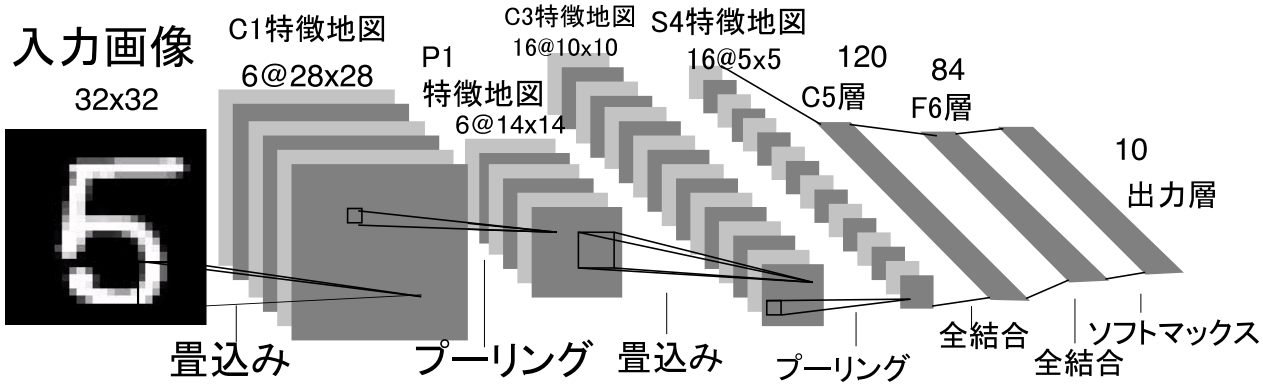


Fig. 3. Illustration showing the input interconnections to the cells within a single cell-plane

26 LeNet

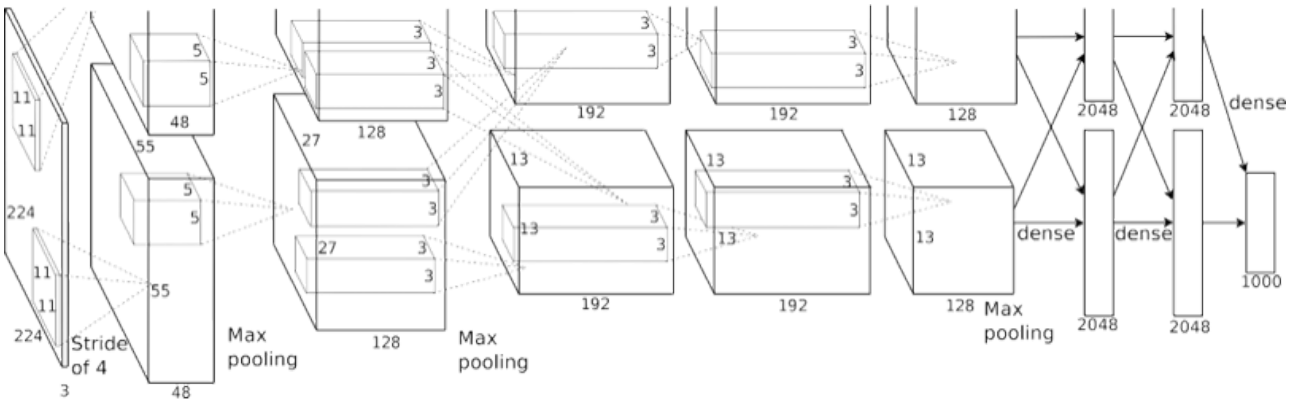
(LeCun et al. 1998) は手書き数字認識モデル LeNet を提案した。LeNet は畳込み演算とサブサンプリング (sub sampling) の繰り返しである。 サブサンプリングは後にプーリング (pooling) に置き換えられている。



最左の入力画像の大きさは高さ，幅すなわち縦横の画素数が 32×32 の濃淡画像 (ゆえに奥行きが1) である。 続く C1 特徴地図層 (第 1 層) は高さと幅が 28×28 の特徴地図であり， 特徴数は 6 （従って奥行きは 6） である。 次の P1 特徴地図層は，C1 層をプーリングした層であり，高さ 14, 幅 14，奥行き 6 (従って特徴数は6)である。 同じ処理がもう一度繰り返され，畳込みとプーリングが行われる。 さらに 10 種類の手書き文字を識別するために全結合層 が 2 つ，C5層 (ニューロン数120)， F6層 (ニューロン数84) が用意され最終層は 10 ニューロンである。 これら 10 個のニューロンはそれぞれ 0 から 9 までの 10 種類の手書き数字に対応している。

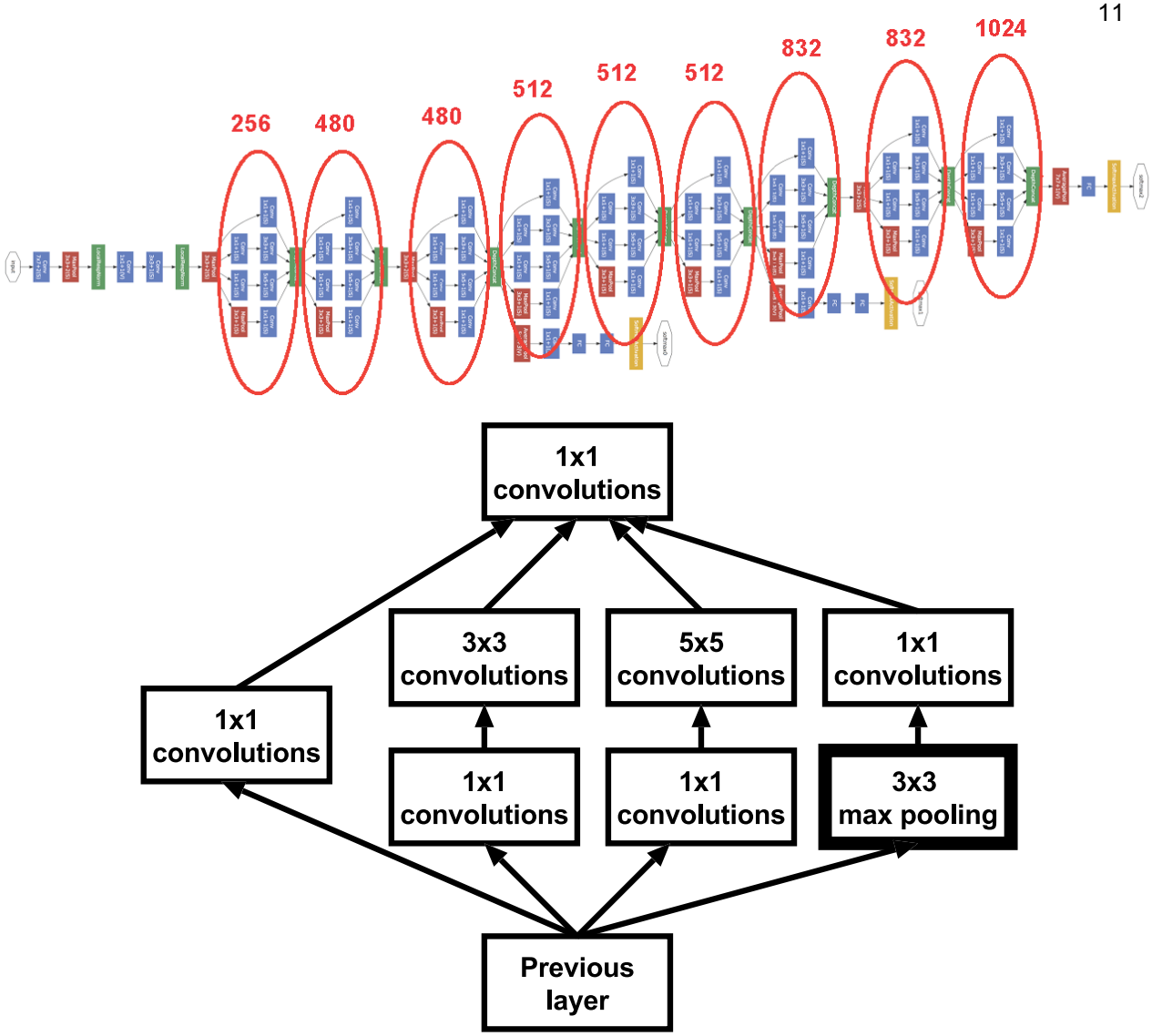
27 AlexNet

2012年の大規模画像認識コンテスト ImageNet において当時 SOTA (state of the art) であった SVM を 10% 以上凌駕したモデルが AlexNet (Krizhevsky, Sutskever, and Hinton 2012)である。 第一著者の名前からの呼び名である。 AlexNet の特徴としては，畳込み，ドロップアウト, データ拡張，GPU の利用，局所正規化，が挙げられる。 ImageNet の分類課題は，画像を 1000 種のカテゴリに分類することが求められる。 このとき上位5候補を出力して，この 5 カテゴリの中に正解が含まれているか否かで性能を競う。 SVM では上位 5 候補の誤判別率が約 26% であった。 一方 AlexNet は 16% を達成した。 ネットワークの構成は LeNet と同様であるが規模が大きい。



28 GoogLeNet インセプション

GoogLeNet(Szegedy et al. 2015) は ImageNet2014 の優勝モデル。 Google が開発し LeNet に敬意を表して GoogLeNet と表記される。 GoogLeNet は複数のカーネルを並列に用いたインセプション(inception)モジュールを基本単位とする。 インセプションモジュールとは，複数のカーネルを並列に用いて基本単位とする インセプションモジュール内では結合が構造化されているので，全結合を考えるより総結合数が少なくて済む。 実際 AlexNet では総結合数(従って推定すべきパラメータ数)が約 600 万であったが，GoogLeNet では 40万。



Inception モジュール

29 インセプション (2)

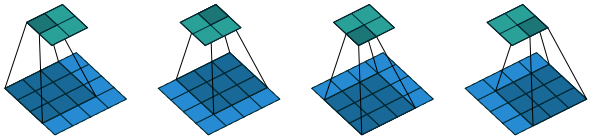


Figure 2.1: (No padding, no strides) Convolving a 3×3 kernel over a 4×4 input using unit strides (i.e., $i = 4$, $k = 3$, $s = 1$ and $p = 0$).

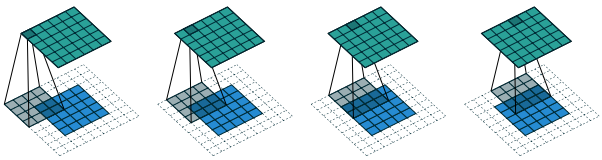


Figure 2.2: (Arbitrary padding, no strides) Convolving a 4×4 kernel over a 5×5 input padded with a 2×2 border of zeros using unit strides (i.e., $i = 5$, $k = 4$, $s = 1$ and $p = 2$).

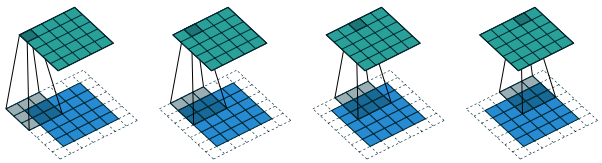


Figure 2.3: (Half padding, no strides) Convolving a 3×3 kernel over a 5×5 input using half padding and unit strides (i.e., $i = 5$, $k = 3$, $s = 1$ and $p = 1$).

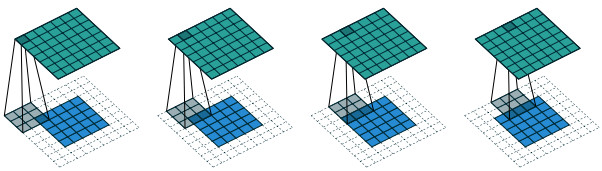


Figure 2.4: (Full padding, no strides) Convolving a 3×3 kernel over a 5×5 input using full padding and unit strides (i.e., $i = 5$, $k = 3$, $s = 1$ and $p = 2$).

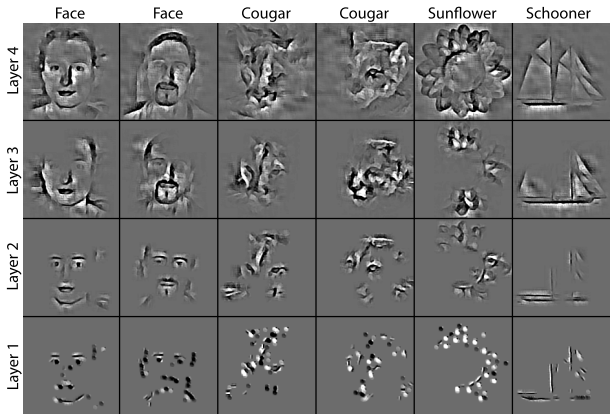
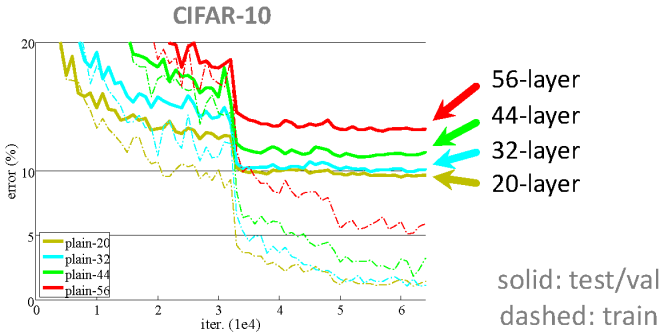


Figure 1. Top-down parts-based image decomposition with an adaptive deconvolutional network. Each column corresponds to a different input image under the same model. Row 1 shows a single activation of a 4^{th} layer feature map projected into image space. Conditional on the activations in the layer above, we also take a subset of 5, 25 and 125 active features in layers 3, 2 and 1 respectively and visualize them in image space (rows 2-4). The activations reveal mid and high level primitives learned by our model. In practice there are many more activations such that the complete set sharply reconstructs the *entire* image from each layer.

30 ResNet

ResNet については既述したが，その他にショートカットと FastRCNN(Girshick 2015), FasterRCNN(Ren et al. 2015) の手法を用いて関心領域の切り出しを行っている。

ResNet はさらにバッチ正則化 (Ioffe and Szegedy 2015)も取り入れている。 バッチ正則化を用いたため ドロップアウト［Hinton et al. (2012)] は用いられなかった。



31 ResNet のスキップ結合

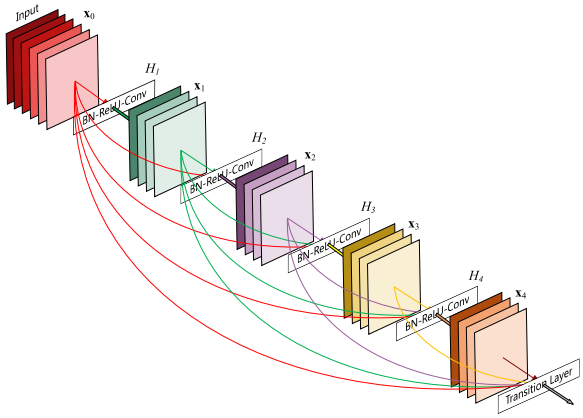
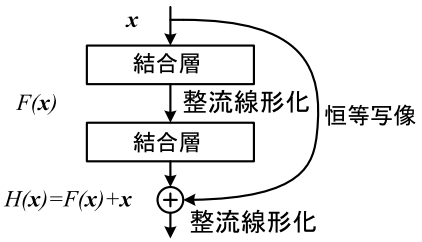
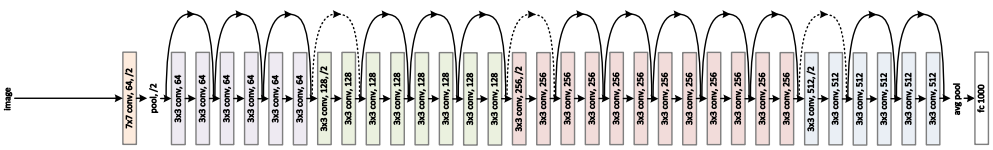


Figure 1: A 5-layer dense block with a growth rate of $k = 4$. Each layer takes all preceding feature-maps as input.

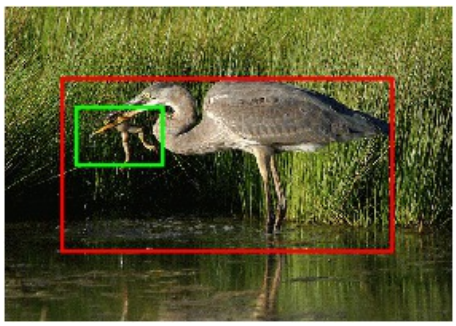
(Huang, Liu, and Maaten 2018) Fig. 1





32 R-CNN

ImageNet コンテストにはクラス分類課題とロケーション課題とが存在する。 クラス分類課題とは画像データが与えられたとき，その画像が何であるかを問う課題である。 一方，与えられた画像中のどの位置に物体が存在するかを問う課題をロケーション課題という。 データ拡張などを用いて訓練しても CNN では画像の判別は行えても，画像上で物体が占める位置を問うことは難しいと考えられてきた。



一般画像認識の難しさは，対象が部分的に他の対象に隠蔽されていたり，画像中に存在する物体の個数についての事前知識を仮定できないことなどが挙げられる。

33 Linsker の最大情報量保存原理

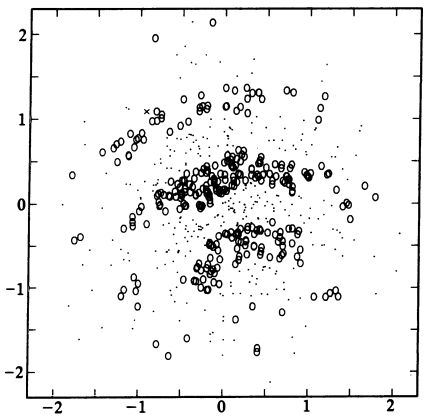


FIG. 1. Synaptic positions and connection strengths at maturity for a single cell of layer G having 600 synapses, placed randomly according to a two-dimensional Gaussian distribution. Parameter values are $n_{EG} = 0.5$, $r_G/r_F = 4$, $k_1 = 0$, $k_2 = -3$. Connection strengths are indicated as ovals, for $c = -0.5$, and dots, for $c = +0.5$; x represents intermediate c (one point only). Axis values are in units of r_G.

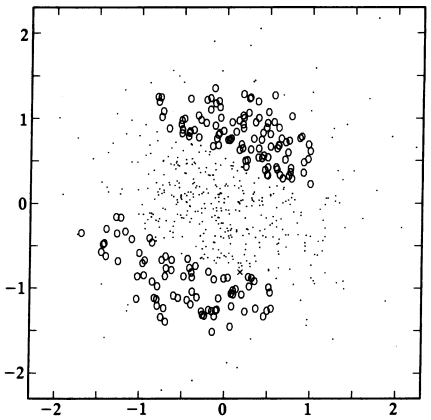


FIG. 2. A bilobed G cell. Parameter values are $n_{EG} = 0.5$, $r_G/r_F = 1.8$, $k_1 = 0.6$, $k_2 = -3$. Symbols are as in Fig. 1.

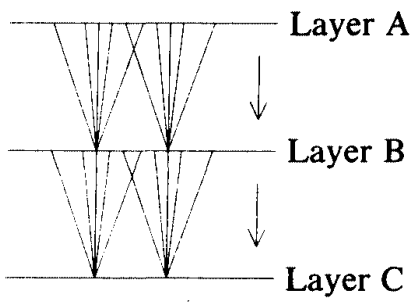


FIG. 1. Modular self-adaptive network diagram for the system discussed in this paper.

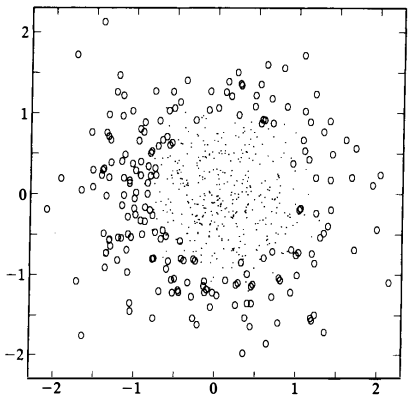
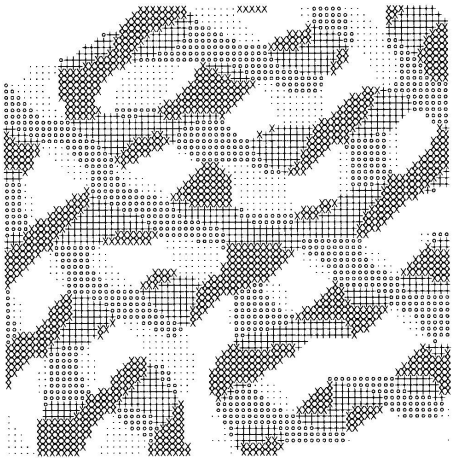
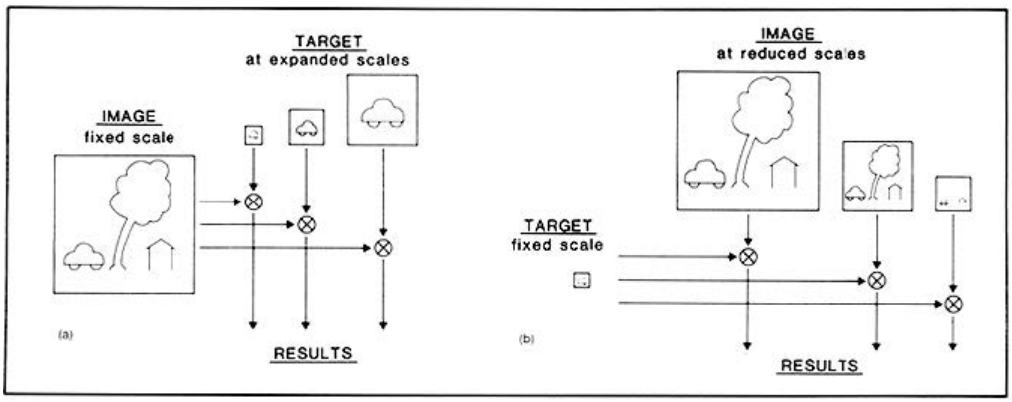
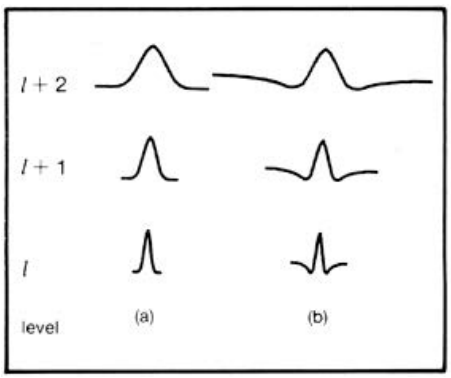


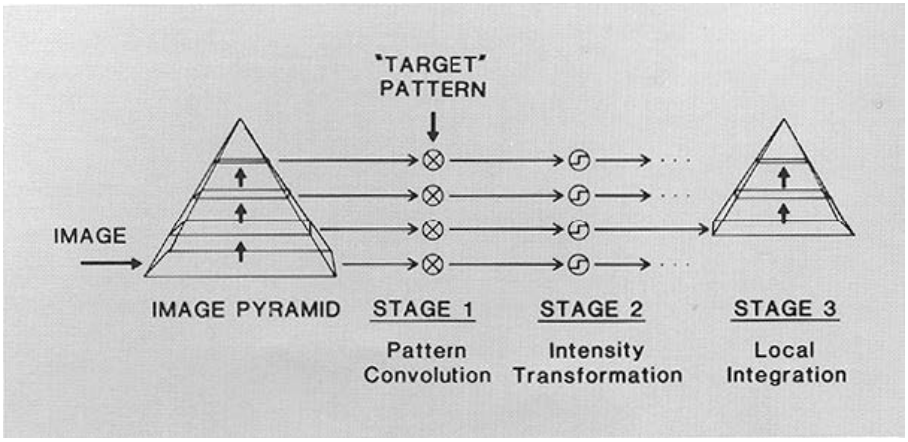
FIG. 2. Synaptic positions and mature connection strengths for a single cell of layer C having 600 synapses. Parameter values are $k_1 = 0.45$, $k_2 = -3$, $r_C/r_B = 3_{1/2}$, and each c value is allowed to range between -0.5 and $+0.5$. Random initial c values are chosen from uniform distribution on the interval -0.5 to $+0.5$. Values of Q^B_{nm} are appropriate to random placement of A-to-B and B-to-C synapses; layer uniformity is not assumed (see text). At maturity, every c reaches an extreme value: 0.5 (indicated by an oval) or -0.5 (dot). Axes are labeled by distance from cell center (in units of r_C).



34 ガウシアンピラミッド







Efficient procedure for computing integrated image properties at many scales. Each level of the image pyramid is convolved with a pattern to enhance an elementary image characteristic, step 1. Sample values in the filtered image may then be passed through a nonlinear transformation, such as a threshold or power function, step 2. Finally, a new "integration" pyramid is built on each of the processed image pyramid levels, step 3. Node values then represent an average image characteristic integrated within a Gaussian-like window.

35 標準正則化理論とマックスプーリング

(Poggio, Torre, and Koch 1985,@1995GirosiPoggio,@1999Riesenhuber_Poggio,@2005Serre_Poggio) と正則化L1, L2, L0, ElasticNet

(Riesenhuber and Poggio 1999) は マックスプーリングが生理学データと合致すると主張した

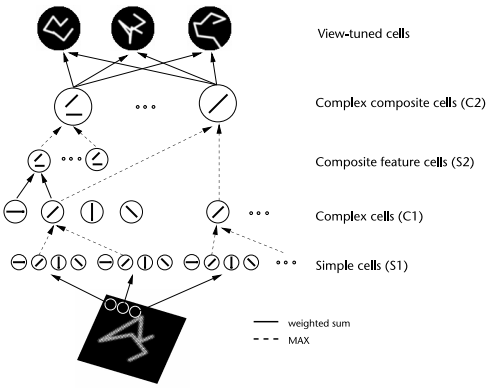


Fig. 2. Sketch of the model. The model was an extension of classical models of complex cells built from simple cells, consisting of a hierarchy of layers with linear ('S' units in the notation of Fukushima, performing template matching, solid lines) and non-linear operations ('C' pooling units, perform-ing a 'MAX' operation, dashed lines). The nonlinear MAX operation—which selected the maximum of the cell's inputs and used it to drive the cell—was key to the model's proper-ties, and differed from the basically linear summation of inputs usually assumed for complex cells. These two types of operations provided pattern specificity and invariance to translation, by pooling over afferents tuned to different posi-tions, and to scale (not shown), by pooling over afferents tuned to different scales.

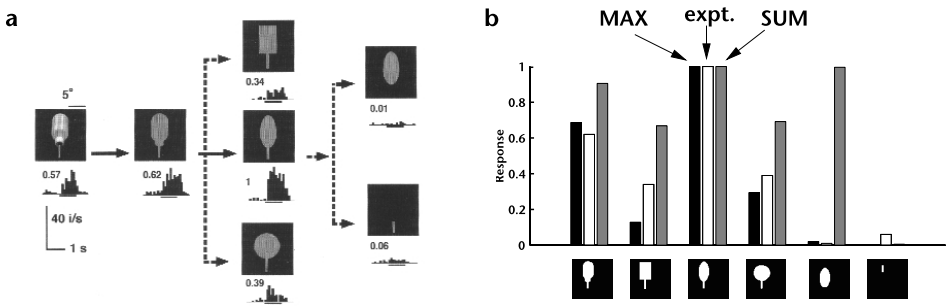


Fig. 3. Highly nonlinear shape-tuning properties of the MAX mechanism. (**a**) Experimentally observed responses of IT cells obtained using a 'simplifi-cation procedure'[26] designed to determine 'optimal' features (responses normalized so that the response to the preferred stimulus is equal to 1). In that experiment, the cell originally responded quite strongly to the image of a 'water bottle' (leftmost object). The stimulus was then 'simplified' to its monochromatic outline, which increased the cell's firing, and further, to a paddle-like object consisting of a bar supporting an ellipse. Whereas this object evoked a strong response, the bar or the ellipse alone produced almost no response at all (figure used by permission). (**b**) Comparison of experiment and model. White bars show the responses of the experimental neuron from (**a**). Black and gray bars show the response of a model neu-ron tuned to the stem-ellipsoidal base transition of the preferred stimulus. The model neuron is at the top of a simplified version of the model shown in Fig. 2, where there were only two types of S1 features at each position in the receptive field, each tuned to the left or right side of the transition region, which fed into C1 units that pooled them using either a MAX function (black bars) or a SUM function (gray bars). The model neuron was con-nected to these C1 units so that its response was maximal when the experimental neuron's preferred stimulus was in its receptive field.

36 正則化項 ラグランジェ乗数

$\mathbf{X}$ をデータ行列とする。一行一データ，各列には，変数があるとする。画像であれば，各画素，調査データであれば，各調査項目，脳画像であれば各画素あたりの画素値である。 このとき $\mathbf{w}$ をもちいて合成変量 $\mathbf{y} = \mathbf{X}\mathbf{w}$ を作って，$\left(\mathbf{y}^{\mathsf{T}}\mathbf{y}\right)$ すなわち $\mathbf{y}$ の分散を最大化することを考える。 ここで T は行列の転置を表すものとする。

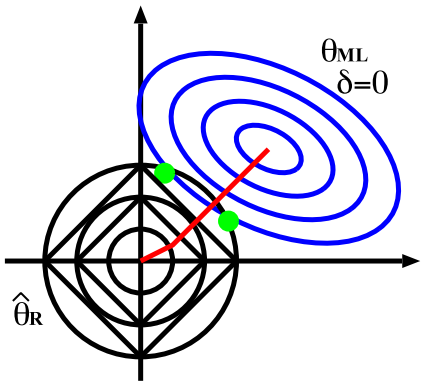
$$\arg\max \operatorname{var}(\mathbf{y}) = (\mathbf{X}\mathbf{w})^{\mathsf{T}}\mathbf{X}\mathbf{w} = \mathbf{w}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}\mathbf{w} \tag{5}$$

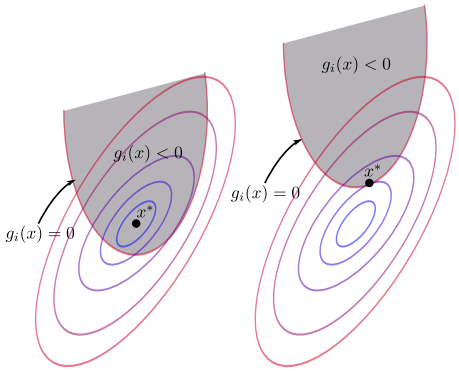
このとき $\mathbf{w}$ に制約を設けなければ，$\mathbf{y}$ の分散際限なく大きくなってしまう。そこで $\mathbf{w}^\top \mathbf{w} = 1$ という条件の元で where, $\mathbf{w}^\top \mathbf{w} = 1$

Lagrange multiplier λ を用いれば次式のように固有値問題に帰結する:

$$\lambda \mathbf{X}^\top \mathbf{X} = \mathbf{X}^\top \mathbf{X} \mathbf{w} \tag{6}$$

正則化項は，条件付き最適化とみなすことができる。したがって，L1, L2 正則化も VAE で用いられる Klause-Kuhn-Tucker 条件も条件付き正則化と考えられる。





37 オイラー=ラグランジェ Euler Lagrange 方程式と正則化項

Original:
https://ja.wikipedia.org/wiki/%E3%82%AA%E3%82%A4%E3%83%A9%E3%83%BC%EF%BC%9D%E3%83%A9%E3%82%B0%E3%83%A9%E3%83%B3%E3%82%B8%E3%83%A5%E6%96%B9%E7%A8%8B%E5%BC%8F

3次元デカルト座標 $\boldsymbol{x} = (x, y, z)$ の場合を考える。このとき時間微分 $\dot{\boldsymbol{x}} = \boldsymbol{v} = (v_x, v_y, v_z)$ は速度である。また ポテンシャルは速度には依らないものとする。 ラグランジアン L は『**運動エネルギー − ポテンシャル**』の形をしており [L({},{},t)= ({v_{x}}^{2}+{v_{y}}{2}+{v_{z}}^{2})-V()] である。

このとき，ラグランジュの運動方程式は [m{} = -V({})] となり ニュートンの運動方程式に一致する。

ニュートン力学においては 関数 u_i は一般化座標 q_i であり，その変数は時間 t である。一般化座標の次元 f を系の（力学的な）自由度という。 関数 F はラグランジアン L がその役割を果たす。 オイラー＝ラグランジュ方程式は [(q(t),{}(t),t) - {}({} (q(t),{}(t),t)) = 0] となる。なお，ドットは時間による微分を表す。この式を特に **ラグランジュの運動方程式** と呼ぶこともある。 **一般化運動量** は [p_{i}(q(t),{}(t),t) = (q(t),(t),t)] で定義され，これを使うとオイラー＝ラグランジュ方程式は [_{i}= (q(t),{}(t),t)] と書き換えられる。上式右辺を一般化力と呼ぶ事にすると，上述の方程式は「**一般化運動量の微分 ＝ 一般化力**」を意味する。

ニュートン方程式は「運動量の微分＝力」であったので，オイラー＝ラグランジュ方程式はニュートン方程式を一般化座標に拡張したものであるとみなす事ができる。

38 Poggio の標準正則化理論

The regularization of the ill-posed problem of finding z from the 'data' y

$$Az = y \tag{1}$$

requires the choice of norms $\|\cdot\|$ and of a stabilizing functional $\|Pz\|$. In standard regularization theory, A is a linear operator, the norms are quadratic and P is linear. Two methods that can be applied are[8,13]: (1) among z that satisfy $\|Az - y\| \leqslant \varepsilon$ find z that minimizes (ε depends on the estimated measurement errors and is zero if the data are noiseless)

$$\|Pz\|^2 \tag{2}$$

(2) find z that minimizes

$$\|Az - y\|^2 + \lambda \|Pz\|^2 \tag{3}$$

where λ is a so-called regularization parameter.

The first method computes the function z that is sufficiently close to the data and is most 'regular', that is minimizes the 'criterion' $\|Pz\|^2$. In the second method, λ controls the compromise between the degree of regularization of the solution and its closeness to the data. Standard regularization theory provides techniques for determining the best λ[12,15]. Thus, standard regularization methods impose the constraints on the problem by a variational principle, such as the cost functional of equation (3). The cost that is minimized reflects physical constraints about what represents a good solution: it has to be both close to the data and regular by making the quantity $\|Pz\|^2$ small. P embodies the physical constraints of the problem. It can be shown for quadratic variational principles that under mild conditions the solution space is convex and a unique solution exists. It must be pointed out that standard regularization methods have to be applied after a careful analysis of the ill-posed nature of the problem. The choice of the norm $\|\cdot\|$, of the stabilizing functional $\|Pz\|$ and of the functional spaces involved is dictated both by mathematical properties and by physical plausibility. They determine whether the precise conditions for a correct regularization hold for any specific case.

Variational principles are used widely in physics, economics and engineering. In physics, for instance, most of the basic laws have a compact formulation in terms of variational principles that require minimization of a suitable functional, such as the energy or the lagrangian.

Table 1 Regularization in early vision

Problem	Regularization principle
Edge detection	$\int [(Sf - i)^2 + \lambda (f_{xx})^2]\, dx$
Optical flow (area based)	$\int [i_x u + i_y v + i_t)^2 + \lambda (u_x^2 + u_y^2 + v_x^2 + v_y^2)]\, dx\, dy$
Optical flow (contour based)	$\int [(V \cdot N - V^N)^2 + \lambda ((\partial/\partial_s) V)^2]\, ds$
Surface reconstruction	$\int [S \cdot f - d)^2 + \lambda (f_{xx}^2 + 2 f_{xy}^2 + f_{yy}^2)^2]\, dx\, dy$
Spatiotemporal approximation	$\int [(S \cdot f - i)^2 + \lambda (\nabla f \cdot V + f_t)^2]\, dx\, dy\, dt$
Colour	$\|I^\nu - Az\|^2 + \lambda \|Pz\|^2$
Shape from shading	$\int [(E - R(f, g))^2 + \lambda (f_x^2 + f_y^2 + g_x^2 + g_y^2)]\, dx\, dy$
Stereo	$\int \{[\nabla^2 G * (L(x, y) - R(x + d(x, y), y))]^2 + \lambda (\nabla d)^2\}\, dx\, dy$

Some of the early vision problems that have been solved in terms of variational principles. The first five are standard quadratic regularization principles. In edge detection[26,27] the data on image intensity ($i = i(x)$) (for simplicity in one dimension) are given on a discrete lattice: the operator S is the sampling operator on the continuous distribution f to be recovered. A similar functional may be used to approximate time-varying imagery. The spatio–temporal intensity to be recovered from the data $i(x, y, t)$ is $f(x, y, t)$; the stabilizer imposes the constraint of constant velocity V in the image plane (ref. 61). In area-based optical flow[18], i is the image intensity, u and v are the two components of the velocity field. In surface reconstruction[21,22] the surface $f(x, y)$ is computed from sparse depth data $d(x, y)$. In the case of colour[32] the brightness is measured on each of three appropriate colour coordinates $I^\nu (\nu = 1, 2, 3)$. The solution vector z contains the illumination and the albedo components separately; it is mapped by A into the ideal data. Minimization of an appropriate stabilizer enforces the constraint of spatially smooth illumination and either constant or sharply varying albedo. For shape from shading[19] and stereo (T.P. and A. Yuille, unpublished), we show two non-quadratic regularization functionals. R is the reflectance map, f and g are related to the components of the surface gradient, E is the brightness distribution[19]. The regularization of the disparity field d involves convolution with the laplacian of a gaussian of the left (L) and the right (R) images and a Tikhonov stabilizer corresponding to the disparity gradient.

機械学習の文脈では，重み崩壊 weight decay などと呼ばれてきたが，標準正則化理論とオイラー=ラグランジェ方程式との関係で言うと見通しが良い。 オイラー=ラグランジェ方程式はラグランジェの未定乗数法により変分法，条件付き最適化，になるので，物理学，経済学，などへの応用も盛んである。

重み減衰(Krogh and Hertz 1991) については，古典的なニューラルネットワークで研究されてきた From https://machinelearningmastery.com/how-to-reduce-overfitting-in-deep-learning-with-weight-regularization/

39 Softmax 関数

$$p(x_i) = \frac{\exp(x_i)}{\sum_j \exp(x_j)} \tag{7}$$

ロジスティックシグモイド関数はソフトマックス関数の特別な場合と見なしうる

$$p(x) = \frac{1}{1 + \exp(-x)} \tag{8}$$

Softmax 関数は最上位層で用いられる。したがって特徴抽出を繰り返した多層ニューラルネットワークの特徴表現を **記号接地** symbol grounding する関数と見なしうる。

また，ソフトマックス関数は，交差エントロピー誤差 cross entropy error 関数と親和する。交差エントロピー誤差関数は出力が 2 値の場合に用いられる:

$$\mathrm{CE} = t \log(y) + (1 - t) \log(y) \tag{9}$$

出力が 2 値ではなく一般的な確率とみなせる場合にはソフトマックス関数の対数尤度を最大化することが行われる(Hinton 1989)。

40 腹側経路と背側経路

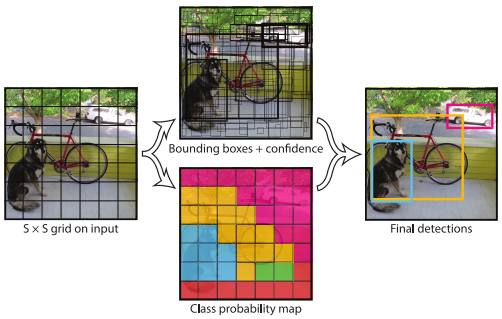


Figure 2: The Model. Our system models detection as a regression problem. It divides the image into an $S \times S$ grid and for each grid cell predicts B bounding boxes, confidence for those boxes, and C class probabilities. These predictions are encoded as an $S \times S \times (B * 5 + C)$ tensor.

YOLO の概念図

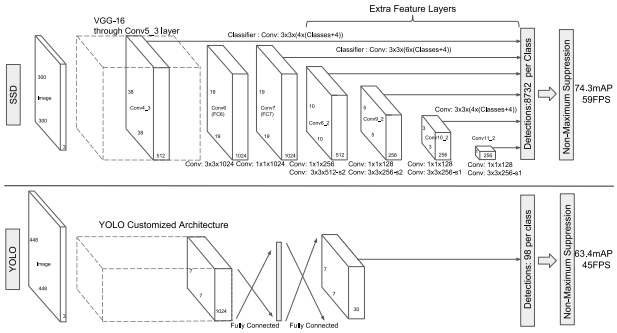
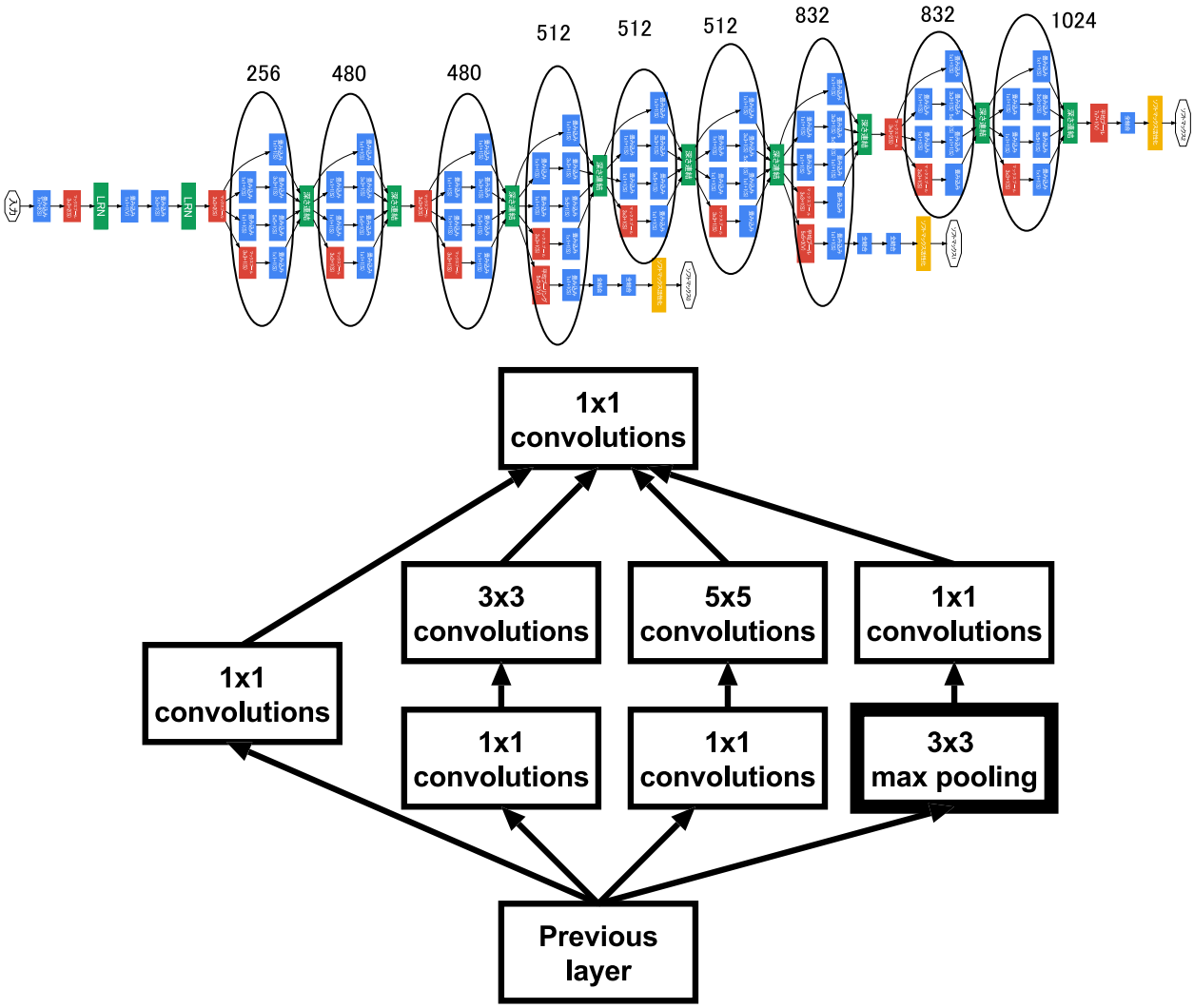


Fig. 2: A comparison between two single shot detection models: SSD and YOLO [5]. Our SSD model adds several feature layers to the end of a base network, which predict the offsets to default boxes of different scales and aspect ratios and their associated confidences. SSD with a 300×300 input size significantly outperforms its 448×448 YOLO counterpart in accuracy on VOC2007 `test` while also improving the speed.



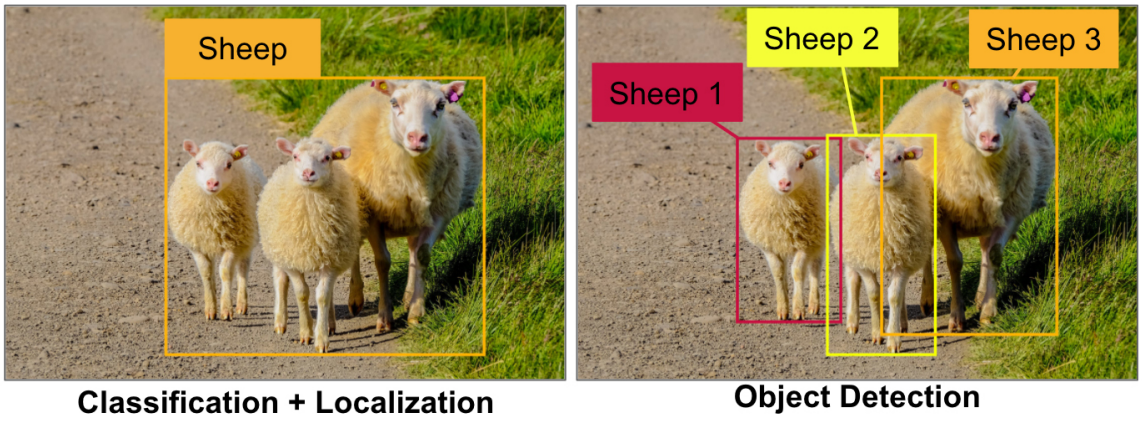
Inception モジュール

41 セマンティックセグメンテーションとインスタンスセグメンテーション

– 2020–0117 追記, セグメンテーションには semantic, instance, object, class の三種類ある。 object と class の例は ## Image segmentation: Object, class, instance http://host.robots.ox.ac.uk/pascal/VOC/voc2011/segexamples/index.html を参照のこと これによれば

- the object segmentation: pixel indices correspond to the first, second, third object etc.
- the class segmentation: pixel indices correspond to classes in alphabetical order (1=aeroplane, 2=bicycle, 3=bird, 4=boat, 5=bottle, 6=bus, 7=car , 8=cat, 9=chair, 10=cow, 11=diningtable, 12=dog, 13=horse, 14=motorbike, 15=person, 16=potted plant, 17=sheep, 18=sofa, 19=train, 20=tv/monitor)

PASCAL VOC は 20 分類なのでこれで全て。物体が切り分けられていれば, 物体切り分け, 切り分けた画素の示す物体が何であるかが特定できれば class 切り分けである。



Source: Introducing capsule networks From

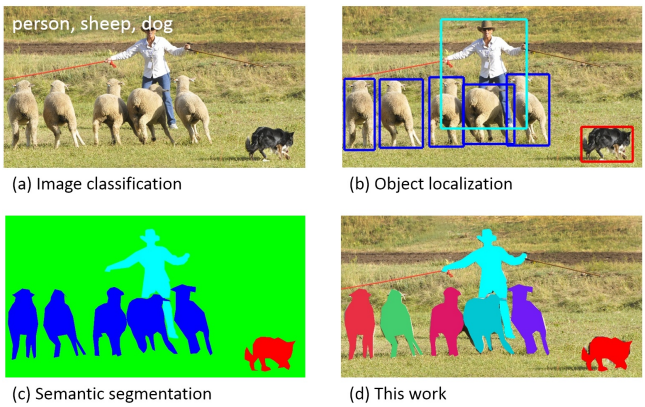
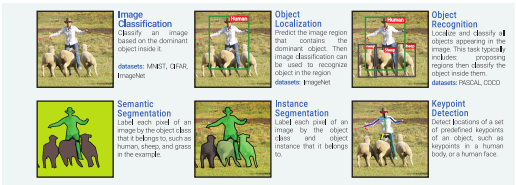
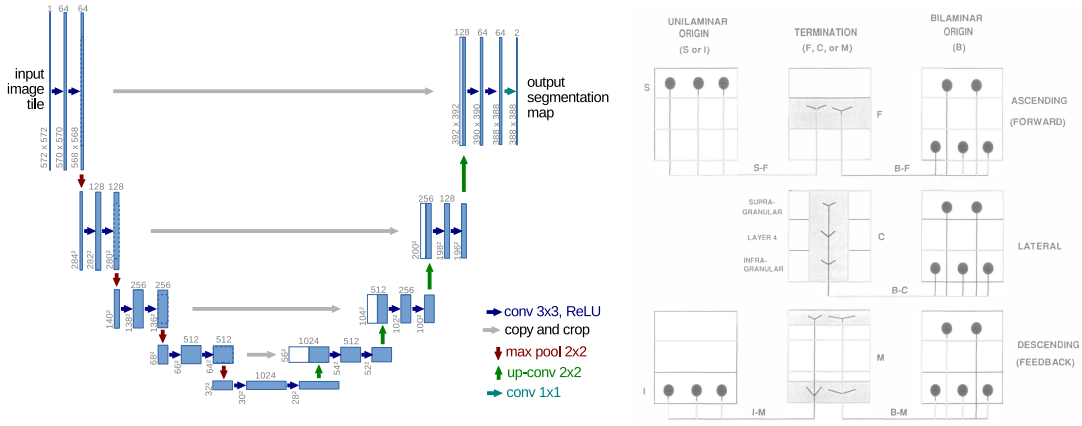


Fig. 1: While previous object recognition datasets have focused on (a) image classification, (b) object bounding box localization or (c) semantic pixel-level segmentation, we focus on (d) segmenting individual object instances. We introduce a large, richly-annotated dataset comprised of images depicting complex everyday scenes of common objects in their natural context.

(Lin et al. 2014) Fig. 1



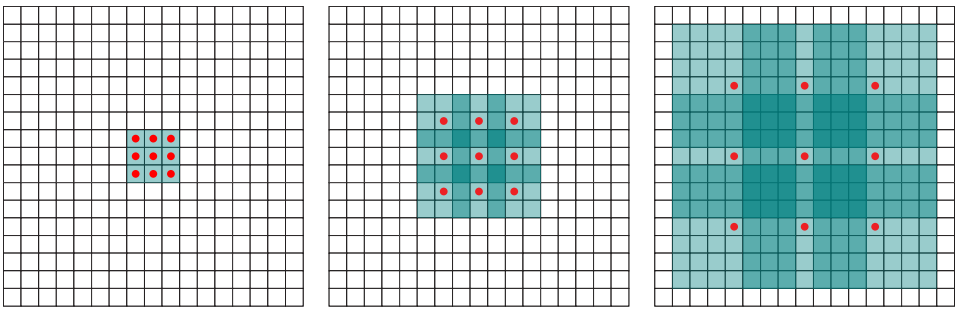
42 U-Net と afferent/efferent connections



Left:(Ronneberger, Fischer, and Brox 2015), right:(Felleman and Essen 1991)

43 拡張畳み込み Dialated Convolutions

アトラス畳み込みとも呼ばれる 拡張畳み込み (dialated convolutions) は解像度を失うことなく指数関数的に拡大する受容野をサポートする クロネッカー因子畳み込みフィルター (KFC: Kronecker Factor Convolutions) の一般化である。拡張畳み込みは，アップサンプリングされたフィルターを使用する通常の畳み込みである。 拡張率 l は，アップサンプリング係数を制御する係数である。 下図に示すように l 拡張畳み込みを積み重ねると受容野は指数関数的に拡大する。しかしフィルタのパラメータ数は線形にしか増大しない。 拡張畳み込みにより，任意の解像度で効率的な高密度特徴抽出が可能となる。 通常の畳み込み操作は $l = 1$ の畳み込みとみなしうる。

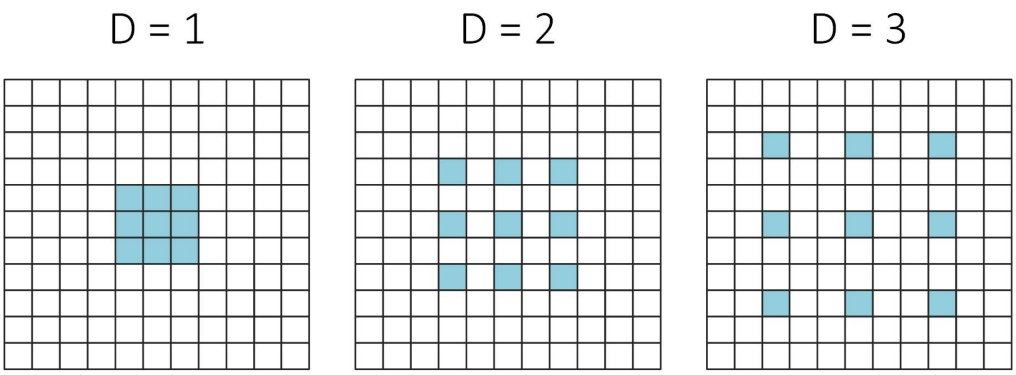


(a) (b) (c)

Figure 1: Systematic dilation supports exponential expansion of the receptive field without loss of resolution or coverage. (a) F_1 is produced from F_0 by a 1-dilated convolution; each element in F_1 has a receptive field of 3×3. (b) F_2 is produced from F_1 by a 2-dilated convolution; each element in F_2 has a receptive field of 7×7. (c) F_3 is produced from F_2 by a 4-dilated convolution; each element in F_3 has a receptive field of 15×15. The number of parameters associated with each layer is identical. The receptive field grows exponentially while the number of parameters grows linearly.

From (Yu and Koltun 2016) Fig. 1

実際には，通常の畳み込みを行う前にフィルタを拡張することと同義である。すなわち，空の要素をゼロで埋めながら，膨張率に応じてサイズを拡張する。 フィルタの重みは，膨張率 l が 1 より大きい場合，隣接していない遠くの要素に一致する。 下図は，拡張フィルタの例を示している。



44 クロネッカー積 Kronecker Product

$\mathbf{A} \in \mathbb{R}^{m_1 \times n_1}$ と $\mathbf{B} \in \mathbb{R}^{m_2 \times n_2}$ とを所与の行列とすれば， クロネッカー積 $\mathbf{A} \otimes \mathbf{B}$ は $m \times n$ 行列となる。ここで $m = m_1 m_2, n = n_1 n_2$:

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & \cdots & a_{1n_1}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{m_1 1}\mathbf{B} & \cdots & a_{m_1 n_1}\mathbf{B} \end{bmatrix}. \tag{10}$$

クロネッカー積の行列のサイズを変更する演算 $\mathbf{X}$ を用いて次式で定義される:

$$(\mathbf{A} \otimes \mathbf{B})\text{vect}(\mathbf{X}) = \text{vect}\left(\mathbf{B}\mathbf{X}\mathbf{A}^\top\right), \tag{11}$$

45 情報量最大化原理に基づく正規分布

$p(x)$ を確率密度関数 pdf として，期待値と分散を以下のように定義する

$$\int p(x)\,dx = 1 \tag{12}$$

$$\int xp(x)\,dx = \mu \tag{13}$$

$$\int (x-\mu)^2 p(x)\,dx = \sigma^2 \tag{14}$$

- ラグランジェの未定乗数法を用いて，$p(x)$ の変分を定義:

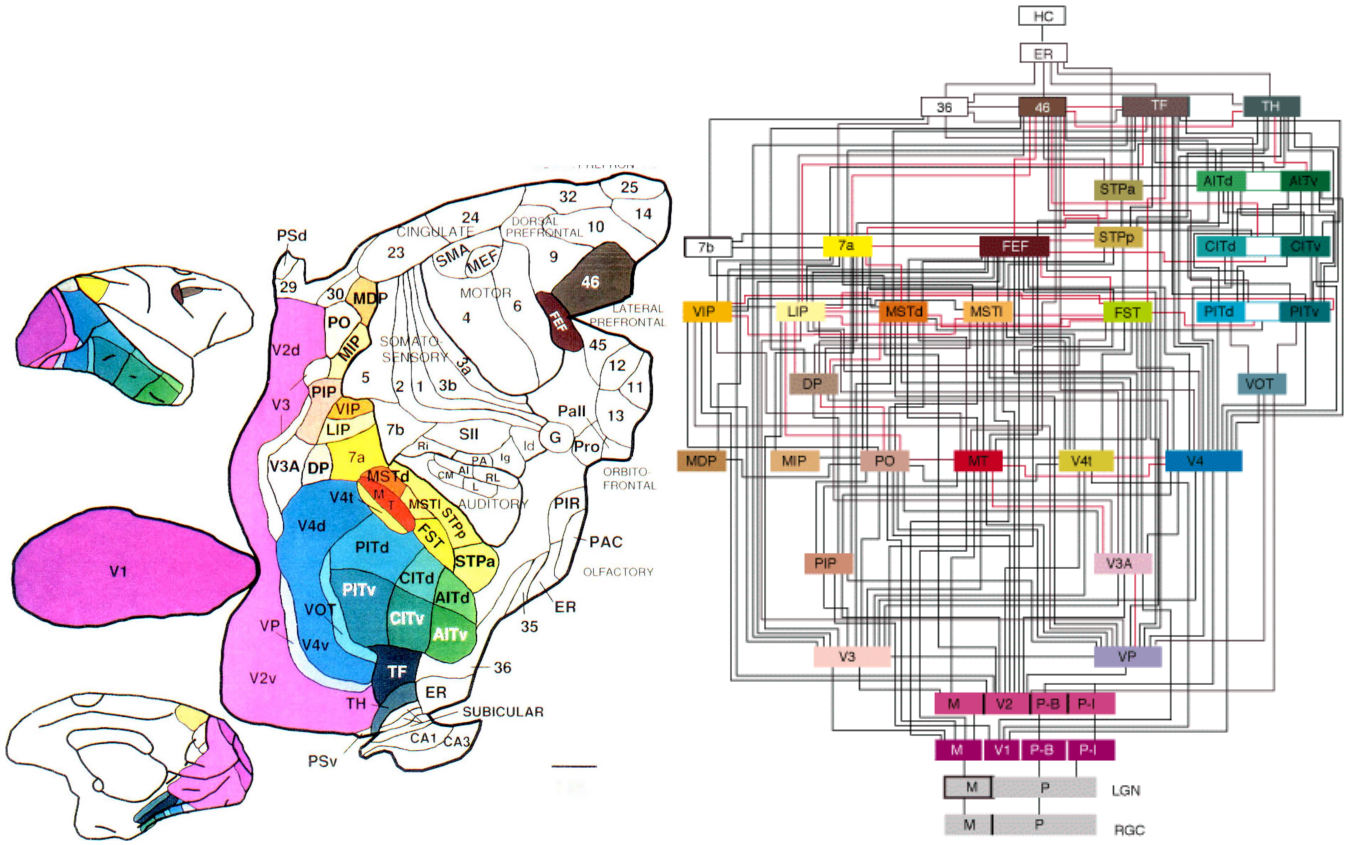
$$-\int p(x)\log p(x)\,dx + \lambda_1\left(p(x)\,dx - 1\right) + \lambda_2\left(\int xp(x)\,dx - \mu\right) + \lambda_3\left(\int (x-\mu)^2 p(x)\,dx - \sigma^2\right) \tag{15}$$

微分してゼロとおいて整理すれば，正規分布を得る:

$$p(x) = \exp\left\{-1 + \lambda_1 + \lambda_2 x + \lambda_3 (x-\mu)^2\right\} \tag{16}$$

すなわち，情報量最大化原理から導出される分布が正規分布である。

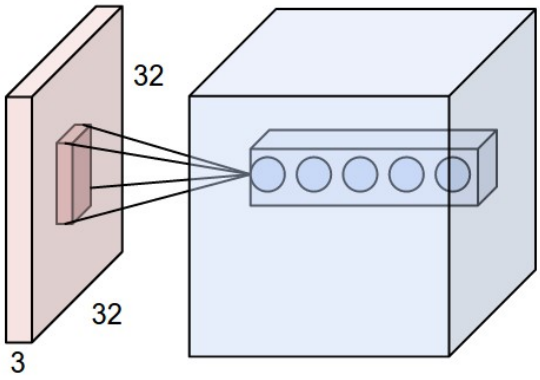
46 ResNet と Van Essen



47 畳み込み層

畳み込み層は ほとんどの計算負荷を処理する畳み込みネットワークの中核要素である。

1. 畳み込み層のパラメータは 学習可能なフィルター群で構成される。
2. 全フィルタは空間的に（幅 x と高さ y ）小さくなる。この空間情報の大きさはすべての深さで同一である。
3. たとえば 畳み込み層一般的なフィルタサイズは 5x5x3（幅 x と高さ y が 5 画素分 画像の深さ，この場合，色チャンネルが 3 ）である。
4. 前向き処理中に 各フィルタを入力情報の幅 x と高さ y にわたってスライド（畳み込み）し フィルタのエントリと任意の位置の入力間のドット積を計算する。
5. 入力ボリュームの幅 x と高さ y でフィルタをスライドさせると すべての空間位置でそのフィルタの応答を提供する 2 次元の活性化マップが生成される。
6. 直感的には ネットワークは 最初の層のある方向のエッジやある色のブロブ(斑点) 最終的にはネットワークの上位層 のハニカムまたはホイールのようなパターン全体 ある種の視覚的特徴を見たときに活性化するフィルタを学習する。
7. 各畳込み層にフィルタ集合が形成され それぞれが個別の 2 次元の活性化地図を学習する。
8. これらの活性化地図を深さ次元に沿って積み上げ 出力量を生成する。

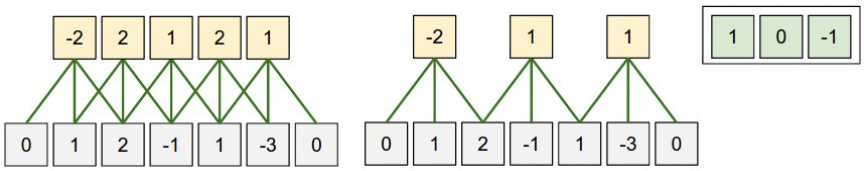


入力画像例（32x32x3） および第一畳み込み層のニューロンの例。 畳み込み層の各神経細胞は 空間的に入力情報の局所領域にのみ接続されている。完全な深さ（つまり，すべての色チャネル）に接続されている。 深さ（色チャンネル）に沿って複数のニューロン（この例では 5）がありすべて入力内の同じ領域，すなわち同一受容野を形成する。

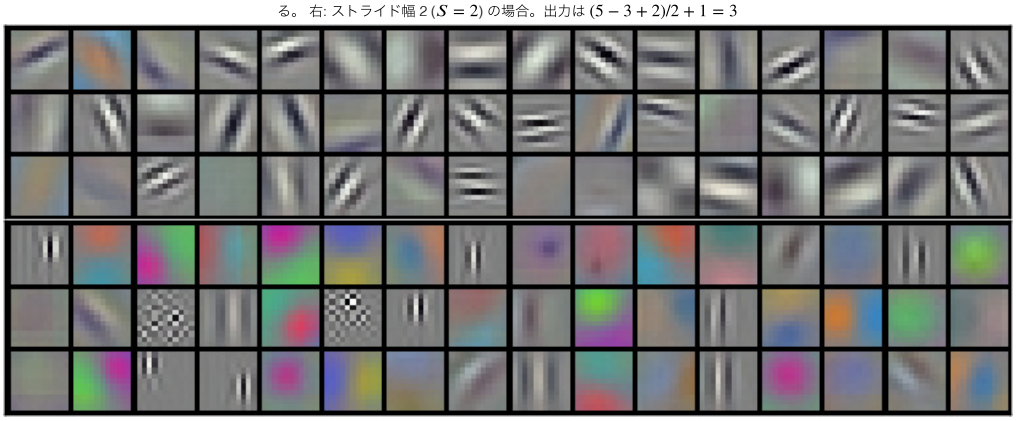
空間配置. 畳み込み層の各神経細胞の入力側の接続性については説明したが 出力側に存在するニューロンの数や配置方法についてはまだ説明されていない。 出力層の大きさは **深さ** **ストライド** および **ゼロパディング** の 3 つのハイパーパラメーターによって制御される。

1. 出力の深さ（特徴数，チャンネル数）はハイパーパラメータである。これは 使用するフィルター数に対応し それぞれが入力で異なる特徴を学習する。 たとえば 最初の畳み込み層が入力として生画像である場合 さまざまな方向のエッジ または 色の塊 が存在すると 深さ次元 に沿った 異なる神経細胞が活性化するようになる。 同じ入力領域を 深度列 (特徴数) とみなしている神経細胞のセットを参照する (ファイバという用語も用いられる）。
2. 次に フィルタをスライドさせる。 ストライドが 1 の場合 フィルターを 1 画素ずつ移動することを意味する。 ストライドが 2 （実際にはまれ）の場合フィルタ は一度に 2 画素ずつ ジャンプする。 これにより 空間的に 小さな出力生成される。
3. 入力ボリュームの境界をゼロで埋めると便利な場合がある。 このゼロパディングのサイズはハイパーパラメータである。 ゼロパディングの優れた機能は 出力ボリュームの空間サイズを制御できることである (ほとんどの場合 入力ボリュームの空間サイズを正確に保持するために使用する)

入力ボリューム W, 畳み込み層ニューロンの受容野サイズ F, ストライド幅 S, 境界で使用されるゼロパディングの幅 Pの関数として 出力ボリュームの空間サイズを計算できる。 ニューロンの数を計算するための式は $(W - F + 2P)/S + 1$ で与えられる。 たとえば入力画像サイズ 7x7，フィルタサイズ 3x3, ストライド幅 1, パディング幅 0 の場合 5x5 出力が得られる。 ストライド 2 ならば 3x3 の出力が得られる。 別の例を以下に示す。



空間配置の例: 簡単のため X 軸のみ表示してある。受容野サイズは 3 ($F = 3$), 入力サイズは 5 ($W = 5$), ゼロパディング幅は 1 ($P = 1$)。 左: ストライド幅 1 ($S = 1$) の場合。出力は $(5 - 3 + 2)/1 + 1 = 5$ となる。 右: ストライド幅 2 ($S = 2$) の場合。出力は $(5 - 3 + 2)/2 + 1 = 3$



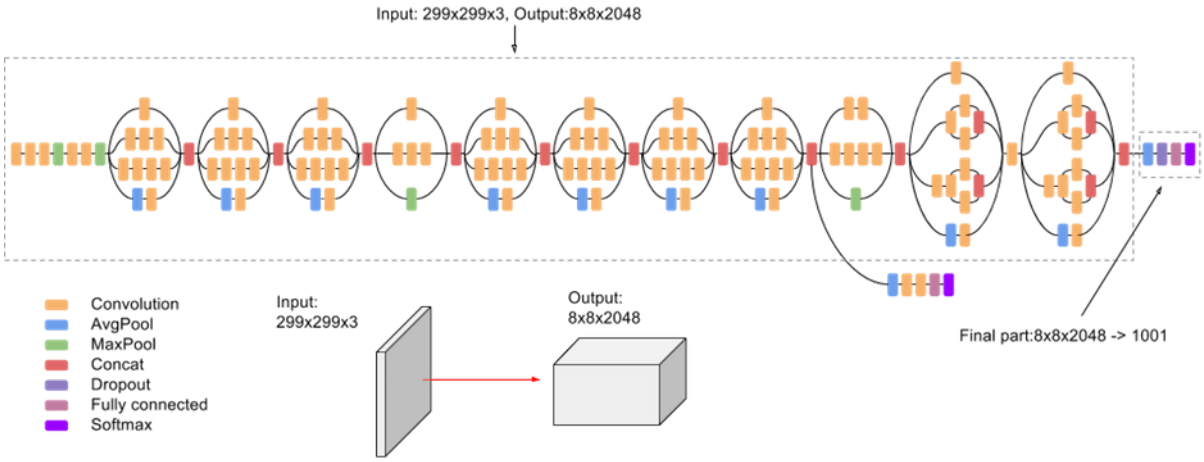
Krizhevsky らによる学習済みフィルターの例: 96 個のフィルタはそれぞれサイズが [11x11x3] それぞれが 1 つの深度スライスの 55x55 ニューロンによって共有されている。 パラメータ共有の仮定は合理的であもある。 画像の特定の場所で水平エッジを検出することが重要な場合 画像の並進不変の構造により 他の場所でも直感的に役立つはずである。 したがって 畳込み層の出力が 55x55 の異なる場所のすべてで水平エッジを検出するために再学習する必要はない。

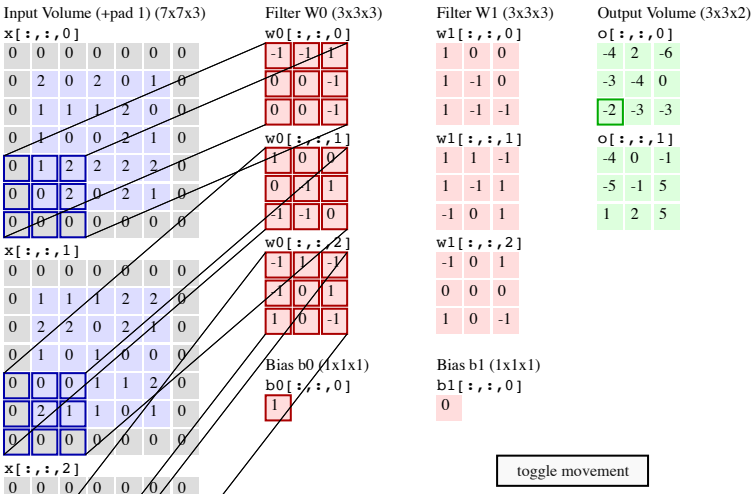
パラメータ共有の仮定が意味をなさない場合も存在する。 特に 畳込みネットワークへの入力画像が特定の中心構造を持っている場合に当てはまる。たとえば 画像の片側で他とはまったく異なる特徴を学習する必要がある場合などである。 実用的な例としては 入力が画像の中心に顔がある場合である。 さまざまな空間固有の場所で さまざまな目固有または髪固有の特徴を学習することが期待される場合があろう。 この場合 パラメータ共有条件を緩和し 代わりに単に層を局所接続層にする場合がある。

誤差逆伝播法: 畳み込み演算（データと重みの両方）の逆方向パスも畳み込み（ただし 空間的に反転したフィルターを使用）。 おもちゃの例を使用して 1 次元の場合に導き出すのは簡単

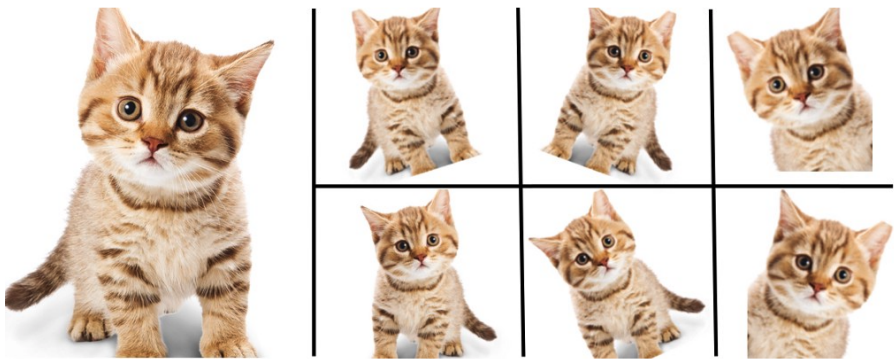
1×1 畳み込み: Network in Network で最初に調査されたように いくつかの論文は 1×1 畳み込みを使用している。 信号処理のバックグラウンドのある人は 1×1 畳込みの解釈に混乱する。 通常信号は 2 次元であるため $11 $の畳み込みは意味をなさないからである（単なる点ごとのスケーリング）。 ただし畳み込みニューラルネットワークでは当てはまらない。3 次元ボリュームを操作することと フィルターが常に入力ボリュームの深さ全体に広がることを覚えておく必要がある。 たとえば入力が $[32 \times 32 \times 3]$ の場合 1×1 の畳み込みを実行すると 事実上 3 次元の内積が実行される（入力の深さが 3 チャネルであるため）。

拡張畳込み: 最近の発展（たとえば Fisher Yu nad Vladlen Koltun を参照）は，*拡張* と呼ばれるもう 1 つのハイパーパラメーターを畳込み層に導入した。 これまでは 連続した畳込みフィルターについてのみ説明してき。 ここで *膨張* と呼ばれる 各セル間にスペースがあるフィルターを使用することが可能である。 例として ある次元では サイズ 3 のフィルター `w` は 入力 `x` に対して次の演算を行う: `w[0] * x[0] + w[1] * x[1] + w[2] * x [2]` . これは 0 の膨張である。 膨張 1 の場合 フィルターは `w[0] * x[0] + w[1] * x[2] + w[2] * x[4]` を計算する。 つまり 畳込み演算を実施する際に間隙 1 が存在する。 これは いくつかの設定で 0 拡張フィルターと併用すると非常に便利である。より少ない層でより積極的に入力全体の空間情報をマージできるからである。 たとえば 2 つの 3×3 の畳込み層を上下に重ねると 2 番目の層のニューロンは入力の 5×5 パッチの関数となる（*有効な受容野* のニューロンは 5×5 ）。 拡張畳み込みを使用すると この効果的な受容野ははるかに速く成長する。





48 データ拡張



Enlarge your Dataset

From Data Augmentation | How to use Deep Learning when you have Limited Data — Part 2

49 SGD

Bottou (2010) (http://leon.bottou.org/projects/sgd)

	()	(%)
	$\lambda = 10^{-4}$	
SVMLight	23,642	6.02
SVMPerf	66	6.03
SGD	1.4	6.02
	$\lambda = 10^{-5}$	
LibLinear ($\rho = 10^{-2}$)	30	5.68
LibLinear ($\rho = 10^{-3}$)	44	5.70
SGD	2.3	5.66

(Bottou 2010), (Bottou and Bousquet 2008)

50 転移学習と蒸留

温度 T のソフトマックス関数:

$$q_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_i/T)} \tag{17}$$

温度が高ければ，鈍った S 字曲線になる。その時の差分を考える

$$\frac{\partial C}{\partial z_i} = \frac{1}{T}(q_i - p_i) = \frac{1}{T}\left(\frac{e^{z_i/T}}{\sum_j e^{z_i/T}} - \frac{e^{v_i/T}}{\sum_j e^{v_j/T}}\right) \tag{18}$$

$$\frac{\partial C}{\partial z_i} \approx \frac{1}{T}\left(\frac{1 + z_i/T}{N + \sum_j z_j/T} - \frac{1 + v_i}{N + \sum_j v_j/T}\right) \tag{19}$$

(3)

$$\frac{\partial C}{\partial z_i} \approx \frac{1}{NT^2}(z_i - v_i) \tag{20}$$

Bibliography

Bottou, Léon. 2010. “Large-Scale Machine Learning with Stochastic Gradient Descent.” In *Proceedings of the 19th International Conference on Computational Statistics (COMPSTAT2010)*, edited by Yves Lechevallier and Gilbert Saporta, 177–87. Paris, France: Springer. http://leon.bottou.org/papers/bottou-2010.

Bottou, Léon, and Olivier Bousquet. 2008. “Learning Using Large Datasets.” In *Mining Massive Datasets for Security, Nato Asi Workshop Series*. Amsterdam, Netherland: IOS Press.

Cadieu, Charles F., Ha Hong, Daniel L. K. Yamins, Nicolas Pinto, Diego Ardila, Ethan A. Solomon, Najib J. Majaj, and James J. DiCarlo. 2014. “Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition.” *ArXiv*.

Cichy, Radoslaw Martin, Aditya Khosla, Dimitrios Pantazis, Antonio Torralba, and Aude Oliva. 2016. “Comparison of Deep Neural Networks to Spatio-Temporal Cortical Dynamics of Human Visual Object Recognition Reveals Hierarchical Correspondence.” *Nauture Scientific Report* 6:27755: 1–13. https://doi.org/10.1038/srep27755.

Essen, David C. Van, and Jack L. Gallant. 1994. “Neural Mechanisms of Form and Motion Processing in the Primate Visual System.” *Neuron* 13: 1–10.

Felleman, Daniel J., and David C. Van Essen. 1991. “Distributed Hierarchical Processing in the Primate Cerebral Cortex.” *Cerebral Cortex* 1: 1–47.

Fukushima, Kunihiko. 1980. “Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position.” *Biological Cybernetics* 36: 193–202.

Girosi, Federico, Michael Jones, and Tomaso Poggio. 1995. “Regularization Theory and Neural Networks Architectures.” *Neural Computation* 7 (2): 219–69. https://doi.org/10.1162/neco.1995.7.2.219.

Girshick, Ross. 2015. “Fast R-CNN.” *ArXiv:1504.08083*.

Girshick, Ross, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. “Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation.” In *Proceedings of Computer Vision and Pattern Recognition Conference (CVPR)*. Columbus, Ohio, USA.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. “Deep Residual Learning for Image Recognition.” *ArXiv:1512.033835*.

Hinton, Geoffrey E. 1989. “Connectionist Learning Procedures.” *Artificial Intelligence* 40: 185–234.

Hinton, Geoffrey E., Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R. Salakhutdinov. 2012. “Improving Neural Networks by Preventing Co-Adaptation of Feature Detectors.” *The Computing Research Repository (CoRR)* abs/1207.0580.

Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. 2015. “Distilling the Knowledge in a Neural Network.” *arXiv Preprint*.

Hu, Ronghang, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. 2017. “Learning to Reason: End-to-End Module Networks for Visual Question Answering.” *ArXiv Preprint*. https://arxiv.org/abs/1704.05526.

Huang, Gao, Zhuang Liu, and Laurens van der Maaten. 2018. “Densely Connected Convolutional Networks.” *ArXiv*.

Hubel, David, and Torsen N. Wiesel. 1959. “Receptive Fields of Single Neurones in the Cat’s Striate Cortex.” *Journal of Physiology* 148: 574–91.

———. 1962. “Receptive Fields, Binocular Interaction and Functional Architecture in the Cat’s Visual Cortex.” *Journal of Physiology* 160: 106–54.

———. 1968. “Receptive Fields and Functional Architecture of Monkey Striate Cortex.” *Journal of Physiology* 195: 215–43.

Ioffe, Sergey, and Christian Szegedy. 2015. “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift.” *ArXiv:1502.03167*.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. 2012. “ImageNet Classification with Deep Convolutional Neural Networks.” In *In Advances in Neural Information Processing Systems 25*, edited by F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger. Montréal, Canada. http://papers.nips.cc/book/advances-in-neural-information-processing-systems-25-2012.

Krogh, Anders, and John A Hertz. 1991. “A Simple Weight Decay Can Improve Generalization.” In *In Proceedings of Advances in Neural Information Processing Systems*, edited by J. E. Moody, S. J. Hanson, and R. P. Lippmann, 4:950–57. Denver, Colorado, USA.

LeCun, Yann, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. “Gradient-Based Learning Applied to Document Recognition.” *Proceedings of the IEEE* 86: 2278–2324. https://doi.org/10.1109/5.726791.

Lin, Tsung-Yi, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2014. “Microsoft Coco: Common Objects in Context.” *ArXiv*.

Liu, Wei, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. 2016. “SSD: Single Shot Multibox Detector.” *arXiv Preprint arXiv:1512.02325*. https://github.com/weiliu89/caffe/tree/ssd.

Marblestone, Adam H., Greg Wayne, and Konrad P. Kording. 2016. “Towards an Integration of Deep Learning and Neuroscience.” *BioRxiv*. https://doi.org/http://dx.doi.org/10.1101/058545.

Poggio, Tomaso, Vincent Torre, and Christof Koch. 1985. “Computational Vision and Regularization Theory.” *Nature* 317: 314–19.

Redmon, Joseph, and Ali Farhadi. 2016. “YOLO9000: Better, Faster, Stronger.” *arXiv Preprint arXiv:* http://pjreddie.com/yolo9000/.

Ren, Shaoqing, Kaiming He, Ross Girshick, and Jian Sun. 2015. “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks.” *ArXiv::1504.01497*.

Riesenhuber, Maximilian, and Tomaso Poggio. 1999. “Hierarchical Models of Object Recognition in Cortex.” *Nature Neuroscience* 2 (11): 1019–25.

Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. 2015. “U-Net: Convolutional Networks for Biomedical Image Segmentation.” *ArXiv Preprint*, arXiv:1505.04597v1 [cs.CV].

Santosh Divvala, Joseph Redmon adn, Ross Girshick, and Ali Farhadi. 2016. “You Only Look Once: Unified, Real-Time Object Detection.” *arXiv Preprint arXiv:1612.08242*. http://pjreddie.com/yolo/.

Serre, Thomas, Lior Wolf, and Tomaso Poggio. 2005. “Object Recognition with Features Inspired by Visual Cortex.” In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 994–1000. San Diego, CA, USA. https://doi.org/10.1109/CVPR.2005.254.

Simonyan, Karen, and Andrew Zisserman. 2015. “Very Deep Convolutional Networks for Large-Scale Image Recognition.” In *Proceedings of the International Conference on Learning Representations (ICLR)*, edited by Yoshua Bengio and Yann LeCun. San Diego, CA, USA.

Szegedy, Christian, Wei Liu, Yangqing Jia, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. “Going Deeper with Convolutions.” In *Computer Vision and Pattern Recognition (CVPR)*. Boston, MA, USA.

Tan, Mingxing, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V. Le. 2019. “MnasNet: Platform-Aware Neural Architecture Search for Mobile.” *ArXiv Preprint*. arXiv:1807.11626v3 [cs.CV].

Thorpe, Simon J., and Michéle Fabre-Thorpe. 2001. “Seeking Categories in the Brain.” *Science* 291: 260–62.

Yamins, Daniel L. K., and James J DiCarlo. 2016. “Using Goal-Driven Deep Learning Models to Understand Sensory Cortex.” *Nature Neuroscience* 19 (3): 356–65.

Yamins, Daniel L. K., Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. 2014. “Performance-Optimized Hierarchical Models Predict Neural Responses in Higher Visual Cortex.” *Proceedings National Academy of Science. USA, Neurobiology*, June, 8619–24.

Yu, Fisher, and Vladlen Koltun. 2016. “Multi-Scale Context Aggreation by Dilated Convolutions.” In *Proceedings in the International Conference on Learning Representations (ICLR)*. San Juan, Puerto Rico.