

視覚系のモデルとしての畳み込みニューラルネットワーク: 過去, 現在, そして未来

Convolutional Neural Networks as a Model of the Visual System: Past, Present, and Future

Grace W. Lindsay

要約

畳み込みニューラルネットワーク (CNN) は, 生物学的な視覚研究の初期の成果から着想を得たものである。以来, CNN はコンピュータビジョンの分野で成功を収め, 視覚タスクにおける神経活動と行動の両方の最先端モデルとなっている。このレビューでは, CNN の文脈において, 計算論的神経科学における優れたモデルとはどういうものか, また, モデルが洞察をもたらすさまざまな方法について紹介する。具体的には, CNN の起源と, 生物学的視覚のモデルとして検証する方法を取り上げている。さらに, CNN を理解し実験することで生物学的視覚について何を学ぶことができるかを詳しく説明し, 基本的な物体認識を超えて視覚研究に CNN を使用する新たな機会について論じている。

1. はじめに

計算モデルは神経科学においていくつかの役割を果たす。計算モデルはシステムがどのように機能するかについての直感を直接テストする方法を提供することでその直感を検証することができる。また, 細部まで制御・計測できる理想的な実験場で, 新たな仮説を探る手段にもなります。さらにモデルは数学的な分析を行うことで問題となっているシステムを新たな領域で理解することができる。近年では, 畳み込みニューラルネットワーク (CNN) が視覚システムのモデルとしてこれらの役割をすべて果たしている。

この概説論文では CNN の起源, 視覚システムのモデルとして CNN を検証する方法, CNN を使って実験することで得られるもの, そして視覚研究で CNN を使うための新たな機会について説明する。重要なのは, 本稿では CNN の完全な概要を示すものではなく, 視覚研究における深層学習のすべての用途を広範囲にカバーすることを意図していないことである (この点については他の概説論文が読者にとって有用であろう; Kietzmann, McClure, & Kriegeskorte, 2019; Kriegeskorte & Golan, 2019; Serre, 2019; Storrs & Kriegeskorte, 2019; Yamins & DiCarlo, 2016; Kriegeskorte, 2015)。むしろ, 生物学的な視覚についての洞察や理解を得るために, モデルとしての CNN がどのような戦略で使われるのかを示すことを目的としています。

Kay (2018) によれば「機能論的モデル」はシステムに提供された同じ入力を与えられたシステムの出力を一致させようとするだけであるのに対し, “機械論的モデル”はシステムの実際の物理的な構成要素を並列にしたコンポーネントも使用しようとする」とのことである。これらの定義を用いてこのレビューでは視覚システムの“機械論的モデル”としての CNN の使用を対象とする。つまり2つのシステムの出力間の全体的な一致に加えて CNN の下位部分が視覚システムの 下位部分 と一致するように意図されていると仮定し論じていく。

2. CNN の来歴

CNN の歴史は, 神経科学と人工知能の両方に通じている。一般的な人工ニューラルネットワークと同様, CNN は脳から着想を得たアイデアが, コンピュータサイエンスやエンジニアリングとの相互作用によって結実した例である。

2.1 モデルの起源

20 世紀半ば ヒューベルとヴィーゼルは, 猫の一次視覚野 (V1) に 2 つの主要なタイプの細胞を発見した (Hubel & Wiesel, 1962)。1 つ目のタイプは単純細胞であり, 特定の空間的位置に置かれた明暗の棒に反応する。それぞれの細胞には, 最も発火しやすい棒の向きがあり, 棒の角度がこの選好の向きから変わると反応が低下する (向きの「同調曲線」)。第 2 のタイプである複雑細胞は反応のプロファイルがそれほど厳密ではない。これらの細胞は, 選好方向性を持っているが, 選好刺激近隣の複数の異なるバーに対しても同じように強く反応する。ヒューベルとウィーゼルは, これらの複雑細胞は, 選好方向は同じだが, 選好場所がわずかに異なる複数の単純細胞から入力を受けている可能性が高いと結論づけた (図1左)。

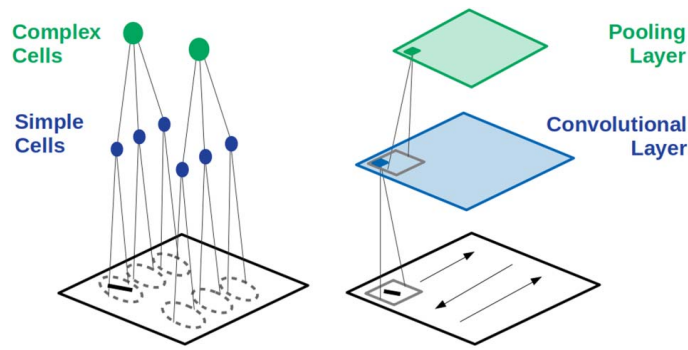


図 1. 視覚システムの構成要素とCNNの基本動作との関係。ヒューベルトウィーゼル (1962) は、単純細胞 (左図, 青) は、特定の方向の棒に最も強く反応する画像内の選好位置 (破線の楕円) を持つことを発見した。複雑細胞 (緑) は、多くの単純な細胞からの入力を受けるためより空間的に不変な反応をする。これらの操作は CNN (右図) でも再現される。最初の畳み込み層 (青) は、画像に畳み込み演算を適用して作られる。具体的には、画像のすべての場所に小さなフィルタ (灰色のボックス) を適用することで特徴地図を作成する。畳み込み層は適用されたフィルタの数だけ特徴地図を持つ (ここでは 1 つだけ)。それぞれの特徴地図 (灰色の箱) の小さな部分で最大の活性化を取ると、画像がダウンサンプリングされ、複雑細胞のような反応 (緑) が得られる。これは「最大値プーリング max-pooling」と呼ばれる操作である。最大値プーリングによって生成されたダウンサンプリングされた特徴地図がプーリング層を構成する。

1980 年 福島 はヒューベルとヴィーゼルの発見を視覚系の機能モデルに変換した (Fukushima, 1980)。このモデル「ネオコグニトロン」は現代の CNN の先駆けとなったモデルである。このモデルには主に 2 種類の細胞が含まれている。S 細胞は単純細胞にちなんで名付けられたもので、その基本的な機能を再現する。具体的には、入力画像の各位置に 2 次元の重みのグリッドを適用して、S 細胞の応答を作成する。S 細胞の「プレーン」は、すべての細胞が同じ視覚的特徴を共有し、1 つの層に複数のプレーンが存在するという、網膜局所的なレイアウトになっている。C 細胞 (複雑細胞にちなんで名付けられた) の応答は、同じ平面の異なる位置から来る複数の S 細胞の非線形関数である。

V1 の基本的な計算を表す単純なセルと複雑なセルの層の後、ネオコグニトロンは単純にそのプロセスを再び繰り返す。つまり、複雑な細胞の第 1 の層の出力が、第 2 の単純な細胞の層の入力となり、これを繰り返していく。これを何度か繰り返すことで、V1 だけでなく、腹側視覚経路全体の動作を模倣した階層的なモデルができあがります。このネットワークは「自己組織化」されており、ラベルのない画像を繰り返し見ることで重みが増加する。

1990 年代までには、視覚系の多くの同様の階層モデルが検討され、データとの関連付けが行われていました (Riesenhuber & Poggio, 2000)。その中でも最も著名なものの一つである「HMAX」は、単純な細胞の活動に対する単純な「最大」演算を用いて C 細胞の反応を決定するもので、画像の変化に対して非常にロバストなものでした。これらのモデルは、人間の心理物理学実験で使用されたものと同じ画像に適用できるため、モデルの動作は、人間が迅速に視覚的分類を行う能力と直接比較することができました。これにより、これらの階層モデルと視覚処理の最初の 100~150 ミリ秒との間に対応関係が見出されました (Serre et al. 2007)。また、このようなモデルは、複雑な形状の刺激に対する V4 ニューロンの反応を捉えるのにも適していた (Cadieu et al. 2007)。

2.2 コンピュータビジョンにおける CNN

今日私たちが知っている CNN はコンピュータビジョンの分野から生まれたものだが、そこにはヒューベルとヴィーゼルの研究からのインスピレーションがはっきりと見て取れる (図 1: Rawat & Wang, 2017)。現代の CNN は、まず入力画像に一連のフィルターを畳み込み、出力を整流することで、ネオコグニトロンの S 細胞の平面に似た「特徴地図」を生成する。その後、マックスプーリングを行い、複雑な細胞のような反応を作り出します。このパターンを何度か繰り返した後、非畳み込み完全連結層を追加し、最後の層にはタスクのカテゴリー数と同じ数のユニットを入れて、画像のカテゴリーラベルを出力する (図 2 下)。

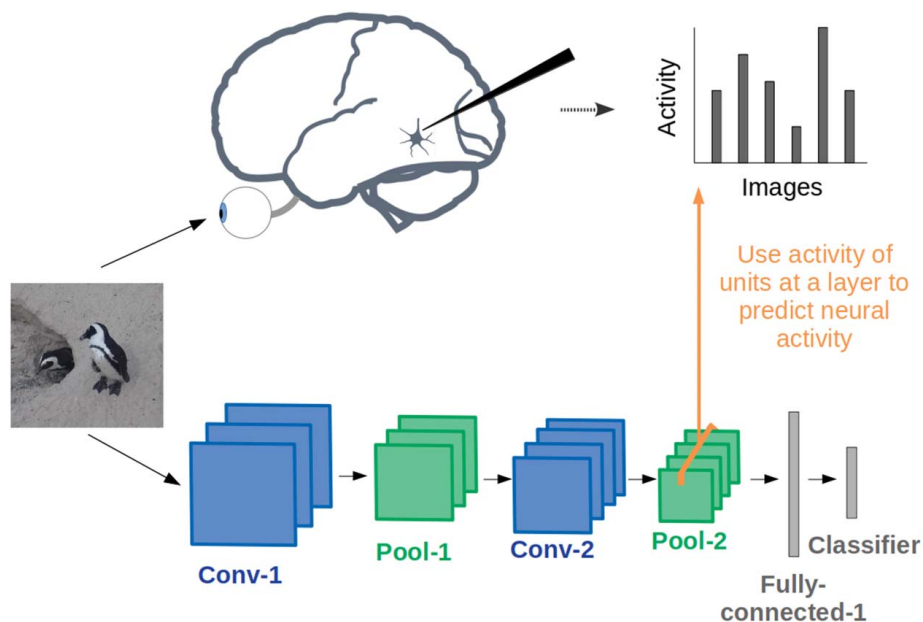


図2. CNN の表現を脳の表現と比較することで、モデルとしての妥当性を検証する。物体認識を行うために学習された CNN (下) は、いくつかの畳み込み層 (ここでは Conv-1 と Conv-2、それに続く最大プーリング層 Pool-1 と Pool-2) を含み、最後にカテゴリ数と同じ数のユニットを持つ完全連結層で終わる。同じ画像を CNN と動物に見せた場合 (上)、それぞれの反応を比較することができる。ここでは、脳の視覚野から記録された単一ユニットの活動が、特定の層の人工ユニットの活動から予測される。

CNN の能力が最初に大きく示されたのは1989年のことで、バックプロパゲーション・アルゴリズムを使って監視下で学習した小型の CNN が、手書きの数字を分類できることが示された (LeCun et al., 1989)。しかし、CNN が本格的に普及するのは 2012 年になってからのことで、バックプロパゲーション法で学習した8層のネットワーク (図3では“AlexNet”, “Standard Architecture”と呼ばれている) が、ImageNet の課題で最先端の性能をはるかに上回る結果を出した。ImageNet のデータセットは 100 万枚以上の実世界の画像で構成されており、1 枚の画像を 1,000 個の物体カテゴリに分類することが求められている。このネットワークの成功は、神経科学者が発見した視覚システムの基本的な特徴が、適切な学習アルゴリズムとデータを必要とするだけで、実際に視覚をサポートできることを示している。

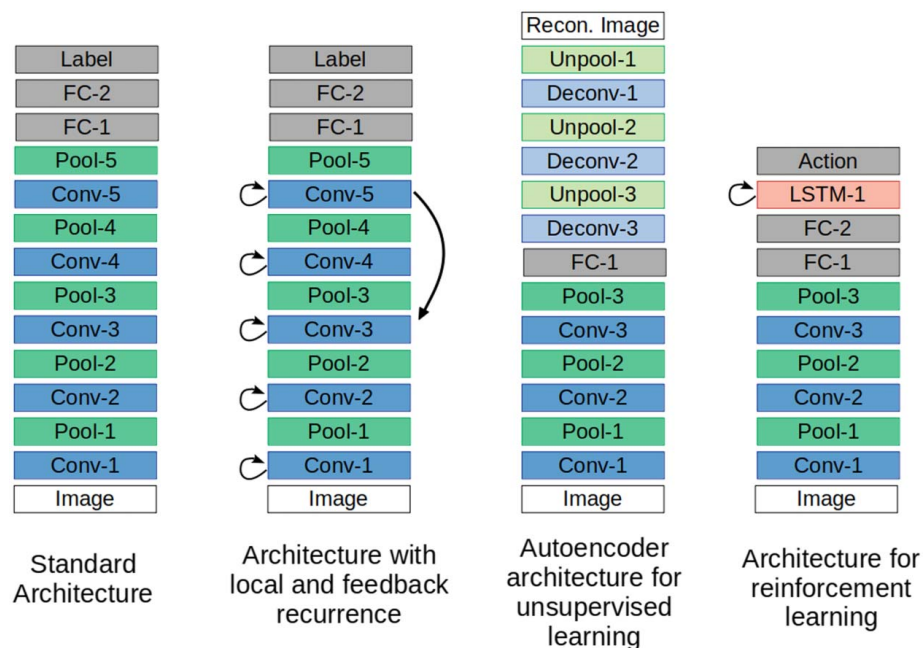


図3. アーキテクチャの例。ここで示されている標準的なアーキテクチャは、オリジナルの AlexNet に似ており、赤/緑/青の画像が入力として取り込まれ、最終層での活動がカテゴリラベルを決定します。2番目のアーキテクチャでは、各畳み込み層で局所再帰性が示され、フィードバック再帰性は第5層から第3層へと続く。このネットワークは、バックプロパゲーション・スルースタイム・アルゴリズムで学習される。3つ目のネットワークは、オートエンコーダーです。ここでのネットワークの目的は、入力画像を再現することです。重要なのは、完全連結層が画像よりもかなり低い次元であることで、ネットワークは関連する画像統計のよりコンパクトな表現を見つけなければならない。最後に、強化学習のためのアーキテクチャを示します。ここでは、ネットワークの目的は、現在の状態に関する情報を表す入力画像を、ネットワークが取るべき行動に変えることです。このアーキテクチャ

は、LSTM と呼ばれる特定のタイプの局所再帰ユニットによって強化されており、これによって記憶が埋め込まれています。

このデモから数年の間に、さまざまな CNN アーキテクチャが模索されており、主なパラメータとしては、ネットワークの深さ、プーリング層の配置、層ごとの特徴地図の数、学習手順、層を飛ばした残留接続が存在するかどうかなどが変化している (Rawat & Wang, 2017)。コンピュータビジョンのコミュニティでこれらのパラメータを探索する目的は、標準的な画像ベンチマークでより良い性能を発揮するモデルを作成することであり、ネットワークをより小さくしたり、より少ないデータで訓練したりすることが二次的な目標となっています。生物学との対応は推進要因ではない。

3. 視覚系のモデルとして CNN を検証する

CNN のアーキテクチャは、視覚システムのアーキテクチャと (意図的に) 直接的な類似性を持っています。これらのネットワークに入力される画像は、通常、まず正規化され、3 つの異なるカラーチャンネル (赤、緑、青) に分けられる。畳み込み-非線形性-プールの各スタックバンドルは、1 つの視覚領域 (通常 V1, V2, V4, IT などの腹側ストリームに沿ったもの) に近似していると考えることができ、それぞれが独自の網膜性と特徴地図を持っています。このような積み重ねにより、個々のニューロンの受容野が形成され、ネットワークの深部になるほどサイズが大きくなり、反応する画像の特徴も複雑になります。これまで述べてきたように、これらのアーキテクチャは、訓練を受ければ、画像を取り込み、人間の判断と一致するカテゴリラベルを出力することができる。

これらの特徴のすべてが、CNN を視覚システムのモデルに適した候補としている。しかし、これらの特徴のすべてが CNN に明示的に組み込まれているわけではない。CNN が視覚システムと同じような計算をしていることを検証するには、工学的ではない追加的な方法で視覚システムと一致させる必要がある。つまり、モデルに入れられた仮定から、データのさらなる特徴が落ちてくるはずだ。実際、神経レベルと行動レベルで多くのさらなる対応関係が発見されている。

3.1 神経レベルでの比較

最近、神経科学者の間で人工ニューラルネットワークへの関心が再燃している大きな原因のひとつは、CNNが腹側の流れに沿った視覚情報の表現を再現できることがわかったことである。特に、CNNと動物に同じ画像を見せた場合 (図 2)、人工ユニットの活動を使って本物のニューロンの活動を予測することができ、その精度はこれまでの方法を上回る。

これを示した初期の研究の 1 つ (Yamins et al, 2014) は 2014 年に発表されたもので、複雑な物体画像を見ている間のマカクの細胞外活動を記録しました。実際の V4 ニューロンや IT ニューロンの活動を、ネットワーク内の人工的なユニットの活動に回帰させたところ (そして、予測能力をホールドアウトしたテストセットで交差検証したところ)、物体認識課題で良い結果を出したネットワークは、神経活動の予測にも優れていることがわかった (この関係はビデオ分類を用いても見られた: Tacchetti, Isik, & Poggio, 2017)。さらに、ネットワークの最後の層のユニットの活動が IT の活動を最もよく予測し、最後の層が V4 を最もよく予測しました。このように、ネットワークの後半の層が腹側ストリームのより高い領域をよりよく予測するというモデルと脳の関係は、ヒトの fMRI (Güçlü & van Gerven, 2015)、MEG (Eeliger et al, 2018)、刺激として静止画像ではなく動画を用いた研究 (Eickenberg, Gramfort, Varoquaux, & Thirion, 2017) など、他のいくつかの研究でも見出されています。

異なる集団間の対応関係を確認するもう一つの方法として、表象類似性分析 (RSA; Kriegeskorte, Mur, & Bandettini, 2008) があります。この方法では、まず母集団ごとに、画像のペアごとにその母集団の反応がどれだけ似ていないかを表す行列を作成します。これは、その集団の表現特性の特徴として機能します。そして、2 つの異なる集団間の類似性は、それらの非類似度行列の相関として測定されます。この方法は、2014 年に (Khaligh-Razavi & Kriegeskorte, 2014)、ImageNet で学習した AlexNet ネットワークの後段が、人間の視覚系の複数の高次領域やサル IT と、これまで使用されていたモデルよりもよく一致することを実証しました。

非類似度行列は、行動出力を含むあらゆる種類の応答から作成できるため、RSA は異なる実験手法やモデルのデータを比較する手段として広く適用できる。また、回帰法が一度に 1 つのニューロンやボクセルに焦点を当てるのに対し、RSA は全集団の反応を取り入れて比較するための簡単な方法です。一方、回帰法では、データの適合に最も関連するモデルの特徴を選択的に重み付けすることができるため、「教師なし」の RSA 法よりも情報量が多い可能性があります。いずれの場合も、手法や解釈の詳細は慎重に検討する必要があります (Kornblith, Norouzi, Lee, & Hinton, 2019; Thompson, Bengio, Formisano, & Schönwiesner, 2018)。

CNN と生物学的データを比較した研究の多くは、V4 や IT などの後期視覚野を説明する能力を強調している。というのも、これらの領域の複雑な反応特性のために、一次視覚野に比べて CNN の適合が難しいことが知られていたからだ。しかし、最近の研究では、タスクトレーニングされた CNN の初期から中間層も、より伝統的な V1 モデルの能力を超えて V1 活動を予測できることが示されている (Cadena, Denfield, et al. 2019)。

神経活動を予測したり、全体的な表現を一致させたりするだけでなく、CNN はシステム神経科学で伝統的に用いられているより具体的な特徴を用いて神経データと比較することもできる。ネットワーク内のユニットと個々のニューロンの間には、応答の疎密やサイズのチューニングについて類似性が見出されているが、物体選択性や方向のチューニングについては違いが見出されている (Tripp, 2017)。他の研究では、形状やカテゴリへのチューニング (Zeman, Ritchie, Bracci, & Op de Beeck, 2019) や、ポーズ、位置、サイズの変化に伴って応答がどのように変化するかを調べています (Murty & Arun, 2017, 2018; Hong, Yamins, Majaj, & DiCarlo, 2016)。

一般的に、物体認識を行うために CNN をトレーニングしたときに現れる実際の神経活動との類似性は、このアーキテクチャとタスクが実際に視覚システムのアーキテクチャと目的にある程度類似していることを示唆している。

3.2 行動レベルでの比較

実世界の画像分類において、CNN がこれまでの視覚系モデルを凌駕している限り、CNN は人間の行動によくマッチしていると考えられる。しかし、標準的な ImageNet タスクでの総合的な精度は、CNN の動作を示す指標のひとつにすぎず、それはネットワークが明示的に最適化されたものである。

これらのネットワークが犯す誤りを見ることで、より深い比較が可能になります。正しい方法は一つしかありませんが、人間とモデルが分類を間違える方法は数多くあり、それぞれの仕組みを理解するのに役立ちます。例えば、混同行列は、あるカテゴリーの画像が別のカテゴリーに分類される頻度を表すもので、モデルと動物の行動を比較することができます。大規模な研究では、いくつかの異なるディープCNNアーキテクチャが、モデルが画像レベルまで予測できないにもかかわらず、人間とサルの物体分類に類似した一致を示すことが示された (Rajalingham et al., 2018)。

他の研究では、画像間の類似性を評価するよう参加者に明示的に求め、人間の判断とモデル表現を比較しています (King, Groen, Steel, Kravitz, & Baker, 2019; Jozwik, Kriegeskorte, Storrs, & Mur, 2017)。例えば、Rosenfeld, Solbach, and Tsotsos (2018) の研究では、大規模なデータセットが Web サイト “Totally Looks Like” から取得されました。多くの類似性研究では、CNN から良好なマッチングが得られますが、この研究では、CNN が再現するのに苦労した、より困難な類似性の要素が示されました。

類似性に加えて、CNN でテストされた他の心理学的概念には、典型性 (Lake, Zaremba, Fergus, & Gureckis, 2015), ゲシュタルト原理 (Kim, Reif, Wattenberg, & Bengio, 2019), アニマシー (Bracci, Ritchie, Kalfas, & Op de Beeck, 2019) などがある。Jacob, Pramod, Katti, and Arun (2019) では、視覚心理物理学の知見に触発された一連のテストを CNN に適用したところ、およそ半数のテストによれば、CNN は生物学的な視覚に類似していることがわかったという。

動物や CNN の行動を探索もう1つの方法は、画像品質を劣化させることで分類タスクをより困難にすることです。いくつかの研究では、標準的な画像にさまざまな種類のノイズ、隠蔽、またはぼかしを加えて、分類性能の低下を観察している (Geirhos, Temme, et al., 2018; Tang et al., 2018; Geirhos et al., 2017; Wichmann et al., 2017; Ghodrati, Farzmaadi, Rajaei, Ebrahimpour, & Khaligh-Razavi, 2014)。重要なのは、この性能低下は通常、CNN では人間よりも深刻であり、生物学的視覚には劣化を克服するメカニズムがあることを示唆している。このようなミスマッチは、今後の研究を進める上で重要なポイントとなります (「代替データセット」のセクションを参照)。カプセルネットワーク (Roy, Ghosh, Bhattacharya, & Pal, 2018) と呼ばれる特定の CNN アーキテクチャは、劣化に対してより頑健であることが示されたが、このアーキテクチャの「カプセル」を視覚系の一部とどのように関連付けるかは正確には明らかになっていない。

CNN の行動分析におけるもうひとつの新たな発見は、テクスチャーへの依存度である。CNN が人間の形状感度のモデルであるという議論がなされているが (Kubilius, Bracci, & Op de Beeck, 2016), 他の研究では、CNN が画像を分類する際に、テクスチャーに頼りすぎて、形状には十分に頼らないことが実証されている (Baker, Lu, Erlikhman, & Kellman, 2018; Geirhos, Rubisch, et al. 2018)。

興味深いことに、ある種の深層ネットワークがいくつかのタスクで人間を凌駕することは可能です (Kheradpisheh, Ghodrati, Ganjtabesh, & Masquelier, 2016)。このことは、コンピュータビジョンの目標と神経科学の目標との間の重要な緊張関係を浮き彫りにしています。前者では人間のパフォーマンスを超えることが望ましいが、後者ではモデルとデータのミスマッチとしてカウントされてしまう。

CNN の動作を研究する際に留意すべきことは、標準的なフィードフォワード CNN アーキテクチャは、さまざまな種類の再帰的処理が行われる前の、視覚処理のごく初期の段階を表していると考えられていることだ。そのため、CNN の挙動を動物の行動と比較する際には、高速な刺激提示と後方マスクングが推奨され、これらは再帰的処理の多くの段階を妨げることが知られているからだ (Tang et al. 2018)。

3.3 妥当性検証のための他の形態

神経と行動の比較は、CNN が視覚システムのモデルとして検証される主な方法であるが、他のアプローチもさらなる裏付けとなる。

ネットワークの異なる層のユニットを駆動する画像特徴を可視化する手法 (図4, Olah, Mordvintsev, & Schubert, 2017, Zeiler & Fergus, 2014) により、神経科学で見られるものと一致する好ましい視覚パターンが明らかになった。例えば、CNN の最初の層には Gabor のようなフィルタがあるが、後の層では部分的な物体の特徴に反応し、最終的には顔のような完全な特徴に反応する。このことは、CNN の処理ステップが腹側ストリームのそれと一致しているという考えを支持している。また、CNN は実際のニューロンに最適な刺激を作り出すためにも使われている。前述の CNN による神経活動予測の手順を出発点として、ニューロンの発火率を最大限に駆動することを意図した刺激を生成した (Bashivan, Kar, & DiCarlo, 2019)。結果として得られた刺激は、ネットワークが学習した画像とは無関係の不自然なものであったにもかかわらず、実際にニューロンを通常のレート以上に駆動させる効果があったという事実は、CNN が視覚処理ストリームに関する基本的な何かを捉えているという考えをさらに裏付けるものである。

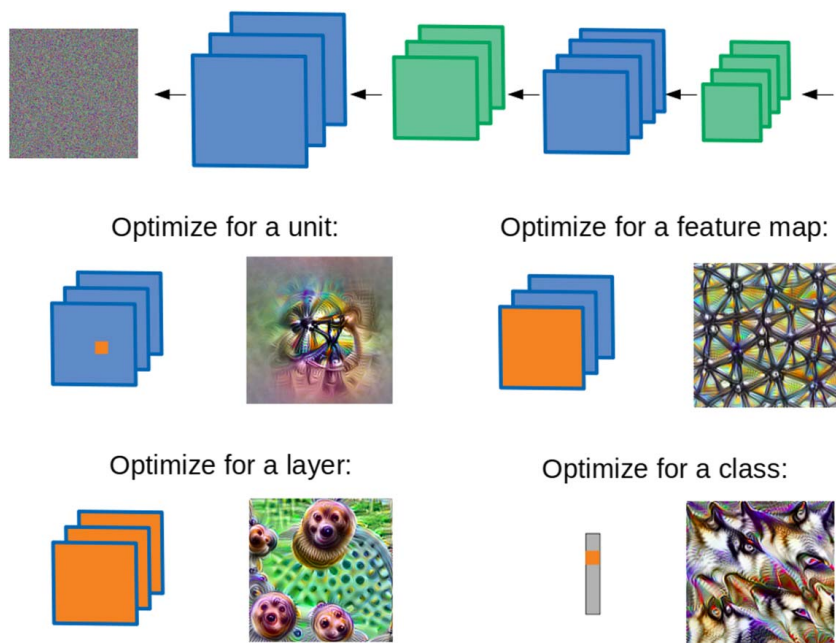


図 4. 異なるネットワークコンポーネントの好みを視覚化する。これらの画像を作成するための一般的な方法を上部に示す。具体的には、モデルのコンポーネントが最適化のために選択され、ネットワークを通じて送られてくる勾配計算によって、最初はノイズの多い画像の画素が選択されたコンポーネントを最も有効にするためにどのように変化すべきかが示される。コンポーネントには 個々のユニット、特徴地図全体、DeepDream アルゴリズムを用いた層、最終層の特定のクラスを表すユニットなどがある。Olah et al. (2017) より

これに関連して、CNN を使って神経活動を解読し、参加者に提示されている刺激を再現する研究も行われている (Shen, Horikawa, Majima, & Kamitani, 2019)。

4. モデル変更から学ぶこと

上述の研究では、主に標準的なフィードフォワード CNN アーキテクチャに焦点を当て、教師付き学習によって学習させ、物体認識を行ってきた。しかし、これらのモデルを完全に制御することができれば、データセット、アーキテクチャ、学習手順など、さまざまなバリエーションを検討することができる。これらの変更によってモデルのデータへの適合性がどのように良くなったり悪くなったりするかを観察することで、生物学的な視覚処理の特徴がどのようにして、またなぜ存在するのかについての洞察を得ることができる。

4.1 データセットの代替

ImageNet データセットは、MNIST や CIFAR-10 のような以前の小さくて単純なデータセットとは異なり、様々なタスクに適応できる基本的な視覚的特徴のセットを学習するのに非常に有用であることが証明されています。しかし、このデータセットには物体に焦点を当てた画像が含まれており、結果的に物体認識経路の研究に最も適しています。後頭葉領域のような場面処理領域を研究するために、いくつかの研究では、代わりに場面画像を使って学習を行っています。Bonner と Epstein (2018) の研究では、シーンを分類するように訓練されたネットワークを用いて、後頭葉場所領域の反応を予測することができました。さらに著者らは、ネットワークが捉えた特徴を、シーン内のナビゲーションアフォーダンスに関連付けることができた。Cichy, Khosla, Pantazis, and Oliva (2017) の研究では、シーンを学習したネットワークが、MEGを用いて収集したシーンサイズの表現を捉えることができた。このような研究は一般的に、異なる脳領域の表現が、異なる特殊なデータセットで訓練されたネットワークとどの程度一致するかに基づいて、異なる脳領域の進化的または発達の決定された役割を特定するのに役立ちます。異なる画像セットに対する fMRI の応答のオープンデータセットについては Chang et al. (2019) を参照してください。

また、CNN が人間の行動と一致しない点を修正することを明確に意図したデータセットも開発されている。例えば、異なる画像劣化でトレーニングすることで、ネットワークをその劣化に対してよりロバストにすることができる (Geirhos, Temme, et al.2018)。しかし、これは特別に訓練されたノイズモデルに対してのみ有効であり、新しいノイズの種類には一般化しない。また Geirhos, Rubisch, et al. (2018) の研究では、抽象的な形状を保持しつつ、低レベルのテクスチャ要素を変化させたデータセットが、CNN のテクスチャバイアスを減少させることが示された。動物が優れた視覚性能を持つのは、発達期に同様の多様なデータに触れることによるものなのか、それとも代わりに内蔵されたブライアによるものなのかは、まだ解明されていない。

4.2 アーキテクチャの代替

機械学習コミュニティや神経科学者の間では、さまざまな純粋なフィードフォワード・アーキテクチャが研究されています。データとの比較では、AlexNet がよく使われており、性能も良い。しかし、VGGモデルや ResNets などのやや深いアーキテクチャに負けることがあります (Schrimpf et al.2018; Wen, Shi, Chen, & Liu, 2018)。非常に深いネットワークでは、画像タスクでは優れた性能を発揮するかもしれないが、ネットワーク内の層と脳内の領域の関係が崩れ、データへの適合性が悪くなる可能性がある。しかし、非常に深いネットワークでの処理は、脳内の複数段階のリカレント処理と同等と考えられる場合もあります (Kar, Kubilius, Schmidt, Issa, & DiCarlo, 2019)。

霊長類の視覚では、最初の 2 つの処理段階 (網膜と外側帯状核) に、中心にある光のパターンと中心から外れた光のパターンに優先的に反応する細胞が含まれており、V1 でのみ方位のチューニングが支配的になります。しかし、ガボールフィルタは CNN の第 1 層で学習されることが多く、V1 のような表現になってしまう。視神経によって生物学的に制約されているように、初期層からの接続を修正したアーキテクチャを使用すると、初期段階では中心にある応答と中心から外れた応答があり、方向性のチューニングは後になってから行われるCNNができます (Lindsey, Ocko, Ganguli, & Deny, 2019)。このように、霊長類の視覚で見られる選択性のパターンは、解剖学的な制約の結果である可能性がある。

Kell, Yamins, Shook, Norman-Haignere, McDermott (2018) の研究では、一対の聴覚タスク (音声認識と音楽認識) を実行するために、さまざまなアーキテクチャが訓練されました。これを通じて著者らは、2 つのタスクが 3 つの層の処理を共有できることを発見し、その後、ネットワークが両方のタスクで良好なパフォーマンスを発揮するために専門的なストリームに分割する必要があることを明らかにしました。最近では、同様の手順を視覚タスクに適用し、顔の処理に特化した経路が視覚系で生じる理由の説明に用いられた (Dobs, Kell, Palmer, Cohen, & Kanwisher, 2019)。また、同様の問題は、単一のネットワークを 2 つのタスクを実行するように訓練し、訓練されたネットワークの中で出現するサブネットワークを探すことでもアプローチされました (Scholte, Losch, Ramakrishnan, de Haan, & Bohte, 2018)。このような研究は、背側と腹側のストリームの存在や、視覚アーキテクチャのその他の詳細を説明することができます。

生物学にヒントを得て、多くの研究では、局所的な再帰性とフィードバック的な再帰性の両方が有益な役割を果たすことが検討されています (図3)。局所的再帰性とは、1 つの視覚野の中での水平方向のつながりのことである。これらの接続を CNN に追加することで、これらの再帰的な接続がネットワークをより困難なタスクに適したものにするという研究結果がある (Hasani, Soleymani, & Aghajan, 2019; Montobbio, Bonnasse-Gahot, Citti, & Sarti, 2019; Spoerer, Kietzmann, & Kriegeskorte, 2019; Tang et al. 2018)。また、これらの接続は、特に困難な画像や応答の後の時点で、CNN 表現を神経データとよりよく一致させるのに役立ちます (Kar et al., 2019; Kubilius et al., 2019; Rajaei, Mohsenzadeh, Ebrahimpour, & Khaligh-Razavi, 2019; Shi, Wen, Zhang, Han, & Liu, 2018; McIntosh, Maheswaranathan, Nayeibi, Ganguli, & Baccus, 2016)。これらの研究は、これらの局所的な接続の計算上の役割を強く主張しています。

フィードバック接続は、前頭葉や頭頂葉の領域から視覚系の領域へ、あるいは高次の視覚領域から低次の領域へと戻る (Wyatte, Jilk, & O'Reilly, 2014)。水平方向の再帰性と同様に、生物の視覚ではよく見られることが知られています。前頭葉と頭頂葉からの接続は、目標に向けた選択的注意を実現すると考えられている。このようなフィードバックは、ネットワークモデルに追加され、手がかり付き検出タスクを実装している (Thorat, van Gerven, & Peelen, 2019; Wang, Zhang, Song, & Zhang, 2014)。高次の視覚野から低次の視覚野に戻るフィードバックは、ノイズ除去のようなより即時的で一般的な画像処理を実装すると考えられている。いくつかの研究では、局所的な再帰性に加えてこれらの接続を追加し、同様にパフォーマンスを助けることができることがわかっています (Kim, Linsley, Thakkar, & Serre, 2019; Spoerer, McClure, & Kriegeskorte, 2017)。異なるフィードバックとフィードフォワードのアーキテクチャを比較した研究では、フィードバックが細胞の有効な受容野サイズを増やすことで一部役立つことが示唆されています (Jarvers & Neumann, 2019)。

4.3 訓練手続きの代替

バックプロパゲーションを用いた教師付き学習は、CNN を学習する最も一般的な方法であるが、他の方法でも、視覚システムの優れたモデルをもたらす可能性がある。例えば、物体認識のパフォーマンスと神経反応を捉える能力の間に相関関係があることを最初に示した2014年の研究 (Yaminsら, 2014) では、バックプロパゲーションではなく、モジュール式の最適化手順を用いていた。

ネットワークが入力と出力を一致させるのではなく、入力データの関連する統計を捉えることを目的とする教師なし学習も、ニューラルネットワークの学習に使用することができる (Fleming & Storrs, 2019; 図3)。これらの方法は、高次元の視覚入力の特徴にある低次元の特徴のセットを特定するのに役立ち、その結果、動物が世界をよりよく理解できるようになり、おそらく有用な因果関係モデルを構築できるようになるかもしれない (Lake, Ullman, Tenenbaum, & Gershman, 2017)。さらに、教師ありの学習には大量のラベル付きデータポイントが必要なため、脳は教師なしの学習を利用しなければならないと考えられている。しかし、今のところ、教師なしの手法では、教師ありの手法と同様に神経表現を捉えるモデルが得られない。行動学的には、生成モデルは、人間の画像分類を捉える上で、教師ありモデルよりもperformmuch悪いことが示されました (Peterson, Abbott, & Griffiths, 2018)。また、予測符号化の概念に基づいて学習されたモデル (Lotter, Kreiman, & Cox, 2016) は、物体の動きを予測し、動きの錯覚を再現することができました (Watanabe, Kitaoka, Sakamoto, Yasugi, & Tanaka, 2018)。全体として、視覚モデルの教師なし訓練の限界と利点については、さらなる調査が必要である。

興味深いことに、最近の研究では、クラス条件付きの生成モデルが、分類を行うように訓練されたモデルよりも、論争的になっている刺激に対する人間の判断とよく一致することがわかった (Golan, Raju, & Kriegeskorte, 2019)。この生成モデルは依然としてクラスラベルに依存して学習しており、教師なしではないが、人間の知覚の側面を再現する能力は、視覚システムが単に画像を物体のカテゴリーに漏らすのではなく、視覚世界の分布を把握することを部分的に目的としているという考え方を支持するものである。

教師なし法と教師あり法の妥協点が「半教師あり」学習であり、最近ではより生物学的に現実的なネットワークを作るための手段として探求されている (Zhuang, Yan, Nayebi, & Yamins, 2019; Zhuang, Zhai, & Yamins, 2019)。

強化学習は、機械学習における3番目の主要な学習方法です。これらのシステムでは、人工的なエージェントは、報酬を含む環境からの情報に応じて行動の結果を生み出すことを学習しなければならない。このような人工システムのいくつかは、世界に関する視覚情報を処理するために、フロントエンドに畳み込みアーキテクチャを使用している (図3 ; Merel et al.2018 ; Paine et al.2018 ; Zhu et al. 2018)。これらのモデルの文脈で学習された表現を、他のメカニズムで学習された表現やデータと比較することは興味深いことです。

ネットワークのトレーニングの重要性を理解する簡単な方法は、同じアーキテクチャだがランダムな重みを持つネットワークと比較することです。Kim, Reif, et al.(2019)では、ネットワークが知覚的なクロージングを行う能力は、ランダムなネットワークではなく、自然な画像で訓練されたネットワークにのみ存在した。

上記の方法は、神経記録から得られない一連の画像などのトレーニングデータに依存しています。しかし、これらのアーキテクチャをトレーニングして、神経活動を直接再現することは可能です。そうすることで、ニューロンの発火に最も関与する入力画像の特徴を特定することができます (Cadena, Denfield, et al. 2019; Kindel, Christensen, & Zylberberg, 2019; Sinz et al., 2018)。理想的には、ネットワークが分類タスクで訓練されるときに行われるように、モデルの構成要素を当該回路の解剖学的特徴にも関連付けることができる (Günthner et al., 2019; Maheswaranathan et al., 2018)。

最後のハイブリッドオプションは、分類タスクでトレーニングしながら、ニューラルデータを使って中間表現を脳のようなものに制約することです (Fong, Scheirer, & Cox, 2018)。その結果、ニューラルデータとよりよくマッチし、タスクを実行できるネットワークができあがります。

5. CNN の理解の仕方

ネットワークの構造やトレーニングを変化させることは、ネットワークがどのように機能するかを探るための1つの方法です。さらに、訓練されたネットワークは、通常、脳に適用される技術やモデルでしか利用できない技術を使って直接調べることができる (Samek & Müller, 2019)。いずれにしても、CNN が視覚システムのモデルとして検証されていると考えれば、その仕組みを探ることで得られる知見は、生物学的な視覚にも応用できるかもしれない。

5.1 実験的方法

標準的な神経科学のツールボックスである、病変、記録、解剖学的トレーシング、刺激、サイレンシングなどのツールは、すべて人工ニューラルネットワークで容易に利用できる Barrett, Morcos, & Macke, 2019)、これらのネットワークの動きに関する疑問に答えるために使用することができます。

例えば CNN の個々のユニットを“アブレイティング”することで、分類精度に影響を与えることができますが、特定のユニットをアブレイティングすることの影響は、そのユニットの選択性の特性とは強い関係がありません (Morcos, Barrett, Rabinowitz, & Botvinick, 2018; Zhou, Sun, Bau, & Torralba, 2018)。

Bonner and Epstein (2018) の研究では、画像を操作したり、隠蔽したりして、どの特徴がシーンに対する CNN の反応を担っているかを、本物のニューロンの機能を探るのと同様に調べた。

他の研究 (Lindsey et al., 2019; Spoerer et al., 2019) では、訓練されたネットワークの接続特性を分析して、視覚皮質の解剖学的特徴を再現しているかどうかを確認しています。

視覚処理の各段階で行われていることを説明するための一つの概念的枠組みとして、"untangling" があります。ピクセルや網膜の表現の中で絡み合っていた高レベルの概念が、後の表現の中で分離しやすいクラスターを形成するために引き離される。この理論は、生物学的データを用いて開発されてきたが (DiCarlo & Cox, 2007), 最近では、これらのクラスター (または多様体) の形状を記述する技術が開発され、深層ニューラルネットワークにおけるアンタングリングプロセスの理解に用いられている (Cohen, Chung, Lee, & Sompolinsky, 2019; Chung, Lee, & Sompolinsky, 2018)。本研究では、これらの多様体の分類に関連する特徴と、それらが学習および処理段階を通じてどのように変化するかを明らかにしています。これは、データの中からどの応答の特徴を探すべきかを特定するのに役立ちます。

5.2 数学的分析

CNN を定義する方程式に完全にアクセスできれば、現在は脳に適用できない多くの数学的手法を実行できる。そのような一般的なツールのひとつが、グラデーションの計算である。グラデーションは、ネットワーク内の特定のコンポーネントが、遠く離れた他のコンポーネントにどのような影響を与えるかを示す。グラデーションは、ある層の重みをどのように変えれば出力でのエラーが減少するかを判断し、ネットワークを訓練するために使用されます。しかし、ネットワーク内のあるユニットの特徴を視覚化するためにも使用されます。その場合、勾配の計算は入力画像にまで及び、特定のユニットの活動を高めるために個々のピクセルをどのように変化させるべきかを決定することができる (Olah et al. 2017; Figure 4)。また、特徴の可視化の複数のバリエーションは、例えば、異なる層にどの不変性が存在するかを示すことで、ニューラルネットワークの機能を探るためにも使用されている (Nguyen, Yosinski, & Clune, 2019) (Cadena, Weis, Gatys, Bethge, & Ecker, 2018)。

勾配は、人工ニューロンが分類タスクで果たす役割を判断するためにも使用できます。LindsayとMiller (2018) の研究では、あるユニットが画像を特定のカテゴリーのものとして分類する際に果たす役割は、そのカテゴリーの画像にどれだけ強く反応するかとは密接に相関していないことがわかりました。上記のアブレーションの研究と同様に、これはチューニングと機能の解離を示しており、神経科学者たちは、脳のチューニングに基づく分析がどれほど有益であるかについて躊躇するはずです。

機械学習の研究者や数学者も、CNN を伝統的な数学的分析に適したものにすることを目指している。たとえば、情報理論 (Tishby & Zaslavsky, 2015) やウェーブレット散乱などの概念がこの目的のために使われてきた (Mallat, 2016)。数学的分析のもう一つの実りあるアプローチは、この近似がより多くの分析を可能にするとして、深層線形ニューラルネットワークを研究することである (Saxe, McClelland, & Ganguli, 2013)。

計算論的神経科学におけるいくつかの研究では、単純なりカレントニューラルネットワークを訓練してタスクを実行させ、その後、訓練したネットワークの特性を調べて、その仕組みを知る手がかりとしている (Barak, 2017; Foerster, Gilmer, Sohl-Dickstein, Chorowski, & Sussillo, 2017; Sussillo & Barak, 2013)。今後は、CNN の学習された表現や接続パターンを理解するための、より洗練された技術を開発する必要がある。これは、これらのネットワークがどのように機能するかについての洞察を提供すると同時に、どのような実験手法やデータ解析方法を追求することが実りあることなのかを示すことができる。

5.3 CNN は理解可能か？

研究者のなかには、CNN は複雑さと解釈可能性の間の避けられないトレードオフの存在であると考える人もいる。実世界のタスクを実行できるほど複雑なモデルを作るためには、システム神経科学の多くに固有の目標である、各ステージがどのように動作するかを簡単に説明するという欲求を犠牲にしなければならない。このような理由から、計算に関する単純な記述に頼らずにコンパクトにネットワークを記述する別の方法が提案されている (Lillicrap & Kording, 2019; Richards et al. 2019)。この視点では、具体的な計算を記述するのではなく、ネットワークのアーキテクチャ、最適化機能、学習アルゴリズムを記述することに焦点を当てています。なぜなら、具体的な計算や表現は、これら3つの要素から単純に生まれてくると考えられるからです。

確かに、個々のニューロンやニューロン群の役割を言語ベースで記述するという歴史的な目的 (たとえば“方向に同調する”とか“顔検出器”とか) は、CNN の本質的な計算を捉える方法としては非常に不完全であるように思える。しかし、これらのネットワークの機能を記述するには、もっとコンパクトな方法があり、そのようなシンプルな記述を見つけることで、さらなる理解を得ることができるとも考えられる。たとえば、ある種の重みのセットだけが、実世界の分類タスクでネットワークをうまく動作させることができる。今のところ、この「良い」重みについては、最適化によって得られるということしかわかっていません。しかし、ネットワークの説明をより凝縮したものにするために、特定できる本質的な特性はあるのでしょうか？宝くじのような学習方法では、一部の重みだけを使って高密度なネットワークと同じように機能するネットワークを作り出すことができます。最近の研究では、モデルの重みの95%もが残りの5%から推測できることが指摘されています (Frankle & Carbin, 2018; Denil, Shakibi, Dinh, Ranzato, & de Freitas, 2013)。このような知見は、高パフォーマンスのネットワークの計算について、より凝縮された (したがって、より理解しやすい可能性のある) 記述が可能であることを示唆している。

アーキテクチャの同様の分析は、接続性のどの広範な特徴が良好なパフォーマンスを生み出すのに十分であるかを判断するために行うことができる。例えば、ランダムな重みを用いた最近の研究では、特定の重みの値に対するアーキテクチャの役割を分離するのに役立つ (Gaier & Ha, 2019)。高パフォーマンスのネットワークの本質的な特徴をよりコンパクトに記述することは、機械学習と神経科学の両方にとっての目標である。

深層学習の紛れもない非コンパクトなコンポーネントの一つがデータセットです。これまで述べてきたように、生物学的な視覚の性能と神経表現を一致させるには、大規模な現実世界のデータセットが必要です。私たちは、視覚がどのように機能するかを説明するために、ImageNetデータセットを持ち歩くことを運命づけられているのでしょうか？それとも、十分な統計量のセットを定義することができるのでしょうか？自然なシーンの統計は、計算論的神経科学とコンピュータビジョンの両方において、歴史的に大きな役割を果たしてきました (Simoncelli & Olshausen, 2001)。この研究の多くは、

物体認識に関連する特徴を捉えるには不十分な低次の相関関係に焦点を当ててきましたが、高次の画像統計を探索するには機が熟しているように思われます。特に、生成モデリング (Salakhutdinov, 2015) の進歩により、自然画像の完全で複雑な特徴をモデルに凝縮する能力が指摘されています。

いずれにしても、CNN の解釈可能性に対する批判のほとんどすべてが、生物学的ネットワークにも同様に適用できる。したがって、CNN は、神経システムにおける理解のあり方を決めるためのよい実験場となる。

6. 基礎を超えて

2014 年以降、CNN のさまざまな機能を変化させることで、CNN の特性やデータ照合能力がどのように変化するのか、多くの疑問が爆発的な研究によって解決されてきました。これらの発見は、生物学的視覚の「方法」と「理由」の両方を理解する上での助けとなりました。さらに、視覚システムの「画像計算可能」なモデルを利用できるようになると、視覚の中核部分だけでなく、さまざまなモデリングの可能性が開けてくる。

6.1 認知的課題の探索

画像符号化と記憶性の関係は、Jaegle ら (2019) で探索された。著者らは CNN の後の層の応答の全体的な大きさが、画像が実験的にどのように記憶されるかに相関していることを示した。

Devereux, Clarke, and Tyler (2018) では、意味的特徴に基づくアトラクタネットワークが CNN アーキテクチャの末尾に追加された。この追加的な処理段階は、意味的な課題の間の脊髄周囲皮質活動を説明することができた。

視覚的注意は、困難な視覚的課題の成績を向上させることが知られている。注意の神経調節効果を CNN のユニットに適用すると、これらのネットワークでも成績が向上することが示され、初期の層よりも後の層に適用するとより効果的であった (Lindsay & Miller, 2018)。また、このような課題遂行型モデルの使用は、神経データのみ由来するものよりも、注意がどのように機能しうるかについてのより良い理論につながった (Thorat et al. 2019)。

最後に CNN は、細粒の知覚学習のいくつかの行動的・神経的效果を再現することもできた (Wenliang & Seitz, 2018)。

6.2 生物学的な詳細の追加

上述したアーキテクチャの変形の一部、特に局所再帰とフィードバック再帰の追加は、生物学的な着想を得たもので、困難な課題を実行するネットワークの能力を高めると想定している。また、機械学習コミュニティでは、脳から着想を得た他の細部を、困難な課題に有用であるとの期待から検討したことがある。これには、画像分類のための foveation と saccading (Mnih, Heess, & Graves, 2014) が含まれる。

もう一つの生物学的な詳細を追加する理由は、ネットワークを「悪化」させるかもしれないという信念からであろう。CNN が多くの生物学的な詳細を欠いているという事実は、必ずしも視覚系の悪いモデルになるわけではなく、単に抽象的なものになるだけである。しかし、抽象的なモデルと詳細なモデルを一緒にして CNN の高レベルの計算が脳にある機械を使ってどのように実行されるかを示すことが究極の目的であるはずである。この目的のために、スパイク CNN (Cao, Chen, & Khosla, 2015) や確率的ノイズを加えた CNN (McIntosh et al., 2016) の研究が行われている。これらの試みは、これらの生物学的な詳細のうち、計算に有用な側面を特定することもできる。

視覚野の回路をモデル化してきた長い歴史は、どのような詳細をどのように取り込むべきかについて、より多くのアイデアを提供することができる。例えば V1 の解剖学と機能の既存の回路モデル (Rubin, Van Hooser, & Miller, 2015) は、CNN のアーキテクチャに入れられ、視覚的注意の効果を再現するのに使われた (Lindsay, Rubin, & Miller, 2019)。生物学的な詳細を加えるより極端なアプローチは Tschopp, Reiser, and Turaga の研究 (2018) に見られ、ハエの視覚系のコネクトームがモデルのアーキテクチャを定義し、そのモデルは視覚課題を実行するために訓練された。これを超えて、視覚系の

画像計算可能なモデルにアクセスすることは、コアビジョン以上のものを探索できる多くのモデリングの機会への扉を開く (図5)。

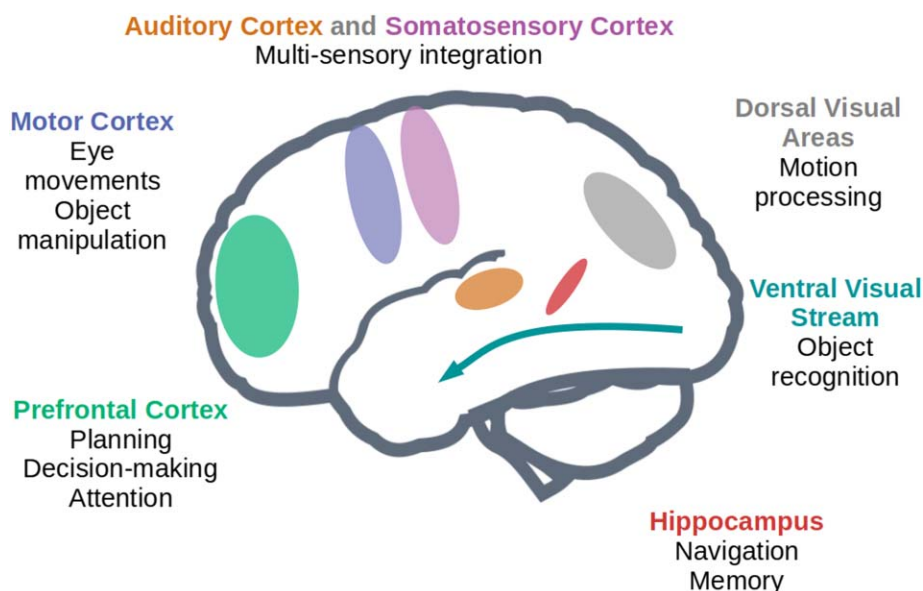


図5. 脳による視覚情報のさまざまな利用法の一例。視覚処理モデルとしての CNN の多くは 腹側視覚経路に焦点を当てている。しかし 視覚系は脳の他の多くの部分と相互作用し、さまざまな目標を達成している。将来のモデルでは CNN をこの大きな図式に統合する

作業が必要である。

7. 限界と将来の方向性

どのモデルにも言えることだが CNN の現在の限界と欠点は、これらのモデルをより生物学と一致させるための将来の研究の方向性を指し示すはずである。CNN の基本構造は重みの共有を仮定している。つまり、特徴地図は下層の各位置に全く同じフィルタ重みが適用された結果である。方向などの視覚的特徴に対する選択性は網膜地図の至る所に現れるが、これはいかなる種類の明示的な重み共有の結果でもないことは明らかである。遺伝的プログラミングによって空間全体で同じ特徴が検出されるようになったか、あるいはこの性質が暴露されることによって学習されたかのどちらかである。「翻訳耐性」に関する研究は、後者が真実である可能性を示している (Dill & Fahle, 1998)。重みの共有は CNN の学習を容易にするが、理想的にはより生物学的に妥当な方法でフィルタを当てはめれば、同じ結果が得られるだろう。

さらに、ほとんどの CNN ではデールの法則は尊重されない。つまり、同じニューロンに対して抑制性 (負) と興奮性 (正) の両方の重みを与えることができる。視覚系では領域間の接続は興奮性セルからだけもたらされる傾向がある。生物学と矛盾しないように、負のフィードフォワードの重みは、局所的な抑制性ニューロンに作用する興奮性フィードフォワード接続と解釈することができる。しかし、興奮性フィードフォワード接続と局所的な抑制性再帰性の必要性との間のこの関係は、これらのネットワークに生物学的詳細を追加することの複雑さを指摘している。これらの生物学的詳細の中には、他の生物学的詳細と照らし合わせて初めてうまく機能したり、意味をなすものがあるため、それらを一緒に追加する必要があるのだ。

また、これらのネットワークの学習方法にも問題があるとの指摘がある。バックプロパゲーションアルゴリズムは、視覚系が実際に学習する方法の近似として、生物学的に十分な信憑性があるとは考えられていない。しかし、計算論的神経科学におけるモデルフィッティングのほとんどの手法は、生物学的学習を模倣することを意図しておらず、バックプロパゲーションは単なるパラメータフィッティング手法の 1 つと考えることができる。とはいえ、いくつかの研究者は、脳がバックプロパゲーションのようなことを行うことができる手段を研究している (Whittington & Bogacz, 2019; Bartunov et al., 2018; Sacramento, Costa, Bengio, & Senn, 2018; Roelfsema, van Ooyen, & Watanabe, 2010)。より生物学的に妥当な技術を用いて訓練されたモデルを標準的な教師あり学習と比較することで (上述の教師なし学習や強化学習アプローチと同様に) 表現の決定における学習の役割に関する洞察を提供することができるだろう。

CNN を生物学的視覚と比較する研究の大半は、ヒトまたはヒト以外の霊長類からのデータを用いている。神経科学の研究で最もよく使われる動物群の 1 つであるげっ歯類と CNN を比較する試みがなされたところ、結果は霊長類の場合ほど強力ではない (Cadena, Sinz, et al., 2019; de Vries et al., 2019; Matteucci, Marotti, Riggi, Rosselli, & Zoccolan, 2019)。何が CNN を齧歯類の視覚の良いモデルに変えることができるのかを理解することは、霊長類と齧歯類の視覚の違いを理解する上で大きな助けとなり、ここで述べた探索戦略に齧歯類の視覚を開放することになるであろう。また CNN は分類課題におけるハトの行動パターンと比較されている (Love, Guest, Slomka, Navarro, & Wasserman, 2017)。

霊長類の視覚においてさえ、単純な物体やシーンの分類課題は、視覚系が自然にできることや使われることのごく一部に過ぎない。視覚処理の多様性と他の脳領域との関連性を把握するためには、ナビゲーション、物体操作、視覚的推論など、より倫理的に関連性の高い具体的な課題が必要であろう。このアイデアの初期バージョンはすでに検討されている (Cichy, Kriegeskorte, Jozwik, van den Bosch, & Charest, 2019; Dwivedi & Roig, 2018)。昆虫の視覚の研究は、歴史的にこのようなより全体的なアプローチをとっており、有用なインスピレーションになるかもしれない (Turner, Giraldo, Schwartz, & Rieke, 2019)。

7.2 結論

CNN の物語は、一次視覚野の個々のニューロンの同調特性に関する研究から始まった。しかし、視覚システムの研究に CNN を用いることの影響の 1 つは、単一ニューロンの解釈可能な反応に焦点を当てることから、視覚情報がどのように表現され、視覚課題を実行するために変換されるかについての集団レベルの記述へと、この分野を押しやることであった。実際に何かを「行う」モデルへのシフトは、視覚系の研究にまつわる疑問の再構築を余儀なくされた。神経科学者は、この新しいスタイルの説明と、それに伴うさまざまな期待に適応している (Hasson, Nastase, & Goldstein, 2019)。

重要なことは、これらのモデルによって、視覚の研究における既存の目標のいくつかに到達することも可能になったということです。例えば 2007 年 (Pinto, Cox, & DiCarlo, 2008)、物体認識の研究に関する展望記事では“物体認識に対する脳の解法の理解の進歩には、最終的に我々自身の視覚能力を模倣することを目的とした人工認識系の構築が必要であり、多くの場合、生物学のインスピレーションを得た”、“稼働中の認識系の実態化は物体認識の理解の上で特に有効な成功指標となる”と主張している。このように、視覚系のモデルとしての CNN は成功なのである。

もちろん CNN 単体では生物学的視覚について何も学べず、むしろ反復によってのみ学べる。CNN の実験から得られた洞察は、研究室での今後の実験を形成し、ひいては次世代のモデルに反映されるはずである。

文獻

- Baker, N., Lu, H., Erlikhman, G., & Kellman, P. J. (2018). Deep convolutional networks do not classify based on global object shape. *PLoS Computational Biology*, 14, e1006613. <https://doi.org/10.1371/journal.pcbi.1006613>. PMID: 30532273. PMCID: PMC6306249.
- Barak, O. (2017). Recurrent neural networks as versatile tools of neuroscience research. *Current Opinion in Neurobiology*, 46, 1–6. <https://doi.org/10.1016/j.conb.2017.06.003>. PMID: 28668365.
- Barrett, D. G. T., Morcos, A. S., & Macke, J. H. (2019). Analyzing biological and artificial neural networks: Challenges with opportunities for synergy? *Current Opinion in Neurobiology*, 55, 55–64. <https://doi.org/10.1016/j.conb.2019.01.007>. PMID: 30785004.
- Bartunov, S., Santoro, A., Richards, B. A., Marris, L., Hinton, G. E., & Lillicrap, T. P. (2018). Assessing the scalability of biologically-motivated deep learning algorithms and architectures. In *Advances in neural information processing systems* (pp. 9390–9400). Montreal, Canada.

- Bashivan, P., Kar, K., & DiCarlo, J. J. (2019). Neural population control via deep image synthesis. *Science*, 364, eaav9436. <https://doi.org/10.1126/science.aav9436>. PMID: 31048462.
- Bonner, M. F., & Epstein, R. A. (2018). Computational mechanisms underlying cortical responses to the affordance properties of visual scenes. *PLoS Computational Biology*, 14, e1006111. <https://doi.org/10.1371/journal.pcbi.1006111>. PMID: 29684011. PMCID: PMC5933806
- Bracci, S., Ritchie, J. B., Kalfas, I., & Op de Beeck, H. P. (2019). The ventral visual pathway represents animal appearance over animacy, unlike human behavior and deep neural networks. *Journal of Neuroscience*, 39, 6513–6525. <https://doi.org/10.1523/JNEUROSCI.1714-18.2019>. PMID: 31196934. PMCID: PMC6697402.
- Cadena, S. A., Denfield, G. H., Walker, E. Y., Gatys, L. A., Tolia, A. S., Bethge, M., et al. (2019). Deep convolutional models improve predictions of macaque V1 responses to natural images. *PLoS Computational Biology*, 15, e1006897. <https://doi.org/10.1371/journal.pcbi.1006897>. PMID: 31013278. PMCID: PMC6499433.
- Cadena, S. A., Sinz, F. H., Muhammad, T., Froudarakis, E., Cobos, E., Walker, E. Y., et al. (2019). How well do deep neural networks trained on object recognition characterize the mouse visual system? In *NeurIPS Workshop on Real Neurons and Hidden Units*. Vancouver, Canada.
- Cadena, S. A., Weis, M. A., Gatys, L. A., Bethge, M., & Ecker, A. S. (2018). Diverse feature visualizations reveal invariances in early layers of deep neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 217–232). Munich, Germany. https://doi.org/10.1007/978-3-030-01258-8_14.
- Cadieu, C., Kouh, M., Pasupathy, A., Connor, C. E., Riesenhuber, M., & Poggio, T. (2007). A model of V4 shape selectivity and invariance. *Journal of Neurophysiology*, 98, 1733–1750. <https://doi.org/10.1152/jn.01265.2006>. PMID: 17596412.
- Cao, Y., Chen, Y., & Khosla, D. (2015). Spiking deep convolutional neural networks for energy-efficient object recognition. *International Journal of Computer Vision*, 113, 54–66. <https://doi.org/10.1007/s11263-014-0788-3>.
- Chang, N., Pyles, J. A., Marcus, A., Gupta, A., Tarr, M. J., & Aminoff, E. M. (2019). BOLD5000, a public fMRI dataset while viewing 5000 visual images. *Scientific Data*, 6, 49. <https://doi.org/10.1038/s41597-019-0052-3>. PMID: 31061383. PMCID: PMC6502931
- Chung, S., Lee, D. D., & Sompolinsky, H. (2018). Classification and geometry of general perceptual manifolds. *Physical Review X*, 8, 031003. <https://doi.org/10.1103/PhysRevX.8.031003>.
- Cichy, R. M., Khosla, A., Pantazis, D., & Oliva, A. (2017). Dynamics of scene representations in the human brain revealed by magnetoencephalography and deep neural networks. *Neuroimage*, 153, 346–358. <https://doi.org/10.1016/j.neuroimage.2016.03.063>. PMID: 27039703. PMCID: PMC5542416.
- Cichy, R. M., Kriegeskorte, N., Jozwik, K. M., van den Bosch, J. J. F., & Charest, I. (2019). The spatiotemporal neural dynamics underlying perceived similarity for real-world objects. *Neuroimage*, 194, 12–24. <https://doi.org/10.1016/j.neuroimage.2019.03.031>. PMID: 30894333. PMCID: PMC6547050.
- Cohen, U., Chung, S., Lee, D. D., & Sompolinsky, H. (2019). Separability and geometry of object manifolds in deep neural networks. *bioRxiv*, 644658. <https://doi.org/10.1101/644658>.
- Denil, M., Shakibi, B., Dinh, L., Ranzato, M. A., & de Freitas, N. (2013). Predicting parameters in deep learning. In *Advances in neural information processing systems* (pp. 2148–2156). Lake Tahoe, NV.
- Devereux, B. J., Clarke, A., & Tyler, L. K. (2018). Integrated deep visual and semantic attractor neural networks predict fMRI pattern-information along the ventral object processing pathway. *Scientific Reports*, 8, 10636. <https://doi.org/10.1038/s41598-018-28865-1>. PMID: 30006530. PMCID: PMC6045572.
- de Vries, S. E. J., Lecoq, J. A., Buice, M. A., Groblewski, P. A., Ocker, G. K., Oliver, M., et al. (2019). A large-scale, standardized physiological survey reveals functional organization of the mouse visual cortex. *Nature Neuroscience*, 23, 138–151. <https://doi.org/10.1038/s41593-019-0550-9>. PMID: 31844315. PMCID: PMC6948932.
- DiCarlo, J. J., & Cox, D. D. (2007). Untangling invariant object recognition. *Trends in Cognitive Sciences*, 11, 333–341. <https://doi.org/10.1016/j.tics.2007.06.010>. PMID: 17631409. * Dill, M., & Fahle, M. (1998). Limited translation invariance of human visual pattern recognition. *Perception & Psychophysics*, 60, 65–81. <https://doi.org/10.3758/BF03211918>. PMID: 9503912
- Dobs, K., Kell, A. J. E., Palmer, I., Cohen, M., & Kanwisher, N. (2019). Why are face and object processing segregated in the human brain? Testing computational hypotheses with deep convolutional neural networks. Oral presentation at Cognitive Computational Neuroscience Conference, Berlin, Germany. <https://doi.org/10.32470/CCN.2019.1405-0>.
- Dwivedi, K., & Roig, G. (2018). Task-specific vision models explain task-specific areas of visual cortex. *bioRxiv*, 402735. <https://doi.org/10.1101/402735>.
- Eickenberg, M., Gramfort, A., Varoquaux, G., & Thirion, B. (2017). Seeing it all: Convolutional network layers map the function of the human visual system. *Neuroimage*, 152, 184–194. <https://doi.org/10.1016/j.neuroimage.2016.10.001>. PMID: 27777172.
- Fleming, R. W., & Storrs, K. R. (2019). Learning to see stuff. *Current Opinion in Behavioral Sciences*, 30, 100–108. <https://doi.org/10.1016/j.cobeha.2019.07.004>. PMID: 31886321. PMCID: PMC6919301.

- Foerster, J. N., Gilmer, J., Sohl-Dickstein, J., Chorowski, J., & Sussillo, D. (2017). Input switched affine networks: An RNN architecture designed for interpretability. In Proceedings of the 34th International Conference on Machine Learning (Vol. 70, pp. 1136–1145). Sydney, Australia.
- Fong, R. C., Scheirer, W. J., & Cox, D. D. (2018). Using human brain activity to guide machine learning. *Scientific Reports*, 8, 5397. <https://doi.org/10.1038/s41598-018-23618-6>. PMID: 29599461. PMCID: PMC5876362.
- Frankle, J., & Carbin, M. (2018). The lottery ticket hypothesis: Finding sparse, trainable neural networks. arXiv preprint arXiv:1803.03635.
- Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36, 193–202. <https://doi.org/10.1007/BF00344251>. PMID: 7370364.
- Gaier, A., & Ha, D. (2019). Weight agnostic neural networks. arXiv preprint arXiv:1906.04358.
- Geirhos, R., Janssen, D. H., Schütt, H. H., Rauber, J., Bethge, M., & Wichmann, F. A. (2017). Comparing deep neural networks against humans: Object recognition when the signal gets weaker. arXiv preprint arXiv:1706.06969.
- Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2018). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. arXiv preprint arXiv:1811.12231.
- Geirhos, R., Temme, C. R., Rauber, J., Schütt, H. H., Bethge, M., & Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. In *Advances in neural information processing systems* (pp. 7538–7550).
- Ghodrati, M., Farzmaadi, A., Rajaei, K., Ebrahimpour, R., & Khaligh-Razavi, S.-M. (2014). Feedforward object-vision models only tolerate small image variations compared to human. *Frontiers in Computational Neuroscience*, 8, 74. <https://doi.org/10.3389/fncom.2014.00074>. PMID: 25100986. PMCID: PMC4103258.
- Golan, T., Raju, P. C., & Kriegeskorte, N. (2019). Controversial stimuli: Pitting neural networks against each other as models of human recognition. arXiv preprint arXiv:1911.09288.
- GüçlÜ, U., & van Gerven, M. A. J. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35, 10005–10014. <https://doi.org/10.1523/JNEUROSCI.5023-14.2015>. PMID: 26157000. PMCID: PMC6605414.
- Günthner, M. F., Cadena, S. A., Denfield, G. H., Walker, E. Y., Tolia, A. S., Bethge, M., et al. (2019). Learning divisive normalization in primary visual cortex. bioRxiv, 767285. <https://doi.org/10.32470/CCN.2019.1211-0>.
- Hasani, H., Soleymani, M., & Aghajan, H. (2019). Surround modulation: A bio-inspired connectivity structure for convolutional neural networks. In *Advances in neural information processing systems* (pp. 15877–15888). Vancouver, Canada.
- Hasson, U., Nastase, S. A., & Goldstein, A. (2019). Robust-fit to nature: An evolutionary perspective on biological (and artificial) neural networks. bioRxiv, 764258. <https://doi.org/10.1101/764258>.
- Hong, H., Yamins, D. L. K., Majaj, N. J., & DiCarlo, J. J. (2016). Explicit information for category-orthogonal object properties increases along the ventral stream. *Nature Neuroscience*, 19, 613–622. <https://doi.org/10.1038/nn.4247>. PMID: 26900926.
- Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, 160, 106–154. <https://doi.org/10.1113/jphysiol.1962.sp006837>. PMID: 14449617. PMCID: PMC1359523.
- Jacob, G., Pramod, R. T., Katti, H., & Arun, S. P. (2019). Do deep neural networks see the way we do? bioRxiv, 860759. <https://doi.org/10.1101/860759>.
- Jaegle, A., Mehrpour, V., Mohsenzadeh, Y., Meyer, T., Oliva, A., & Rust, N. (2019). Population response magnitude variation in inferotemporal cortex predicts image memorability. *eLife*, 8, e47596. <https://doi.org/10.7554/eLife.47596.015>. <https://doi.org/10.7554/eLife.47596>. PMID: 31464687. PMCID: PMC6715346.
- Jarvers, C., & Neumann, H. (2019). Incorporating feedback in convolutional neural networks. In *Proceedings of the Cognitive Computational Neuroscience Conference* (pp. 395–398). Berlin, Germany. <https://doi.org/10.32470/CCN.2019.1191-0>.
- Jozwik, K. M., Kriegeskorte, N., Storrs, K. R., & Mur, M. (2017). Deep convolutional neural networks outperform featurebased but not categorical models in explaining object similarity judgments. *Frontiers in Psychology*, 8, 1726. <https://doi.org/10.3389/fpsyg.2017.01726>. PMID: 29062291. PMCID: PMC5640771.
- Kar, K., Kubilius, J., Schmidt, K., Issa, E. B., & DiCarlo, J. J. (2019). Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature Neuroscience*, 22, 974–983. <https://doi.org/10.1038/s41593-019-0392-5>. PMID: 31036945.
- Kay, K. N. (2018). Principles for models of neural information processing. *Neuroimage*, 180, 101–109. <https://doi.org/10.1016/j.neuroimage.2017.08.016>. PMID: 28793238.
- Kell, A. J. E., Yamins, D. L. K., Shook, E. N., Norman-Haignere, S. V., & McDermott, J. H. (2018). A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, 98, 630–644. <https://doi.org/10.1016/j.neuron.2018.03.044>. PMID: 29681533.

- Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Computational Biology*, 10, e1003915. <https://doi.org/10.1371/journal.pcbi.1003915>. PMID: 25375136. PMCID: PMC4222664.
- Kheradpisheh, S. R., Ghodrati, M., Ganjtabesh, M., & Masquelier, T. (2016). Deep networks can resemble human feed-forward vision in invariant object recognition. *Scientific Reports*, 6, 32672. <https://doi.org/10.1038/srep32672>. PMID: 27601096. PMCID: PMC5013454.
- Kietzmann, T. C., McClure, P., & Kriegeskorte, N. (2019). Deep neural networks in computational neuroscience. In *Oxford research encyclopedia of neuroscience*. Oxford: Oxford University Press. <https://doi.org/10.1093/acrefore/9780190264086>. 013.46.
- Kim, J., Linsley, D., Thakkar, K., & Serre, T. (2019). Disentangling neural mechanisms for perceptual grouping. *arXiv preprint arXiv:1906.01558*. <https://doi.org/10.32470/CCN.2019.1130-0>.
- Kim, B., Reif, E., Wattenberg, M., & Bengio, S. (2019). Do neural networks show Gestalt phenomena? An exploration of the law of closure. *arXiv preprint arXiv:1903.01069*.
- Kindel, W. F., Christensen, E. D., & Zylberberg, J. (2019). Using deep learning to probe the neural code for images in primary visual cortex. *Journal of Vision*, 19, 29. <https://doi.org/10.1167/19.4.29>. PMID: 31026016. PMCID: PMC6485988.
- King, M. L., Groen, I. I. A., Steel, A., Kravitz, D. J., & Baker, C. I. (2019). Similarity judgments and cortical visual responses reflect different properties of object and scene categories in naturalistic images. *Neuroimage*, 197, 368–382. <https://doi.org/10.1016/j.neuroimage.2019.04.079>. PMID: 31054350. PMCID: PMC6591094.
- Kornblith, S., Norouzi, M., Lee, H., & Hinton, G. (2019). Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning* (pp. 3519–3529). Long Beach, CA.
- Kriegeskorte, N. (2015). Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, 1, 417–446. <https://doi.org/10.1146/annurev-vision-082114-035447>. PMID: 28532370.
- Kriegeskorte, N., & Golan, T. (2019). Neural network models and deep learning. *Current Biology*, 29, R231–R236. <https://doi.org/10.1016/j.cub.2019.02.034>. PMID: 30939301.
- Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis—Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2, 4. <https://doi.org/10.3389/neuro.06.004.2008>. PMID: 19104670. PMCID: PMC2605405.
- Kubilius, J., Bracci, S., & Op de Beeck, H. P. (2016). Deep neural networks as a computational model for human shape sensitivity. *PLoS Computational Biology*, 12, e1004896. <https://doi.org/10.1371/journal.pcbi.1004896>. PMID: 27124699. PMCID: PMC4849740.
- Kubilius, J., Schrimpf, M., Kar, K., Hong, H., Majaj, N. J., Rajalingham, R., et al. (2019). Brain-like object recognition with high-performing shallow recurrent ANNs. *arXiv preprint arXiv:1909.06161*.
- Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40, e253. <https://doi.org/10.1017/S0140525X16001837>. PMID: 27881212.
- Lake, B. M., Zaremba, W., Fergus, R., & Gureckis, T. M. (2015). Deep neural networks predict category typicality ratings for images. In R. Dale, C. Jennings, P. Maglio, T. Matlock, D. Noelle, A. Warlaumont, & J. Yoshimi (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.
- LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., et al. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1, 541–551. <https://doi.org/10.1162/neco.1989.1.4.541>.
- Lillicrap, T. P., & Kording, K. P. (2019). What does it mean to understand a neural network? *arXiv preprint arXiv:1907.06374*.
- Lindsay, G. W., & Miller, K. D. (2018). How biological attention mechanisms improve task performance in a large-scale visual system model. *eLife*, 7, e38105. <https://doi.org/10.7554/eLife.38105>. <https://doi.org/10.7554/eLife.38105.030>.
- Lindsay, G. W., Rubin, D. B., & Miller, K. D. (2019). A simple circuit model of visual cortex explains neural and behavioral aspects of attention. *bioRxiv*, 875534. <https://doi.org/10.1101/2019.12.13.875534>.
- Lindsey, J., Ocko, S. A., Ganguli, S., & Deny, S. (2019). A unified theory of early visual representations from retina to cortex through anatomically constrained deep CNNs. *bioRxiv*, 511535. <https://doi.org/10.1101/511535>.
- Lotter, W., Kreiman, G., & Cox, D. (2016). Deep predictive coding networks for video prediction and unsupervised learning. *arXiv preprint arXiv:1605.08104*.
- Love, B. C., Guest, O., Slomka, P., Navarro, V. M., & Wasserman, E. (2017). Deep networks as models of human and animal categorization. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society* (pp. 1457–1458). London, UK.
- Maheswaranathan, N., McIntosh, L. T., Kastner, D. B., Melander, J., Brezovec, L., Nayebi, A., et al. (2018). Deep learning models reveal internal structure and diverse computations in the retina under natural scenes. *bioRxiv*, 340943.
- Mallat, S. (2016). Understanding deep convolutional networks. *Philosophical Transactions of the Royal Society of London, Series A: Mathematical, Physical and Engineering Sciences*, 374, 20150203. <https://doi.org/10.1098/rsta.2015.0203>. PMID: 26953183. PMCID: PMC4792410.

- Matteucci, G., Marotti, R. B., Riggi, M., Rosselli, F. B., & Zoccolan, D. (2019). Nonlinear processing of shape information in rat lateral extrastriate cortex. *Journal of Neuroscience*, 39, 1649–1670. <https://doi.org/10.1523/JNEUROSCI.1938-18.2018>. PMID: 30617210. PMCID: PMC6391565.
- McIntosh, L. T., Maheswaranathan, N., Nayeibi, A., Ganguli, S., & Baccus, S. A. (2016). Deep learning models of the retinal response to natural scenes. In *Advances in neural information processing systems* (pp. 1361–1369). Barcelona, Spain.
- Merel, J., Ahuja, A., Pham, V., Tunyasuvunakool, S., Liu, S., Tirumala, D., et al. (2018). Hierarchical visuomotor control of humanoids. *arXiv preprint arXiv:1811.09656*.
- Mnih, V., Heess, N., & Graves, A. (2014). Recurrent models of visual attention. In *Advances in neural information processing systems* (pp. 2204–2212). Montreal, Canada.
- Montobbio, N., Bonnasse-Gahot, L., Citti, G., & Sarti, A. (2019). KerCNNs: Biologically inspired lateral connections for classification of corrupted images. *arXiv preprint arXiv:1910.08336*.
- Morcos, A. S., Barrett, D. G. T., Rabinowitz, N. C., & Botvinick, M. (2018). On the importance of single directions for generalization. *arXiv preprint arXiv:1803.06959*.
- Murty, N. A. R., & Arun, S. P. (2017). A balanced comparison of object invariances in monkey IT neurons. *eNeuro*, 4, ENEURO.0333-16.2017. <https://doi.org/10.1523/ENEURO.0333-16.2017>. PMID: 28413827. PMCID: PMC5390242.
- Murty, N. A. R., & Arun, S. P. (2018). Multiplicative mixing of object identity and image attributes in single inferior temporal neurons. *Proceedings of the National Academy of Sciences, U.S.A.*, 115, E3276–E3285. <https://doi.org/10.1073/pnas.1714287115>. PMID: 29559530. PMCID: PMC5889630.
- Nguyen, A., Yosinski, J., & Clune, J. (2019). Understanding neural networks via feature visualization: A survey. *arXiv preprint arXiv:1904.08939*. https://doi.org/10.1007/978-3-030-28954-6_4.
- Olah, C., Mordvintsev, A., & Schubert, L. (2017). Feature visualization. *Distill*, 2, e7. <https://doi.org/10.23915/distill.00007>.
- Paine, T. L., Colmenarejo, S. G., Wang, Z., Reed, S., Aytaç, Y., Pfaff, T., et al. (2018). One-shot high-fidelity imitation: Training large-scale deep nets with RL. *arXiv preprint arXiv:1810.05017*.
- Peterson, J. C., Abbott, J. T., & Griffiths, T. L. (2018). Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cognitive Science*, 42, 2648–2669. <https://doi.org/10.1111/cogs.12670>. PMID: 30178468.
- Pinto, N., Cox, D. D., & DiCarlo, J. J. (2008). Why is real-world visual object recognition hard? *PLoS Computational Biology*, 4, e27. <https://doi.org/10.1371/journal.pcbi.0040027>. PMID: 18225950. PMCID: PMC2211529.
- Rajaei, K., Mohsenzadeh, Y., Ebrahimpour, R., & Khaligh-Razavi, S.-M. (2019). Beyond core object recognition: Recurrent processes account for object recognition under occlusion. *PLoS Computational Biology*, 15, e1007001. <https://doi.org/10.1371/journal.pcbi.1007001>. PMID: 31091234. PMCID: PMC6538196.
- Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., & DiCarlo, J. J. (2018). Large-scale, high-resolution comparisons of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *Journal of Neuroscience*, 38, 7255–7269. <https://doi.org/10.1523/JNEUROSCI.0388-18.2018>. PMID: 30006365. PMCID: PMC6096043.
- Rawat, W., & Wang, Z. (2017). Deep convolutional neural networks for image classification: A comprehensive review. *Neural Computation*, 29, 2352–2449. https://doi.org/10.1162/neco_a_00990. PMID: 28599112.
- Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., et al. (2019). A deep learning framework for neuroscience. *Nature Neuroscience*, 22, 1761–1770. <https://doi.org/10.1038/s41593-019-0520-2>. PMID: 31659335. PMCID: PMC7115933.
- Riesenhuber, M., & Poggio, T. (2000). Computational models of object recognition in cortex: A review (CBCL Paper 190/AI Memo 1695). Cambridge, MA: MIT Press. <https://doi.org/10.21236/ADA458109>.
- Roelfsema, P. R., van Ooyen, A., & Watanabe, T. (2010). Perceptual learning rules based on reinforcers and attention. *Trends in Cognitive Sciences*, 14, 64–71. <https://doi.org/10.1016/j.tics.2009.11.005>. PMID: 20060771. PMCID: PMC2835467.
- Rosenfeld, A., Solbach, M. D., & Tsotsos, J. K. (2018). Totally looks like—How humans compare, compared to machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 1961–1964). Salt Lake City, UT. <https://doi.org/10.1109/CVPRW.2018.00262>.
- Roy, P., Ghosh, S., Bhattacharya, S., & Pal, U. (2018). Effects of degradations on deep neural network architectures. *arXiv preprint arXiv:1807.10108*.
- Rubin, D. B., Van Hooser, S. D., & Miller, K. D. (2015). The stabilized supralinear network: A unifying circuit motif underlying multi-input integration in sensory cortex. *Neuron*, 85, 402–417. <https://doi.org/10.1016/j.neuron.2014.12.026>. PMID: 25611511. PMCID: PMC4344127.
- Sacramento, J., Costa, R. P., Bengio, Y., & Senn, W. (2018). Dendritic cortical microcircuits approximate the backpropagation algorithm. In *Advances in neural information processing systems* (pp. 8735–8746). Montreal, Canada.
- Salakhutdinov, R. (2015). Learning deep generative models. *Annual Review of Statistics and Its Application*, 2, 361–385. <https://doi.org/10.1146/annurev-statistics-010814-020120>.
- Samek, W., & Müller, K.-R. (2019). Towards explainable artificial intelligence. In W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, & K.-R. Müller (Eds.), *Explainable AI: Interpreting, explaining and visualizing deep learning* (pp. 5–22). Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-030-28954-6_1.

- Saxe, A. M., McClelland, J. L., & Ganguli, S. (2013). Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. arXiv preprint arXiv:1312.6120.
- Scholte, H. S., Losch, M. M., Ramakrishnan, K., de Haan, E. H. F., & Bohte, S. M. (2018). Visual pathways from the perspective of cost functions and multi-task deep neural networks. *Cortex*, 98, 249–261. <https://doi.org/10.1016/j.cortex.2017.09.019>. PMID: 29150140.
- Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., et al. (2018). Brain-Score: Which artificial neural network for object recognition is most brain-like? bioRxiv, 407007. <https://doi.org/10.1101/407007>.
- Seeliger, K., Fritsche, M., Güçlü, U., Schoenmakers, S., Schoffelen, J.-M., Bosch, S. E., et al. (2018). Convolutional neural network-based encoding and decoding of visual object recognition in space and time. *Neuroimage*, 180, 253–266. <https://doi.org/10.1016/j.neuroimage.2017.07.018>. PMID: 28723578.
- Serre, T. (2019). Deep learning: The good, the bad, and the ugly. *Annual Review of Vision Science*, 5, 399–426. <https://doi.org/10.1146/annurev-vision-091718-014951>. PMID: 31394043.
- Serre, T., Kreiman, G., Kouh, M., Cadieu, C., Knoblich, U., & Poggio, T. (2007). A quantitative theory of immediate visual recognition. *Progress in Brain Research*, 165, 33–56. [https://doi.org/10.1016/S0079-6123\(06\)65004-8](https://doi.org/10.1016/S0079-6123(06)65004-8).
- Shen, G., Horikawa, T., Majima, K., & Kamitani, Y. (2019). Deep image reconstruction from human brain activity. *PLoS Computational Biology*, 15, e1006633. <https://doi.org/10.1371/journal.pcbi.1006633>. PMID: 30640910. PMCID: PMC6347330.
- Shi, J., Wen, H., Zhang, Y., Han, K., & Liu, Z. (2018). Deep recurrent neural network reveals a hierarchy of process memory during dynamic natural vision. *Human Brain Mapping*, 39, 2269–2282. <https://doi.org/10.1002/hbm.24006>. PMID: 29436055. PMCID: PMC5895512.
- Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience*, 24, 1193–1216. <https://doi.org/10.1146/annurev.neuro.24.1.1193>. PMID: 11520932.
- Sinz, F., Ecker, A. S., Fahey, P., Walker, E., Cobos, E., Froudarakis, E., et al. (2018). Stimulus domain transfer in recurrent models for large scale cortical population prediction on video. In *Advances in neural information processing systems* (pp. 7199–7210). Montreal, Canada. <https://doi.org/10.1101/452672>.
- Spoerer, C. J., Kietzmann, T. C., & Kriegeskorte, N. (2019). Recurrent networks can recycle neural resources to flexibly trade speed for accuracy in visual recognition. bioRxiv, 677237. <https://doi.org/10.32470/CCN.2019.1068-0>.
- Spoerer, C. J., McClure, P., & Kriegeskorte, N. (2017). Recurrent convolutional neural networks: A better model of biological object recognition. *Frontiers in Psychology*, 8, 1551. <https://doi.org/10.3389/fpsyg.2017.01551>. <https://doi.org/10.1101/133330>.
- Storrs, K. R., & Kriegeskorte, N. (2019). Deep learning for cognitive neuroscience. In M. Gazzaniga (Ed.), *The cognitive neurosciences* (6th ed.). Cambridge, MA: MIT Press.
- Sussillo, D., & Barak, O. (2013). Opening the black box: Lowdimensional dynamics in high-dimensional recurrent neural networks. *Neural Computation*, 25, 626–649. https://doi.org/10.1162/NECO_a_00409. PMID: 23272922.
- Tacchetti, A., Isik, L., & Poggio, T. (2017). Invariant recognition drives neural representations of action sequences. *PLoS Computational Biology*, 13, e1005859. <https://doi.org/10.1371/journal.pcbi.1005859>. PMID: 29253864. PMCID: PMC5749869.
- Tang, H., Schrimpf, M., Lotter, W., Moerman, C., Paredes, A., Caro, J. O., et al. (2018). Recurrent computations for visual pattern completion. *Proceedings of the National Academy of Sciences, U.S.A.*, 115, 8835–8840. <https://doi.org/10.1073/pnas.1719397115>. PMID: 30104363. PMCID: PMC6126774.
- Thompson, J. A. F., Bengio, Y., Formisano, E., & Schönwiesner, M. (2018). How can deep learning advance computational modeling of sensory information processing? arXiv preprint arXiv:1810.08651.
- Thorat, S., van Gerven, M., & Peelen, M. (2019). The functional role of cue-driven feature-based feedback in object recognition. arXiv preprint arXiv:1903.10446. <https://doi.org/10.32470/CCN.2018.1044-0>.
- Tishby, N., & Zaslavsky, N. (2015). Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop (ITW)* (pp. 1–5). Jerusalem, Israel. <https://doi.org/10.1109/ITW.2015.7133169>.
- Tripp, B. P. (2017). Similarities and differences between stimulus tuning in the inferotemporal visual cortex and Convolutional networks. In *2017 International Joint Conference on Neural Networks (IJCNN)* (pp. 3551–3560). Anchorage, AK. <https://doi.org/10.1109/IJCNN.2017.7966303>.
- Tschopp, F. D., Reiser, M. B., & Turaga, S. C. (2018). A connectome based hexagonal lattice convolutional network model of the Drosophila visual system. arXiv preprint arXiv:1806.04793.