source: arXiv:1903.01458v1 [q-bio.NC] date: 2019 title: Deep Learning for Cognitive Neuroscience author: Katherine R. Storrs and Nikolaus Kriegeskorte

認知神経科学のためのディープラーニング

Deep Learning for Cognitive Neuroscience

Katherine R. Storrs and Nikolaus Kriegeskorte

要約

ニューラルネットワークモデルは，画像を認識し，テキストを理解し，言語を翻訳し，多くの人間のゲームを人間や超人のレベルでプレイすることができるようになりました。 これらのシステムは高度に抽象化されていますが，生物の脳にヒントを得ており，生物学的に妥当な計算のみを使用しています。 今後数年間で，ニューラルネットワークは，大量のラベル付きデータセットからの学習への依存度を減らし，よりロバストで汎用的なタスクパフォーマンスを発揮するようになるでしょう。 ニューラルネットワークの成功例や失敗例から，脳が得意とするさまざまなタスクに必要な計算量を知ることができます。 また，ディープラーニングは，認知理論を検証するためのツールでもあります。 理論を検証するためには，提案された情報処理システムを大規模に実現し，その実現可能性や出現する行動を評価する必要があります。 ディープラーニングは，原理や回路モデルから，複雑なタスクを実行できるエンドツーエンドの学習可能なモデルへとスケールアップすることができます。 認知神経科学者が深層学習を利用する際には，理論の構築から完全な計算モデルの構築まで，さまざまなレベルで利用することができます。 深層学習が進むことで，認知神経科学の中核をなす壮大な課題である，認知や知覚が脳の中でどのように実現されているかの理解に近づくことができます。

1. はじめに: 神経科学にインスパイアされた AI と AI にインスパイアされた神経科学

脳がどのように推論し，記憶し，知覚し，行動するのかを説明するためには，理論が生物学から行動へと橋渡しする必要があります。 厳密な説明には，神経生物学的に妥当なコンポーネントを使用して認知プロセスを実装するモデルが必要である（Poeppel, 2012; Kriegeskorte & Douglas, 2018; Kriegeskorte & Mok, 2017）。 単純化されたシミュレートされたニューロンで構成された人工ニューラルネットワークは，長い間そのようなモデルを約束してきました (McCulloch & Pitts, 1943; Rumelhart & McClelland, 1986)。 計算機と方法論の進歩のおかげで，ニューラルネットワークは現在，パターン認識問題 (例えば Russakovskyら 2015) だけでなく，言語翻訳 (例えば Wu ら 2016)，視覚的推論 (例えば Santoro ら2017)，ゲームプレー (例えば Mnih ら 2015，Jaderberg ら 2018) などの多くの認知的課題に対して，他のすべての工学的解決策を凌駕しています。 心理学や神経科学からのアイデアはエンジニアにインスピレーションを与え，現代のネットワークにおける多くの機能の基礎となっています（Kriegeskorte, 2015; Hassabis et al.,2017)。 視覚タスクに使用されるネットワークは，哺乳類の視覚野と同様に，空間的に制限された受容野で画像を階層的に処理します（LeCunら，1998年，Russakovskyら，2015年）。 “注意ネットワーク” は，入力の一部を動的に選択し，処理資源を順次投入していく（Xu et al. 2015)。 学習における海馬と大脳新皮質の相補的な役割に関する研究（Kumaran et al.,2016; McLelland, McNaughton & O’Reilly, 1995) は，エピソード記憶に似たものを持つネットワークにインスピレーションを与え，「経験の再生」を介して学習をブートストラップします（例えば Mnih et al.2015)。

一方，認知神経科学者は，深層学習のアイデアや成果に刺激を受け，情報を得ています。 ニューラルネットワークは，視覚的物体認識がどの程度再帰的処理を必要とするかなどの難しい問題に対処するのに役立つ概念実証を提供します (Riesenhuber & Poggio, 1999; O’Reilly et al., 2013; Spoerer, McClure & Kriegeskorte, 2017)。 機械学習は新たな視点を提供し，異なる脳領域に存在する学習目的と，領域特異的なサイトアーキテクチャーに根付いている可能性のある事前の世界知識という観点から，皮質と皮質下の特殊性について考えることを促します (Marblestone & Kording, 2016)。 機械に見たり考えたりすることを教えようとする謙虚な経験は，神経科学者に，これらの達成がいかに計算上困難であるかを印象づけました。 工学的には，比較的単純な計算要素でどれだけのことができるかを実証しました(例: Oliva & Torralba, 2001 の「GIST」モデル)。

ニューラルネットワークモデルは，パターンが静的であるか動的であるかに関わらず，パターン認識やパターン生成のタスクに優れています。 しかし，人間が持っている，細かい部分から総合的に推論したり，一つの経験から新しい概念を学んだり，全く新しい領域にうまく一般化したりする能力は，まだ捉えられていません（Lake et al. 2017, Kriegeskorte 2018）。 記号的表現と確率的推論に基づく，より抽象的な認知計算モデルは，現在のところ，これらの驚くべき人間の認知能力に匹敵するものに近づいていますが，計算の効率性という点では人間の脳には及ばず（Gershman, Horvitz & Tenenbaum, 2015），神経生物学との関連付けも難しくなっています。 人間の認知とその脳への実装を理解するには，認知レベルの計算モデルとニューラルネットワークモデルの両方が必要になります。 ここでは，後者に焦点を当て，認知神経科学者がこれらのモデルを自分の研究に取り入れることを呼びかけます。

2. ニューラルネットワークとは何か？

脳の計算モデルには，モデルニューロンをスパイクさせた回路を生物学的に詳細にシミュレーションしたものから，抽象度の高い認知モデルまでさまざまなものがある。 ニューラルネットワークモデル」という言葉は，通常，抽象度が中間レベルのモデルを指す。 ニューラルネットワークモデルは，相互に接続されたユニットで構成されており，それぞれのユニットが入力の加重和を計算し，それを非線形活性化関数に通します（図1a）。 結果として得られるスカラー出力は，理想化されたニューロン（「ユニット」と呼ばれる）の収縮率と考えることができ，瞬時に反応し，適応期間や不応期間はありません。 このようなユニットのネットワークは，入力（写真など）と出力（各写真に写っている主な物体の名前など）の間の任意の複雑な機能を，線形-非線形のサブファンクションの組み合わせとして実装することができます。

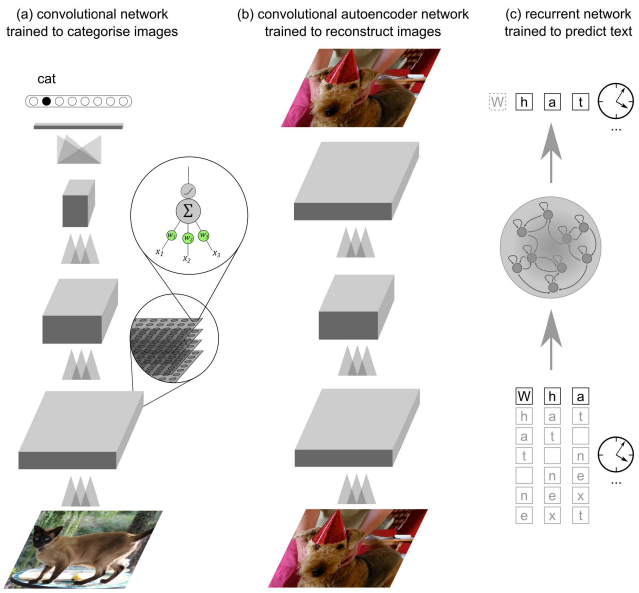

図 1. ニューラルネットワークの種類を示す図。 (a) 4 つの隠れ層（3 つの畳み込み層と 1 つの完全連結層）を持つ畳み込みニューラルネットワーク。 ラベル付きの画像で教師による学習を行った後，学習していない新しい写真の中の物体を分類することができる。 畳み込み層は，複数の「特徴マップ」で構成されており，それぞれのマップには，入力の異なる位置に集中する空間的に限定された受容野を持つユニットが含まれているが，同じ接続重みを共有している。 これにより，各マップは，画像の各位置における特定の特徴の有無を表すことができる。 (b) 3 つの畳み込み隠れ層を持つ畳み込みオートエンコーダー。教師なしの学習により，新しい画像を真ん中の「ボトルネック」層でより簡潔なフォーマットに圧縮し，その後，ほぼ再構成することができる。 (c) リカレント・ニューラル・ネットワークは，テキストを1文字ずつ入力し，教師なしの学習により，見たことのない文の次の文字を予測することができる。

2.1 学習可能な結合係数: ネットワークの柔軟な知識

生物学的なニューラルネットワークのように，このモデルは，環境のすべてを知っているわけではなく，要求されるタスクをどのように実行するかも知らない状態で登場します。 その代わり，学習します。 ネットワーク内の知識は，脳のシナプスの重みに似た，ユニット間の接続に関連する重みによって保持されます。 重みは通常，最初は小さな乱数に設定され，学習アルゴリズムによって段階的に更新されます。

学習アルゴリズムには，大きく分けて「教師あり」「教師なし」「強化」の3種類があります。 教師あり学習では，学習アルゴリズムは，多数の入力パターンに対する出力を，あらかじめ指定された望ましい出力に近づけるように重みを調整する。 コストは，現在の出力が望ましい出力からどれだけ乖離しているかを表します。 重みは，その調整がコストをどれだけ強く減少させるかに比例して調整される。 このためには，各重みに関するコスト関数の導関数を計算する必要がある。 この導関数を計算するための効率的なアルゴリズムがバックプロパゲーションです。 重みに対するコストの導関数は，学習入力に対する出力を目的の出力に近づけるために，各接続の重みをどの方向にどれだけ調整すればよいかを教えてくれます。 機械視覚や言語翻訳など，近年の工学分野における多くの成果は，膨大な量のグランドトゥルース・ラベル付きデータセットを用いた教師付き学習によるものです（Russakovsky et al.2016)。

教師なし学習では，学習データはラベルなしで提供されます。 ネットワークは，入力データの統計情報をモデル化して学習します。 例えば，自己符号化器は，入力パターンを低次元の「コード」層に圧縮することを学習します。 ネットワークは，入力パターンを低次元のコードにマッピングし（符号化器コンポーネント），その後，完全な入力パターンに戻す（復号化器コンポーネント）。 自己符号化器は，各入力パターンの符号化と復号化（再構成）を学習する（例：図1b）。 また，ノイズを除去したり，動的な入力の将来の状態を予測したりすることも可能である。 教師なしの学習信号は非常に豊富であり（画像オートエンコーダーは，再構成を試みるすべてのピクセルに対して学習信号を導き出す），ネットワークは経験データの情報をよりよく活用することができる。 ネットワークは，特定のタスクに関連するものだけでなく，データのすべての規則性を学習しようとします。

強化学習 (Sutton & Barto, 1998) では，ネットワークは行動 (例えば，シミュレーションまたは実環境での動き) を出力し，可能な行動に関連する期待される累積報酬をモデル化して学習します。 環境中の特定のイベントは報酬として定義され，その発生が学習を促進します。 強化学習は深層学習と組み合わせることで，価値関数をニューラルネットワークで表現することができます (例: Deep Q-Learning; Mnih et al. 2015)。 環境中に報酬がほとんどない場合，強化学習は，教師なし学習や教師あり学習に比べて，重みを調整するための直接的な制約が少ないため，難しいとされています。 しかし，生物学や心理学に根ざした強化学習は，生態学的な妥当性が高い。 また最近では，例えばビデオゲームのプレイ領域での成功に見られるように，工学的にも大きな進歩をもたらしています (Mnih et al. 2015; Jaderberg et al. 2018)。

ネットワークは学習データをある程度過学習させるので，常に新しい入力 (画像など) と出力 (カテゴリーラベルなど) の独立したテストセットで検証されます。 生物の脳とは異なり，学習 (接続重みの調整) と知覚 (固定された接続重みを持つネットワークを介して新しい入力を処理すること) は一般的に別々の処理であり，学習は一般的に明確な初期段階で行われます。

2.2 アーキテクチャ: ネットワークの柔軟な構造

異なる構成のユニットを連結することで，無限に多くのネットワーク・アーキテクチャを作ることができます。 フィードフォワードアーキテクチャーでは，ユニットは単一の処理階層を形成し，どのユニットも自分自身や前のユニットに接続されていません。 フィードフォワードネットワークは，フィー

ドフォワードスイープを超える時間的なダイナミクスを持たず，入力から出力への静的なマッピングを計算します。 トップダウンフィードバックのように，有向接続グラフに1つ以上のループを含むネットワークはリカレントであり，その内部状態は離散的なステップで時間とともに進化する。 リカレントネットワークは，ビデオやテキストなどの時系列データの処理によく用いられ，各時間ステップで新しいフレームや文字がネットワークに入力されます（図1c）。

入力データと出力レスポンスの間に「隠れた」1 層のユニットしかないフィードフォワードネットワークは，浅いネットワークと呼ばれています。 浅いフィードフォワードネットワークは，ユニットの数が増えれば増えるほど精度が上がり，すでにあらゆる連続関数を近似することができる (Hornik, 1991)。 しかし，中間処理層を増やすことで，前のユニットが計算した特徴を後のユニットが再利用・再結合することで，同じユニット数でもより複雑な機能を表現できるようになる。 入力ユニットと出力ユニットの間に 2 つ以上の隠れ層が介在している場合，そのネットワークはディープと呼ばれる。 現代のコンピュータビジョンでは，人間に近い画像クラシフィケーション性能を発揮するディープニューラルネットワークは，通常，10 以上の隠れ層と100万以上の個々のユニットを含んでいます (例えば Simonyan & Zisserman, 2015; He et al, 2016)。

特殊なアーキテクチャやユニットタイプは，ドメインやタスクに関する事前の知識を活用することができます。 たとえば，ディープ・コンボリューショナル・ニューラル・ネットワーク（CNN）は，視覚タスクによく使われる (図1a)。 CNN はほ乳類の視覚野にゆるやかにインスパイアされており，空間的に制限された受容野と共有されたウェイトを持つユニットを使用する (LeCun et al.) 重みの共有とは，各層の複数のユニットが接続重みの同じテンプレートを使用し，このテンプレートを入力の異なる位置に適用することを意味します。 これにより，ネットワークが学習しなければならないパラメータの数が減り，同じ特徴（例えば，垂直方向のエッジや鳥の翼など）が画像内のどこにでも現れる可能性があるという事前の確信を得ることができます。 多くのリカレントネットワークでは，「長短期記憶」（LSTM） ユニット (Hochreiter & Schmidhuber, 1997) を使用しています。 他のカスタムユニットは，神経科学的な考え方である正準皮質計算にヒントを得て，入力に対して局所的な応答の正規化や最大プールを行います (Carandini & Heeger, 2012; Riesenhuber & Poggio, 1999)。

3. 認知と知覚のモデルとしてニューラルネットワークを用いる

認知神経科学者は，様々なレベルでニューラルネットワークに関わることができ，技術的なノウハウやリソースの投入の度合いも様々です。 一方では，ネットワークについて読み，考え，その成功，失敗，原理を理論や仮説の種にすることができます。 タスクを実行するモデルを，知覚や認知の計算メカニズムの原理的な証明として考えることもできます。 たとえば，視覚的な物体認識がフィードフォワード・システムでどの程度達成できるかは，歴史的な論争になっています (たとえば Riesenhuber & Poggio, 1999)。 最近の CNN では，自然なシーンでの物体の命名やセグメンテーションが，横方向の接続やフィードバック接続なしで，かなりの程度達成できることが実証されています。 同時に，リカレント接続は，困難な状況下でニューラルネットワークの認識性能を大幅に向上させることができます (O’Reilly et al., 2013; Spoerer, McClure & Kriegeskorte, 2017; Nayebi et al. 2018)。

より深いレベルでは，他の研究者によって構築された事前に訓練されたネットワーク（その多くはオンラインで入手可能）を利用し，それらを行動や内部表現の観点から人間や非人間の被験者と比較することで，DNNモデルを実証的な研究に導入することができます。 脳の表現と行動を説明するモデルの能力に対するアーキテクチャとトレーニングの寄与を分離するために，我々の特定の仮説に関連する刺激と課題で，事前に構築されたアーキテクチャを再トレーニングまたは微調整することができます。

最後に，異なるトレーニングデータ，学習目的，アーキテクチャを模索しながら，新しいモデルを開発することができます。 これにより，環境，学習プロセス，神経構造のさまざまな側面が，認知機能にどのように影響するかを検証することができます。 複雑なニューラルネットワークモデルを設計し，それを評価するための行動実験や神経実験を行うには，幅広い専門知識が必要であり，1つの研究室で統合することは容易ではない。そのため，研究室間での新たな共同作業や共有方法を開発する必要があります。 研究室によっては，モデルの構築に重点を置くところもあれば，共有したモデルを脳や行動のデータで検証することに重点を置くところもあり，また，既存のモデルに残された欠点を明らかにして定量化するためのタスクを開発するところもあるでしょう。

3.1 行動データと脳活動データを用いたニューラルネットワークモデルの検討

行動レベルでは，モデルは対象となるタスクを人間や動物の被験者と同様のレベルで実行できる必要があります。 しかし，人間が得意とするタスク (例：画像の言語的説明の生成，Xu et al. 2015) をモデルが単にうまくこなすだけでは，モデルが人間と同様の計算を行い，同様の内部表現を介して出力に到達するという強力な証拠にはなりません。 優れたモデルは，タスクの異なるインスタンスにおける行動のバリエーションの詳細なパターンを予測することができます。

また，各モデルの内部表現を人間の知覚的判断と比較することもできます。 モデルの内部活性化パターンにおける刺激や条件の類似性は，人間に知覚される類似性を予測するはずです（Kubilius, Bracci & Op de Beeck, 2016）。 モデル内で同一の反応を引き起こす刺激は，人間にも同一に見えるはずである（Wallis et al. 2017)。 刺激が劣化したり歪んだりすると，モデルの性能は人間の性能と定量的に似た形で低下するはずである (これに反して Geirhos ら 2018）は CNN は人間よりも画像ノイズに対するロバスト性がはるかに低いことを発見した)。 モデルは，理想的には単一刺激レベルで，混乱やエラーのパターンを予測できなければならない。 例えば，ある研究では，CNN の分類は，サルとヒトの物体マッチングの混同をオブジェクトカテゴリレベルで正確に予測したが，どの特定の画像が混同したかを予測できなかった（Rajalingham et al. 2018)。

内部表現のレベルでは，モデルは，空間 (脳領域) と時間 (処理の順序) を超えて，同じ表現変換の順序を経るはずです。 モデルと脳の間の内部表現を比較することは，モデルの活動パターンと脳の活動パターンの間の詳細な空間的・時間的な対応関係のマッピングを知らない可能性があるため，複雑である。 Diedrichsen (2018) は，ニューラルネットワークモデルをテストするために，モデルの内部表現と脳活動の測定値を比較することで，表現モデルの枠組みを紹介しています。 簡単に言うと 符号化モデルは，測定された各応答チャネルをニューラルネットワークのユニットの線形結合として予測します (Kay et al. 2008; Dumoulin et al. 2008; Mitchell et al. 2008; Naselaris & Gallant, 2011)。 表象類似性分析 (Kriegeskorte et al., 2008; Nili et al., 2014; Kriegeskorte & Diedrichsen, 2016） およびパターンコンポーネントモデリング（Diedrichsen et al., 2011) は， 非類似性または類似性の刺激ごとの行列を使用して，表象空間を特徴づけ，モデル層を脳領域と比較する。 これらのアプローチには，それぞれ長所と短所があります。 例えば，符号化モデルは，各ボクセルの応答を個別に分析し，これらを大脳皮質上にマッピングするのに適しています (Guclu & van Gerven 2015; Eickenberg et al, 2017)， 表現類似性分析とパターンコンポーネントモデリングは 線形エンコーディングモデルの何千ものパラメータを調べる必要性を回避する一方で，脳の表現における異なる特徴の相対的な有病率をモデル化するために重みを推定する柔軟性を可能にします (例えば Khaligh-Razavi & Kriegeskorte, 2014; Khaligh-

Razavi, Henriksson, Kay & Kriegeskorte, 2017)。 符号化モデル，表現上の類似性分析，パターンコンポーネントモデリングは，研究の目標に応じて適切に組み合わせることができる表現モデリング技術のツールボックスの一部と考えるのが最適である (Diedrichsen & Kriegeskorte, 2016)。

Khaligh-Razavi & Kriegeskorte (2014) は カテゴライズを学習した CNN の後層が，下側頭皮質における自然物画像の表現を予測する27の浅いコンピュータビジョンモデルのどれよりも優れていることを発見した。 また，カテゴライズで学習した CNN は，電気生理学 (Cadieu et al. 2014), fMRI (Guclu & van Gerven 2015; Eickenberg et al. 2017; Wen et al, 2017), MEG (Cichy et al., 2016; Cichy et al., 2017) または EEG (Greene & Hansen, 2018) であり，音声と音楽で訓練されたDNNは聴覚皮質活動を最もよく予測した (Kell et al. 2018)。 神経活動から刺激をデコードまたは再構築することに焦点を当てたこのアプローチのバリエーションは，同様に，CNN が視覚領域の活動をデコードするための優れた特徴空間を提供することを示している (Wen et al. 2017; Horiwaka & Kamitani, 2017)。

どのような方法で DNN モデルを評価するにしても，対照モデルとの比較，および/または，計算や学習の代替理論を具体化する複数の DNN をテストすることが重要です。 刺激処理の些細なモデルであっても，ランダムに重み付けされていない DNN（Cichyら，2016など）と同様に，感覚脳領域の小さいながらも有意な分散を説明する (Khaligh-Razavi & Kriegeskorte, 2014のピクセルベースのコントロールモデルなど)。

3.2 ニューラルネットワークモデルの解釈: シリコンウェハース上での認知神経科学

ある課題を実行し，脳や行動のデータを（ノイズや被験者間のばらつきで決まる限界まで）説明できる DNN ができれば，重要なマイルストーンに到達したことになる。 しかし，我々の仕事はまだ終わっていない。 モデルはタスクで訓練され，多くのパラメータを持っているため，その計算メカニズムを理解していない可能性があります。 そのため，モデル内の表現とダイナミクスを分析する必要がある。 脳と同じように，DNNもさまざまな解像度と抽象度で研究することができる。 脳とは異なり，DNNはシステム全体に完全にアクセスできる。 ネットワークを停止して再起動したり，異なる環境やタスク要求の下で再学習させたり，疲労や損傷を受けることなく継続的にデータを収集したり，構成要素の任意の組み合わせを病変させて復活させたり，刺激の最適化技術を用いてどのような特徴を学習したかを確認したり，さらにはその特性の一部を解析的に証明したりすることが可能である (1)。.

> 脚注1：例えば，群論の数学は，深いフィードフォワードネットワークの連続した層がますます複雑な特徴を学習する理由を説明するために使用されています (Paul & Venkatasubramanian, 2014)。

自由なアクセスが可能なため，DNN は本物の脳よりもはるかに「電気生理学的」手法に適している。 ネットワーク内の個々のユニットが学習した特徴の優先順位を分析するために，実験者は何千，何百万もの刺激を提示して各ユニットの活性化を記録したり（Yosinski et al. 2015)。 図 2c は，映画レビューで訓練されたリカレントテキスト予測ネットワーク内の単一の LSTM ユニットの活性化が，ネットワークが予測している通路が展開するにつれて変化する様子を示しています (Radford, Jozefowicz & Sutskever, 2018)。 この印象的な例では，視覚化されたユニットが，映画レビューがポジティブな感情を表現しているのか，ネガティブな感情を表現しているのかを学習したように見えます。

また，DNN は，脳には適用できないような新しい尋問技術にも適しています。 ニューラルネットワークモデルは微分可能であるため，勾配降下最適化を用いて，個々のユニットに最適な刺激を生成するなど，ネットワーク活性化の特定のパターンを引き出す刺激を作成することが可能である（例えば，Yosinski et al. 2015)。 Olah (2017) が見事にまとめて説明しているように，これを行うためのいくつかの技術が存在します。 図2a は Google の「DeepDream」アルゴリズム (Mordvintsev, Olah & Tyka, 2015) を用いて，2 つのカテゴリー学習済み CNN の 4 つの異なる層のそれぞれを強く活性化するように反復的に最適化されたノイズ画像を示している。 情報の流れをネットワークで追跡すると，ある出力決定をサポートするために刺激のどこで，あるいはいつ証拠が引き出されたのか (これも Olah, 2018) が見事に説明している) あるいは動的な「注意」を持つネットワークが刺激のどの部分に現在リソースを割いているのかを明らかにすることができる (図2b)。

直接的な最適化手法により，深層視覚ネットワークが学習する特徴の階層が，哺乳類の腹側ストリームにおけるものと驚くほど類似していることが明らかになりました (例えば，Yamins & DiCarlo, 2016 を参照)。 また，特徴の好みがトレーニングの過程でどのように発展するかを探ることができ，人間の知覚学習との類似性が示唆されています (Wenliang & Seitz, 2018)。 課題を実行するネットワークは，神経コードがどの程度疎であるか，あるいは分散している可能性があるかといった，認知神経科学における長年の議論に決着をつけるのに役立つかもしれません (Agrawal, Girschick & Malik, 2014; Zhuang, Wang, Yamins & Hu, 2017; Morcos, Barrett, Rabinowitz & Botvinick, 2018)。

ネットワーク表現は，より高い抽象度で要約することもできる。 従来の機能局在化法では，特定のクラスの刺激に対して優先的に選択されるユニットや層を特定することができますが，シーンを分類するためだけに訓練された視覚ネットワークに物体選択ユニットが出現するなど，当初は直感に反する結果となりました (Zhou et al. 2014)。 表象類似性分析 (RSA) は ネットワークがその処理の層 (例えば Khaligh-Razavi & Kriegeskorte, 2014) またはタイムステップ (例えば Cichy et al, 2016) を越えて刺激情報をどのように変換するかを示すのに役立ちます。

ニューラルネットワークモデルにアクセスできるようになったことで，効率的なモデルの検索と照合のための新たな可能性が生まれ，実験計画の立て方が大きく変わるかもしれません。 現在，実験条件は，言葉による理論が示唆する仮説を区別するために，「手探り」で選ばれています。 将来的には，条件をアルゴリズムで最適化して，明示的な計算モデル間での差別化を図ることができるかもしれません (例えば Wang & Simoncelli, 2008)。

4. 多様な認知モデルとしてのニューラルネットワーク

視覚的な物体や顔の認識は，DCNN が静的な刺激と単一のフィードフォワード処理の掃引を用いて人間レベルに近い性能に到達できたことから，エンジニアや認知神経科学者から大きな注目を集めています（He et al. 2015, Taigman et al 2014)。 しかし，深層学習革命は，言語や推論などの私たちの高次の能力を含む，認知と知覚のほぼすべての領域に触れています。

4.1 ニューラルネットワークにおける言語

言語処理を行う DNN は，ウェブサイトやオンライン検索結果の自動翻訳などに広く利用されている。 このようなハイレベルなタスクは，我々がまだ出会っていないよりも複雑な構成のネットワークアーキテクチャによって実現される傾向にある。 最先端の Google Neural Machine Translation システム (Wu et al., 2016) は，2 つの 8 層リカレントニューラルネットワークで構成されています。 1 つはテキストを元の言語からネットワーク内の潜在的な表

現にエンコードし，もう 1 つは潜在的な表現を受け取ってターゲット言語にデコードするもので，「注意」の重みを使って，エンコードされた表現のどの部分を処理するかを各タイムステップで変更します。 潜在表現の形式や復号化時の注意の割り当てなど，ダイナミックなプロセス全体は，数千万の人間が翻訳した文章を用いた学習によりバックプロパゲーションで学習される (Wu et al. 2016)。 同様のアプローチは，画像の言語的説明を生成するためにも用いられている (Xu et al. 2016)。 そこでは 画像処理畳み込みネットワークを使用して画像を高レベルの画像特徴の潜在表現に変換し，再帰ネットワークを学習してそれらの特徴に基づく一連の単語を出力し，空間的注意を使用して各時間ステップで画像のどの部分が最も単語選択を導くかを調節する (図2b)。

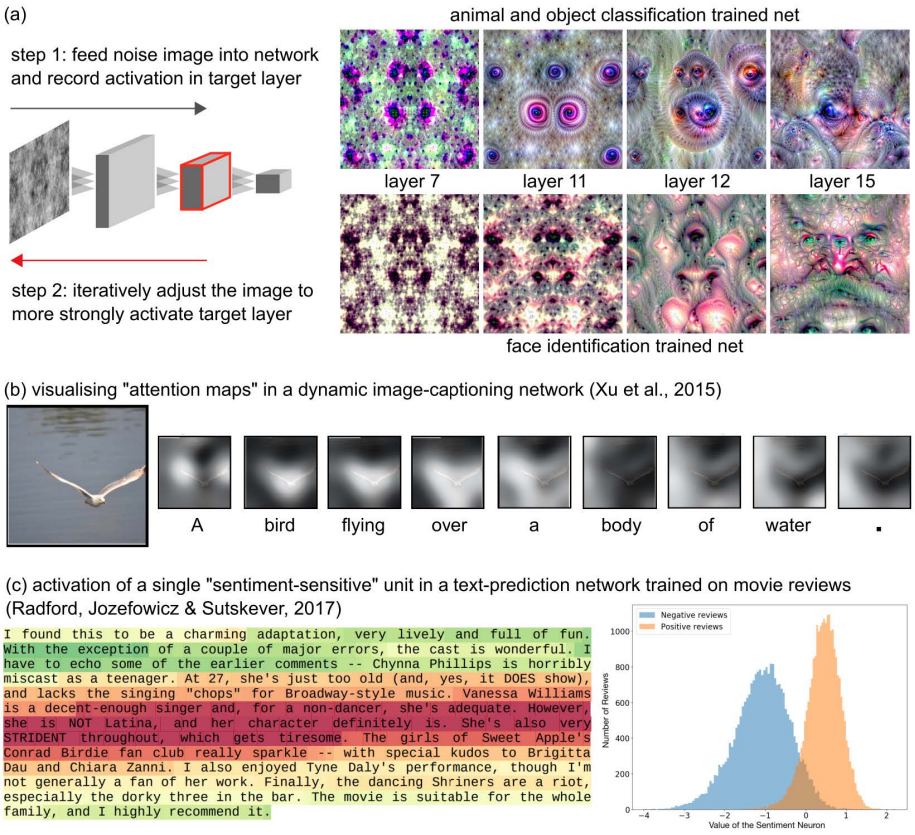

図 2 ディープ・ニューラル・ネットワークの相対的な透明性。 (a) ネットワークの学習に使用されるバックプロパゲーションアルゴリズムは，学習したネットワークの特徴を視覚化するためにも使用できる。 ここでは，ノイズ画像を繰り返し最適化することで，16 層の DCNN の 4 つの層それぞれのユニットの活性化度を高めている。 訓練は，学習される特徴に顕著な影響を与える。 (b) 「注意ネットワーク」は，空間的な注意マスク（入力に対する乗法的な重み付け）を持ち，それを学習することで，訓練課題でのパフォーマンスを向上させるように割り当てられる。 ここでは，画像のテキストキャプションを生成するように訓練されたリカレントネットワークが，時間ステップごとに1つの単語を出力し，各時間ステップでこの画像のどの領域に注意を向けたかを，白いパッチが示している（Xu et al. 2015 の許可を得て複製）。 (c) テキストを1文字ずつ予測するように訓練された再帰ネットワークの単一のLSTMユニットの出力活性化を，ネットワークが見たことのない例文について可視化したもの。活性化が高い場合は緑，低い場合は赤で示される。 このユニットの応答は，レビューの大規模なサンプルを分析したときに，レビューの感情がポジティブかネガティブかを予測した (Radford, Jozefowicz & Sutskever, 2018 から許可を得て転載）。

これらのネットワークは工学的な解決策ではあるが，複数のパターン認識 DNN を構成し，注意誘導 (心理学的に着想された機能) と組み合わせることで，言語翻訳や画像描写などの人間特有のスキルを実現できることは驚くべきことである。 認知神経科学者は，このような原理を人間の認知や言語のニューラルネットワークモデルに取り入れ始めたところである。 Devereux, Clarke and Tyler (2018) は モジュール式の視覚的意味モデルを用いて，人間の観察者が画像内のオブジェクトに名前を付けたときのfMRI活性化パターンを予測した。 このモデルでは，画像はまず，物体の分類について訓練されたフィードフォワードの画像 CNN によって処理され，次に，CNN からの五層目の特徴は，描かれた物体の多数の意味的特性 (「is a fruit」,「grown on trees」など) の有無をエンコードするように訓練された単層の再帰ネットワークに供給された。 CNN 内での表現の類似性は，ヒトの視覚野での類似性を予測するのに役立ち，再帰的意味ネットワーク内での類似性は，意味処理に関連する領域であるヒトの末梢皮質での類似性を予測するのに役立つことがわかった。

4.2 ニューラルネットワークにおける推論

人工的な一般知能はまだ見えていないかもしれませんが，いくつかの深層ネットワークは印象的に柔軟な推論を示しています。 Relation Networks（Santoro et al., 2017）は，画像内のオブジェクト間の関係を推論することを含む，有名な難しいタスクである CLEVR ベンチマークで超人的なパフォーマンスを達成しました。 質問の例としては，”緑の物体と同じ形の物体はいくつありますか？”，”黄色の金属製の円筒と同じ大きさのゴム製のものはありますか？”などがあります。(Santoro et al., 2017) などがある。 フィードフォワードの画像 CNN が画像から高レベルの視覚的特徴を抽出し，リカレントの言語ネットワークが言葉の質問をエンコードする。 次に 関係ネットワークは，質問とともに，画像内のすべての可能な「オブジェクト」のペア（空間的な位置のペアとして運用される）を考慮し，それぞれの可能な応答の可能性を評価する。 関係性ネットワークはCLEVRタスクで95%の精度を達成したが，人間のパフォーマンスは 93% 程度である (Santoro et al. 2017)。 画像，言語，関係推論サブネットワークは，エンドツーエンドの微分可能なシステムを形成し，バックプロパゲーションを用いて単純に一緒にトレーニングされます。

構成的なネットワークアーキテクチャは DNN のインテリジェントな動作を実現するための重要なエンジニアリングトリックであることが証明されています。 Yang, Song, Newsome & Wang (2017) による最近の研究では，課題の要求がどのようにニューラル・アーキテクチャを決定するかを解明し始めている。 リカレント・ネットワークは，スピードド・クラシフィケーションやディレイド・マッチ・トゥ・サンプルといった，人間や動物の認知研究から得られた20の非常に単純なタスクを実行するよう同時に訓練された。 その結果，どのユニットがどの課題に関与しているかを分析することができました。 その結果，ネットワークは 20 の課題の構成的な表現を学習しており，かだいの構成要素（例えば「手掛かりの方向を記憶する」や「手掛かりと反対の方向に反応する」など）に特化したユニットのクラスタが存在し，任意の課題は，その下位課題の構成に応じて，クラスタの連合によって実行されることがわかりました。

4.3 ニューラルネットワークの実体

DNN は必ずしも実体のないものではありません。 例えば，サルが目標に向かって手を伸ばす筋肉の動きをシミュレーションするように訓練した再帰ネットワークは，サルが手を伸ばしている間に記録された運動ニューロンのリングパターンを予測する内部表現を学習した (Sussilo, Churchland, Kaufman & Shenoy, 2015)。 模擬環境でナビゲートするように訓練されたニューラルネットワークは，ラットの内嗅覚皮質で見つかった場所，グリッド，ボーダー，頭の方向を示す細胞と同様のチューニング特性を持つユニットを開発した (Cueva & Wei, 2018; Banino et al. 2018)。 具象化されたモデルは，物理的な体を持つものであれ，仮想的な体を持つものであれ，能動的学習法を用いて有益なトレーニングデータを探し出すことで，体を持たないモデルよりも効率的に学習できる可能性があります (例えば Haber, Mrowca, Fei-Fei & Yamins, 2018)。

5. 認知神経科学におけるニューラルネットワークの今日的挑戦

認知神経科学者にはやるべきことがたくさんあります。 特定の実用的なアプリケーションのために機械学習の文献で紹介されている現在のディープニューラルネットワークのインスタンスは，ほぼ間違いなく，人間の脳で同様の機能がどのように実行されるかを完全に満足させるモデルではないでしょう。 私たちは，DNN の認知をより柔軟で一般的なものにし，学習をより生態学的に妥当なものにし，ニューラルネットワークモデルの予測力を理論的な理解に変える方法を模索しなければなりません。 幸いなことに，これらの目標の多くは，産業界のAI研究者の目標と少なくとも部分的には一致しています。

5.1 より頑健で柔軟な認知と知覚のモデリング

DNN はアーキテクチャ的には深いのですが，概念的には浅いと思われることがあります。 DNN の知能は人間よりも脆弱で，タスクや環境を超えてパフォーマンスを一般化するための因果関係や概念的な理解を深めることができないようです (Marcus, 2017; Lake, Ullman, Tenenbaum & Gershman, 2017)。 視覚的物体認識 DNN は，人間には感知できない微小な画像の摂動を利用して，大幅な分類ミスに陥る可能性があります (Szegedy et al. 2013)。 文章に基づいて質問に答える読解ネットワークも同様に，人間の読者の気を散らさないような小さな文章の変化で騙されることがあります (Jia & Liang, 2017)。. 標準的な DNN は，自分の判断の確信度を見積もるのが苦手で，学習データセットに含まれるものとは異なる統計量の刺激を見せられると，一見無意味な分類を返してしまうことがあります (Nguyen, Yosinski & Clune, 2015; Gharamanhi, 2015)。 強化学習ネットワークは，アーケードビデオゲームのプレイにおいて人間を凌駕することができますが（Mnih et al. 2015），人間が超重要または無関係と考えるような特徴に依存しているようです。

現在の DNN は，生物の脳に比べて計算規模が非常に限られていることを覚えておくと良いでしょう。 最新の DNN は，数十万から数百万のユニットを含んでいるかもしれませんが，これは，一握りの fMRI ボクセル内の皮質ニューロンの数と同じオーダーです[2]。 重要なのは，ニューラルネットワークモデルのユニットは，生物学的なニューロンと同一視することはできないということです。 ニューロンは，構造とダイナミクスがより複雑であり，したがって，より強力な計算能力を持つ可能性があります。 その結果 DNN は，ユニットやニューロンの数から想像されるよりもはるかに大きな因子で不足している可能性があります。 ユニットやニューロンの数を比較しても意味がないのです。

> 脚注2 成功した画像認識ネットワーク AlexNet は約 65 万 8000 ユニットあり（Khrizhevsky, 2012），人間の大脳皮質では 3 mm等方性ボクセルあたり約 63 万個のニューロン密度がありますhttps://cfn.upenn.edu/aguirre/wiki/public:neurons_in_a_voxel

さらに DNN はこれまで，単一のモダリティとタスクの中だけで生活する傾向があり，スケールの面でも不利でした。 よく言えば，特定の感覚処理領域のモデルということになりますが，人間の脳の中のこれらの領域でさえ，他のモダリティ，実行制御，行動との広範な相互作用があります。 上記の「ニューラルネットワークにおける推論」のセクションで説明した例は，コンポジション・アーキテクチャ (Santoro et al., 2017; Xu et al., 2015) や ニューラルネットワークと記号的推論モデルを組み合わせたハイブリッド・ソリューション (Battaglia et al., 2016; Evans & Grefenstette, 2018 など) を含む，可能なソリューションの一部を示唆しています。

認知神経科学は，人間の認知の重要な特徴を捉えるベンチマークタスクを考案するのに適している。 多くの既存の実験パラダイムが理想的である可能性があり，例えば，子供の直感的な物理概念の発達を測定するために使用される「期待の違反」課題は，物理的相互作用に関するモデルの推論を評価するために適用することができます (Piloto, 2018)。 Lake and Baroni (2017) と Marcus (2017) は 機械の認知の奥深さと一般性を評価するために，他の多くのタスクを提案しています。 これらをベンチマークとして定式化することで，コンピュータビジョンで物体認識ベンチマークが行ってきたように，モデリングコミュニティに具体的な目標を与え，進歩を促すことができるでしょう (Kriegeskorte & Douglas 2018)。

5.2 生態学的，及び，生物学的により妥当な方法で学習するネットワークを構築する

教師付き学習も教師なし学習もバックプロパゲーションアルゴリズムに依存していますが，これは従来，生物学的には実現不可能と考えられていました (Glaser, Benjamin, Farhoodi & Kording, 2018)。 最近の理論的な研究では，バックプロパゲーションに似たアルゴリズムを用いた深層学習は，生物学的に実現可能である可能性が示唆されている (Lillicrap et al., 2016; Guerguiev, Lillicrap, and Richards, 2017; Kording & Konig 2001; Scellier & Bengio, 2016; Hinton and McClelland 1988)。 しかし，何百万ものラベル付けされた刺激に対する教師付きトレーニングは，まだ生態学的に非現実的な要件です。 動物の脳は，はるかに多くのデータを効率的に学習することができ，多くの場合，明示的な監督なしで，時には単一の例からでさえも学習することができます。

教師なし，弱い教師あり，ワンショット学習は，機械学習コミュニティの現在の焦点であり (Hassabis et al. 2017) 今後数年間でこれらの分野の進歩が期待できます。 認知神経科学者は，創造的で生物学的に実現可能な解決策のために，脳の理論とデータを利用することができるかもしれません。 神経科学における有力なアイデアの1つは，予測が脳の学習と知覚の方法に根本的に重要であるかもしれないというものです (Hawkins & Blakeslee, 2007)。 予測符号化理論では，知覚の際に，神経活動は主に，予測された感覚情報と実際の感覚情報の違いを符号化すると提案しています(Srinivasan, Laughlin & Dubs, 1982)。 予測符号化の明示的な計算モデルは，これまでに神経回路レベルで定式化されており（Rao & Ballard, 1999），電気生理学的データ (Rao & Ballard, 1999) や fMRI データからも一定の支持を得ている (Muckli et al. 2015)。 機械学習では，予測は，追加の報酬やラベル情報を必要としない，豊富な教師なしのトレーニング信号を提供します。 最新の深層学習フレームワークに実装された予測符号化ネットワークは，自然環境の動画における未来のフレームを予測することができました (Lotter, Kreiman & Cox, 2017; 2018)。 さらに，このネットワークは，その深層において，顔のアイデンティティやポーズなど，動画に描かれたオブジェクトのより高レベルの特性を自発的に発見し (Lotter, Kreiman & Cox, 2017), その個々のユニットは，霊長類の視覚ニューロンの特定の時間的なダイナミクスを再現した (Lotter, Kreiman & Cox, 2018)。 “Curiosity-based” 学習は 動物の学習に動機づけられたもう一つの心躍る方法で，模擬環境に具現化されたネットワークが，学習中にその環境の最も情報量の多い部分を積極的に探し出すものである (Haber, Mrowca, Fei-Fei & Yamins, 2018)。

5.3 暗箱に光をあてる: 予測から説明へ

認知モデルとしてのニューラルネットワークに向けられる批判は，DNN が紛れもなく享受してきた，脳や行動データの予測の成功である (Khaligh-Razavi ら 2014; Cadieu ら 2014; Guclu & van Gerven 2015, Cichy ら 2016; Eickenberg ら, 2017; Kubilius ら, 2016; Wen ら, 2017; 2016; Eickenberg et al., 2017; Kubilius et al., 2016; Wen et al., 2017), 脳や心を説明したり理解したりすることに成功したことにはならない (Kay, 2017)。 あるモデルが，タスクの新しいインスタンスへの一般化を伴って，神経と行動のデータを網羅的に予測していたら，それがある認知機能の正確なモデルであると確信するかもしれません。 しかし，その機能をどのように実行するかをより簡潔に表現できなければ，科学者として十分に満足することはできないでしょう。

以上，ブラックボックスに光を当てるための，In Sillico 電気生理と内部表現の可視化のテクニックを紹介した。 さらに満足度の高いアプローチは，ニューラルネットワークモデルから簡潔な数学的記述に至ることです。 Goncalves と Welchmann (2017) は ステレオ画像のペアからシーン内のオブジェクトの相対的な奥行きを判断するように畳み込みニューラルネットワークを訓練し，古典的な「対応問題」 (2 つの網膜画像内のどの点が，外部のオブジェクト上の同じ点に対応するか) を解決するように要求した。 学習されたネットワークの接続重みの強さと符号を追跡すると，驚くべき計算戦略が発見された。 つまり，ネットワークは，2 つの画像内の点の間の正の一致を見つけるために正の重みを使うのではなく，主に間違った一致を抑制するために負の重みを使っていたのである。 著者らは，この戦略を簡潔な数学的画像補正モデルとしてまとめた。 このモデルは，人間の奥行き知覚におけるいくつかの珍しい，これまで説明されていなかった現象を表示し，少なくとも100年間，計算上のパズルとして立ちはだかっていた対応関係問題の理解を大幅に変えた (Goncalves & Welchmann, 2017)。 ディープニューラルネットワークモデルは，直感的な説明，言葉による理論，簡潔な数学的記述に取って代わるものではありません。 むしろ，複雑な仮説を検証可能にし，言葉で伝えられる理論と神経の実装との間の記述レベルの橋渡しをしてくれます。

6. 結論

深層ニューラルネットワークは，現在，アーティフィシャルインテリジェンスの最先端にあり，行動の多くの側面 (Goncalves & Welchmann, 2017; Kubilius, Bracci & Op de Beeck, 2016; Wallis et al, 2014), および大規模な皮質活動 (Khaligh-Razavi & Kriegeskorte, 2014; Cichy et al., 2016; Cichy et al., 2017; Wen et al., 2017; Guclu & van Gerven 2015; Eickenberg et al., 2017; Greene & Hansen, 2018; Kell et al., 2018; Horiwaka & Kamitani, 2017)。 生物学的に妥当な構成要素と構造を持つこれらのモデルは，脳内で知覚や認知がどのように行われるかについて，エンド・ツー・エンドの明確なモデルに最も近いものです。

したがって 深層学習は，知覚や認知が神経活動からどのように生じるかに関心のあるすべての人に関係があります。 認知神経科学者は，理論的な議論の中でモデリングの文献を検討したり，他の人が作ったモデルをテストしたり，さらには認知や神経科学の理論に基づいて新しいモデルを構築したりと，様々なレベルで深層学習を利用することができます。 逆に，認知神経科学のコミュニティは，より生態学的に実現可能な方法で学習し，より深遠な概念表現を学習し，よりロバストで一般化可能なタスクパフォーマンスを示すDNNを提供することで，エンジニアリングの進歩を促進することができます。 生物学的に妥当なシステムで実世界のタスクを実行できる計算機を機械学習することは，脳からどのようにして知的行動が生じるかを理解する上で大きな役割を果たすだろう。

付録 A1. ディープラーニングを始めるための諸元

A2. 論文

ディープニューラルネットワークモデルの認知神経科学的な応用に関する優れたレビュー記事は数多くある。 神経科学と人工知能の間の補完的な交流に関する最近の概要については，Hassabis et al. ニューラルネットワーク の アーキテクチャ，原理，および脳機能のモデルとしての応用についてのより広範な取り扱いについては，Kriegeskorte (2015) および Kietzmann, McClure & Kriegeskorte (2017) を参照。 感覚神経科学やシステム神経科学におけるニューラルネットワークモデルのレビューについては Yamins & DiCarlo (2016) および Glaser, Benjamin, Farhoodi & Kording (2018) を参照。

A3. チュートリアル，コース，書籍

機械学習とディープニューラルネットワークのわかりやすい入門書としては Geoff Hinton (https://www.coursera.org/learn/neural-networks) と Andrew Ng (https://www.coursera.org/learn/machine-learning)が 現在提供している無料のオンラインコースが貴重である。 Michael Nielsen (http://neuralnetworksanddeeplearning.com/) のオンライン書籍も同様である。 Deep Learning』（Goodfellow, Bengio & Courville, 2016）は，この分野のリーダーたちによって書かれた包括的な書籍である。

A4. ソフトウエア

ディープニューラルネットワークモデルを実装するためのソフトウェアフレームワークは，急速に進化しています。 本稿執筆時点では Python ベースの無償フレームワークが主流となっています。 Tensorflow (Googleが開発)，PyTorch (Facebookが開発)，Caffe (Berkeley大学が開発) などです。 Keras と Sonnet は Tensorflow 上で動作し，多くの高度な機能へのアクセスを容易にする，より高レベルのパッケージです。 MATLAB (MathWorks, Inc.) は，産業界の機械学習チームによるサポートや使用は少ないが，心理学の研究室で人気があるため言及する価値があります。 2017 年以降の MATLAB のバージョンにある Neural Network Toolbox は CPU または GPU 上でのフィードフォワードおよびリカレントモデルの定義と訓練，訓練済みモデルのロード，他のフレームワーク (現在は Keras とCaffe がサポート) で定義したモデルのインポート，訓練済みネットワーク内のユニットの視覚化などのすべてのコア深層学習操作を可能です。

文献

- Agrawal, P., Girshick, R., & Malik, J. (2014). Analyzing the performance of multilayer neural networks for object recognition. In Proceedings of the European Conference on Computer Vision (pp. 329344).
- Banino, A., Barry, C., Uria, B., Blundell, C., Lillicrap, T., Mirowski, P., … & Wayne, G. (2018). Vector-based navigation using grid-like representations in artificial agents. Nature, 557(7705), 429.
- Battaglia, P., Pascanu, R., Lai, M., & Rezende, D. J. (2016). Interaction networks for learning about objects, relations and physics. In Advances in Neural Information Processing Systems (pp. 4502-4510).
- Cadieu, C. F., Hong, H., Yamins, D. L., Pinto, N., Ardila, D., Solomon, E. A., … & DiCarlo, J. J. (2014). Deep neural networks rival the representation of primate IT cortex for core visual object recognition. PLoS Computational Biology, 10(12), e1003963.
- Carandini, M., & Heeger, D. J. (2012). Normalization as a canonical neural computation. Nature Reviews Neuroscience, 13(1), 51.
- Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. Scientific Reports, 6, 27755.
- 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, 346358.
- Cueva, C. J., & Wei, X. X. (2018). Emergence of grid-like representations by training recurrent neural networks to perform spatial localization. arXiv preprint. arXiv:1803.07770
- Devereux, B. J., Clarke, A. D., & 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, 1.
- Diedrichsen, J. (2018). Representational models and the feature fallacy. To appear in M. Gazzaniga (Ed.), The Cognitive Neurosciences (6th Edition). Boston: MIT Press.
- Diedrichsen, J., & Kriegeskorte, N. (2017). Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. PLoS Computational Biology, 13(4), e1005508.
- Diedrichsen, J., Ridgway, G. R., Friston, K. J., & Wiestler, T. (2011). Comparing the similarity and spatial structure of neural representations: a pattern-component model. NeuroImage, 55(4), 1665-1678.
- Dumoulin, S. O., & Wandell, B. A. (2008). Population receptive field estimates in human visual cortex. NeuroImage, 39(2), 647-660.
- 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.
- Evans, R., & Grefenstette, E. (2018). Learning explanatory rules from noisy data. Journal of Artificial Intelligence Research, 61, 1-64.
- Geirhos, R., Temme, C. R. M., Rauber, J., Schuett, H. H., Bethge, M., & Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. arXiv preprint. arXiv:1808.08750
- Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. Science, 349(6245), 273-278. Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. Nature, 521(7553), 452.
- Glaser, J. I., Benjamin, A. S., Farhoodi, R., & Kording, K. P. (2018). The roles of supervised machine learning in systems neuroscience. arXiv preprint. arXiv:1805.08239
- Gonc¸alves, N. R., & Welchman, A. E. (2017). ”What not” detectors help the brain see in depth. Current Biology, 27(10), 1403-1412.
- Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). Deep learning (Vol. 1). Cambridge: MIT press.
- Greene, M. R., & Hansen, B. C. (2018). Shared spatiotemporal category representations in biological and artificial deep neural networks. PLoS Computational Biology, 14(7), e1006327.
- G¨uc¸l¨u, U., & van Gerven, M. A. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. Journal of Neuroscience, 35(27), 10005-10014.
- Guerguiev, J., Lillicrap, T. P., & Richards, B. A. (2017). Towards deep learning with segregated dendrites. eLife, 6, e22901.
- Haber, N., Mrowca, D., Fei-Fei, L., & Yamins, D. L. (2018). Emergence of structured behaviors from curiosity-based intrinsic motivation. arXiv preprint. arXiv:1802.07461
- Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. Neuron, 95(2), 245-258.
- Hawkins, J., & Blakeslee, S. (2007). On intelligence: How a new understanding of the brain will lead to the creation of truly intelligent machines. Macmillan.
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 770-778).
- Hinton, G. E., & McClelland, J. L. (1988). Learning representations by recirculation. In Advances in Neural Information Processing Systems (pp. 358-366).
- Horikawa, T., & Kamitani, Y. (2017). Generic decoding of seen and imagined objects using hierarchical visual features. Nature Communications, 8, 15037.

- Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. Neural Networks, 4(2), 251-257.
- Huys, Q.J., Maia, T.V., & Frank, M.J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. Nature Neuroscience, 19(3), 404.
- Jaderberg, M., Czarnecki, W. M., Dunning, I., Marris, L., Lever, G., Castaneda, A. G., … & Sonnerat, N. (2018). Human-level performance in first-person multiplayer games with population-based deep reinforcement learning. arXiv preprint. arXiv:1807.01281
- Jia, R., & Liang, P. (2017). Adversarial Examples for Evaluating Reading Comprehension Systems. arXiv preprint. arXiv:1707.07328
- Kansky, K., Silver, T., M´ely, D. A., Eldawy, M., L´azaro-Gredilla, M., Lou, X., … & George, D. (2017). Schema networks: Zero-shot transfer with a generative causal model of intuitive physics. arXiv preprint. arXiv:1706.04317
- Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. Nature, 452(7185), 352.
- Kay, K. N. (2017). Principles for models of neural information processing. NeuroImage, 180, 101-109.
- Kell, A. J., Yamins, D. L., 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(3), 630-644.
- Kording, K. P., & K¨ . (2001). Supervised and unsupervised learning with two sites of synaptic integra
- onig, P tion. Journal of Computational Neuroscience, 11(3), 207-215. Khaligh-Razavi, S. M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. PLoS Computational Biology, 10(11), e1003915.
- Khaligh-Razavi, S. M., Henriksson, L., Kay, K., & Kriegeskorte, N. (2017). Fixed versus mixed RSA: Explaining visual representations by fixed and mixed feature sets from shallow and deep computational models. Journal of Mathematical Psychology, 76, 184-197.
- Kietzmann, T. C., McClure, P., & Kriegeskorte, N. (2018). Deep neural networks in computational neuro science. In Oxford Research Encyclopedia of Neuroscience, Oxford University Press.
- Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. Frontiers in Systems Neuroscience, 2, 4.
- Kriegeskorte (2015) Deep neural networks: A new framework for modeling biological vision and brain information processing. Annual Review of Vision Science, 1, 417-446.
- Kriegeskorte, N., & Diedrichsen, J. (2016). Inferring brain-computational mechanisms with models of activity measurements. Philosophical Transactions of the Royal Society B, 371(1705), 20160278.
- Kriegeskorte, N., & Douglas, P. K. (2018). Cognitive computational neuroscience. Nature Neuroscience, 21, 11481160.
- Kriegeskorte, N., & Mok, R. M. (2017). Building machines that adapt and compute like brains. Behavioral and Brain Sciences, 40.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems (pp. 1097-1105).
- 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(4), e1004896.
- Kumaran, D., Hassabis, D., & McClelland, J. L. (2016). What learning systems do intelligent agents need? Complementary learning systems theory updated. Trends in Cognitive Sciences, 20(7), 512-534.
- 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.
- Lake, B. M., & Baroni, M. (2017). Still not systematic after all these years: On the compositional skills of sequence-to-sequence recurrent networks. arXiv preprint arXiv:1711.00350
- LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. In Proceedings of the IEEE, 86(11), 22782324.
- Lillicrap, T. P., Cownden, D., Tweed, D. B., & Akerman, C. J. (2016). Random synaptic feedback weights support error backpropagation for deep learning. Nature Communications, 7, 13276.
- Lotter, W., Kreiman, G., & Cox, D. (2016). Deep predictive coding networks for video prediction and unsupervised learning. arXiv preprint. arXiv:1605.08104
- Lotter, W., Kreiman, G., & Cox, D. (2018). A neural network trained to predict future video frames mimics critical properties of biological neuronal responses and perception. arXiv preprint. arXiv:1805.10734
- Marblestone, A. H., Wayne, G., & Kording, K. P. (2016). Toward an integration of deep learning and neuro science. Frontiers in Computational Neuroscience, 10, 94.
- Marcus, G. (2018). Deep learning: A critical appraisal. arXiv preprint. arXiv:1801.00631
- McCulloch, W. & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. Bulletin of Mathematical Biophysics, 5, 115133.
- McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. Psychological Review, 102, 419457.
- Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K. M., Malave, V. L., Mason, R. A., & Just, M.
- A. (2008). Predicting human brain activity associated with the meanings of nouns. Science, 320(5880), 1191-1195. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., … & Petersen, S. (2015). Human-level control through deep reinforcement learning. Nature, 518(7540), 529. Morcos, A. S., Barrett, D. G., Rabinowitz, N. C., & Botvinick, M. (2018). On the importance of single directions for generalization. arXiv preprint. arXiv:1803.06959
- Mordvintsev, Olah & Tyka (2015) Inceptionism: Going deeper into neural networks. Google technical blog post, retrieved from: https://ai.googleblog.com/2015/06/inceptionism-going-deeper -into-neural.html
- Muckli, L., De Martino, F., Vizioli, L., Petro, L. S., Smith, F. W., Ugurbil, K., … & Yacoub, E. (2015). Contextual feedback to superficial layers of V1. Current Biology, 25(20), 2690-2695.
- Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fMRI. NeuroImage, 56(2), 400-410.
- Nayebi, A., Bear, D., Kubilius, J., Kar, K., Ganguli, S., Sussillo, D., … & Yamins, D. L. (2018). Task-driven convolutional recurrent models of the visual system. arXiv preprint. arXiv:1807.00053

- Nguyen, A., Yosinski, J., & Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 427-436).
- Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., & Kriegeskorte, N. (2014). A toolbox for representational similarity analysis. PLoS Computational Biology, 10(4), e1003553.
- Olah (2017). Feature visualization. Retrieved from: https://distill.pub/2017/feature-visualization/
- Olah (2018). The building blocks of interpretability. Retrieved from: https://distill.pub/2018/building-blocks/
- Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. International Journal of Computer Vision, 42(3), 145-175.
- O'Reilly, R. C., Wyatte, D., Herd, S., Mingus, B., & Jilk, D. J. (2013). Recurrent processing during object recognition. Frontiers in Psychology, 4, 124.
- Paul, A., & Venkatasubramanian, S. (2014). Why does deep learning work?: A perspective from group theory. arXiv preprint. arXiv:1412.6621
- Piloto, L., Weinstein, A., Ahuja, A., Mirza, M., Wayne, G., Amos, D., … & Botvinick, M. (2018). Probing physics knowledge using tools from developmental psychology. arXiv preprint. arXiv: 1804.01128
- Poeppel, D. (2012). The maps problem and the mapping problem: two challenges for a cognitive neuroscience of speech and language. Cognitive Neuropsychology, 29(1-2), 34-55.
- Radford, A., Jozefowicz, R., & Sutskever, I. (2017). Learning to generate reviews and discovering sentiment. arXiv preprint. arXiv:1704.01444
- Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., & DiCarlo, J. J. (2018). Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. Journal of Neuroscience, 38(33), 7255-7269.
- Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nature Neuroscience, 2(1), 79.
- Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. Nature Neuroscience, 2(11), 1019.
- Rumelhart, D.E. & McClelland, J.L. (1986). Parallel distributed processing: explorations in the microstructure of cognition. MIT Press:Cambridge, MA.
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., … & Berg, A. C. (2015). Imagenet large scale visual recognition challenge. International Journal of Computer Vision, 115(3), 211-252.
- Santoro, A., Raposo, D., Barrett, D. G., Malinowski, M., Pascanu, R., Battaglia, P., & Lillicrap, T. (2017). A simple neural network module for relational reasoning. In Advances in Neural Information Processing Systems (pp. 4967-4976).
- Scellier, B., & Bengio, Y. (2016). Equilibrium propagation: Bridging the gap between energy-based models and backpropagation. arXiv preprint. arXiv:1602.05179
- Hochreiter, S. & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 17351780.
- Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. arXiv preprint. arXiv:1409.1556
- Spoerer, C. J., McClure, P., & Kriegeskorte, N. (2017). Recurrent convolutional neural networks: A better model of biological object recognition. Frontiers in Psychology, 8, 1551.
- Srinivasan, M. V., Laughlin, S. B., & Dubs, A. (1982). Predictive coding: A fresh view of inhibition in the retina. Proceedings of the Royal Society of London B, 216(1205), 427-459.
- Sussillo, D., Churchland, M. M., Kaufman, M. T., & Shenoy, K. V. (2015). A neural network that finds a naturalistic solution for the production of muscle activity. Nature Neuroscience, 18(7), 1025.
- Sutton, R. S., & Barto, A. G. (1998). Introduction to reinforcement learning (Vol. 135). Cambridge: MIT press.
- Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. arXiv preprint. arXiv:1312.6199
- Taigman, Y., Yang, M., Ranzato, M. A., & Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1701-1708).
- Wallis, T. S., Funke, C. M., Ecker, A. S., Gatys, L. A., Wichmann, F. A., & Bethge, M. (2017). A parametric texture model based on deep convolutional features closely matches texture appearance for humans. Journal of Vision, 17(12), 5-5.
- Wang, Z., & Simoncelli, E. P. (2008). Maximum differentiation (MAD) competition: A methodology for comparing computational models of perceptual quantities. Journal of Vision, 8(12), 8-8.
- Wen, H., Shi, J., Zhang, Y., Lu, K. H., Cao, J., & Liu, Z. (2017). Neural encoding and decoding with deep learning for dynamic natural vision. Cerebral Cortex, 1-25.
- Wenliang, L., & Seitz, A. R. (2018). Deep neural networks for modeling visual perceptual learning. Journal of Neuroscience, 38(27), 6028-6044.
- Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., … & Klingner, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint. arXiv:1609.08144
- Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., … & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In International Conference on Machine Learning (pp. 2048-2057).
- Yamins, D. L. & DiCarlo, J. J. (2016) Using goal-driven deep learning models to understand sensory cortex. Nature Neuroscience, 19(3), 356.
- Yang, G. R., Song, H. F., Newsome, W. T., & Wang, X. J. (2017). Clustering and compositionality of task representations in a neural network trained to perform many cognitive tasks. bioRxiv preprint. doi:10.1101/183632
- Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., & Lipson, H. (2015). Understanding neural networks through deep visualization. arXiv preprint. arXiv:1506.06579
- Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2014). Object detectors emerge in deep scene CNNs. arXiv preprint. arXiv:1412.6856
- Zhuang, C., Wang, Y., Yamins, D. L., & Hu, X. (2017). Deep learning predicts correlation between a functional signature of higher visual areas and sparse firing of neurons. Frontiers in Computational Neuroscience, 11, 100.