

令和7年度 修士論文

# 食べ物の好みの脳計算過程：深層学習による 表現

青木悠飛

一橋大学大学院ソーシャル・データサイエンス研究科

2025年1月 提出



## 概要

日常の食事選択は、食品に付与する主観的価値に基づいて行われる。しかし、この主観的価値が脳内どのように計算されるかについては、ほとんど解明されていない。本研究では、深層ニューラルネットワーク(DNN)モデルを用いて、食品評価の神経計算過程の解明を目指した。199名の参加者による896枚の食品画像の評価データを用いてDCNNを訓練、視覚-言語モデル(CLIP)の埋め込みを入力とする回帰モデルを学習した結果、DNNは主観的価値を有意に予測し、CLIPが最も高い精度( $r = 0.78$ )を示した。DNN層の活性化パターンの分析から、高次属性(主観的価値、健康性、美味しさ)は後期層で強く表現される一方、低次の色情報は全層にわたって一貫して符号化されることが明らかになった。31名の参加者の機能的磁気共鳴画像法(fMRI)データを用いた表現類似性解析(RSA)では、一次視覚野(V1)においてDNNモデルが約65-69%を説明したが、高次視覚野や価値関連領域(vmPFC)では説明率が低下した。エンコーディング解析では、ConvNeXt(視覚モデル)の階層構造が視覚処理階層と対応し、初期層が一次視覚野、後期層が高次視覚野および価値関連領域と関連することが示された。一方、CLIP(視覚-言語モデル)は中間層で言語・情動・報酬関連領域(IFG・PCC・腹側線条体・島皮質・扁桃体)との対応を示し、食品の主観的価値計算では報酬だけでなく言語・情動などの情報処理が関与していることを示唆する。これらの知見は、視覚情報が階層的に処理され、価値計算と統合される脳内メカニズムを明らかにするものである。

# 目次

<b>1 はじめに</b>	6
<b>2 関連研究</b>	8
2.1 主観的価値 .....	9
2.2 深層ニューラルネットワーク(DNN) .....	10
2.3 DNN と脳の比較 .....	10
<b>3 目的</b>	10
<b>4 方法</b>	11
4.1 オンライン調査 .....	12
4.2 fMRI 実験 .....	12
4.2.1 参加者.....	12
4.2.2 刺激.....	13
4.2.3 fMRI 実験手続き.....	13
4.2.4 使用した DNN モデル.....	13
4.3 DNN モデルの訓練 .....	14
4.3.1 交差検証とデータ拡張.....	14
4.3.2 モデル訓練.....	15
4.4 DNN モデル評価 .....	15
4.4.1 事前学習モデルの評価.....	15
4.4.2 フайнチューニングモデルの評価.....	15
4.5 DNN のデコーディング解析 .....	15
4.6 fMRI データ収集 .....	15
4.7 fMRI データ前処理 .....	16
4.8 fMRI データ解析 .....	16
4.8.1 GLM 分析.....	16
4.8.2 集団レベル解析.....	16
4.8.3 表現類似性解析(RSA).....	17
4.8.4 エンコーディング解析.....	17
<b>5 結果</b>	19
5.1 DNN モデルによる主観的価値予測 .....	20
5.2 DNN の層別情報表現 .....	20
5.3 fMRI 解析結果 .....	21
5.3.1 主観的価値に関連する脳活動.....	21
5.3.2 ROI 解析.....	22
5.3.3 DNN 比較分析.....	22

<b>6 考察</b>	26
<b>7 付録</b>	29
A fMRI の仕組み .....	30
A.a 機能的磁気共鳴画像法(functional Magnetic Resonance Imaging; fMRI).....	30
A.b MRI 信号の仕組み:核磁気共鳴と緩和現象.....	30
A.c BOLD 信号の生成メカニズム.....	30
A.d 血行動態応答と時間特性.....	31
B ニューラルネットワークモデルの構造 .....	31
C ディープニューラルネットワークモデル(DNN) .....	32
C.a CNN アーキテクチャ.....	32
C.b Transformer アーキテクチャ.....	33
D 各層の活性化パターン分析 .....	35
D.a PCA による次元削減.....	35
D.b Ridge 回帰による画像特徴予測.....	35
E fMRI における一般化線型モデル(GLM)分析 .....	36
F 集団解析におけるランダム効果モデル .....	37
F.a ランダム効果解析(Random Effects Analysis).....	37
F.b 要約統計量を使ったランダム効果解析.....	37
G 多重比較補正 .....	38
G.a ファミリー・ワイズ・エラー率(Family-Wise Error Rate; FWE)補正.....	38
H 表現類似性解析(RSA) .....	39
H.a 表現非類似度行列(RDM).....	39
H.b 二重中心化(Double centering).....	39
H.c ノイズ上限値(Noise Ceiling).....	40
I 3 レベル階層的 PCA 分析 .....	40
I.a 層グループの定義.....	40
I.b 3 レベル階層的 PCA.....	40
I.c 直交化の意義.....	41
I.d GLM への適用.....	41
J 補足:事前学習済み ConvNeXt の層別情報表現 .....	42
K 補足:DNN 比較分析 ROI 効果量 .....	42
<b>引用文献</b>	43

# 図表目次

図 1 主観的価値による上位(上)および下位(下)の食品画像の例。	12
図 2 食品評価課題における 1 試行のタイムライン。各試行において、参加者は 1 つの食品に対する「どのぐらい欲しいか」(すなわち主観的価値)を報告した。この評価は支払意思額(Willingness to Pay; WTP)に類似した概念であり、食品に対する動機づけの強さを反映している。評価フェーズでは、キーと選択肢の順序が試行間でランダム化されている。	13
表 1 ConvNeXt および CLIP の層グループ分類	18
図 3 DNN モデルによる主観的価値予測の精度(予測と実際の評価との相関係数)。青:事前学習のみ、緑:ファインチューニング後。CLIP は事前学習済みモデルのみ使用。	20
図 4 DNN モデルの各層における情報表現。左:ファインチューニング後の ConvNeXt、右:事前学習済み CLIP。各線は、各属性の予測と実際の評価との相関係数を示す。	21
図 5 主観的価値(Image × Value)に関する脳活動。クラスターレベル FWE 補正( $p < 0.05$ 、クラスター形成閾値 $p < 0.001$ uncorrected)。カラーバーは T 値を示す。色付きの輪郭線は関心領域(ROI)を示す: vmPFC(青)、OFC(緑)、左線条体(シアン)、右線条体(マゼンタ)。	21
図 6 ROI 解析における主観的価値の効果量( $\beta$ )。すべての ROI で有意な効果が認められた( $p < 0.05$ )。vmPFC:腹内側前頭前皮質、mOFC:内側眼窓前頭皮質。	22
図 7 ROI-RSA 解析結果(Double-centering 適用) a. 各 ROI におけるモデル別説明率の比較。b. 右は視覚野(V1、初期視覚野、LOC、紡錘状回、IT)における平均説明率。左は全 31 ROI における平均説明率。エラーバーは標準誤差(SEM)。*** $p < 0.001$ 、* $p < 0.05$ (対応のある t 検定)。	24
図 8 DNN モデルの層グループ別脳活動(FWE 補正 $p < 0.05$ )。a. ConvNeXt モデル。b. CLIP モデル。各行は異なる層グループ(初期層+共有、中間層+共有、後期層+共有、最終層+共有)と脳活動の対応を示す。	25
図 9 DNN モデルの層グループ別 ROI 効果量。各層(初期・中間・後期・最終)に関する共有成分を加えた効果量を示す。エラーバーは被験者間の標準誤差(SEM)。は有意な効果(SVC FWE $p < 0.05$ )。追加の効果量図は補足 K を参照。	26
図 A ニューラルネットワークの基本構造。入力層、中間層(隠れ層)、出力層から構成される。各ニューロンは、前の層からの入力信号に対して重みを適用し、バイアス項を加えた後、活性化関数を通じて出力信号を生成する。	32
図 B ニューラルネットワークにおける代表的な活性化関数の例。左からシグモイド関数、ReLU 関数、ソフトマックス関数。各関数は、ニューロンの出力信号を生成するために使用される。	32
図 C (Simonyan & Zisserman, 2015); から作成。畳み込みニューラルネットワーク(CNN)の基本構造の例(VGG16)。畳み込み層、プーリング層、全結合層から構成される。畳み込み層は入力画像に対して畳み込みフィルターを適用し、特徴マップを生成する。プーリング層は特徴マップの空間的次元を削減し、計算量を軽減するとともに、位置不变性を持たせる役割を果たす。全結合層は最終的な分類結果を出力するために、抽出された特徴を用いる。	33
図 D 畳み込み操作の例。入力画像に対して畳み込みフィルターを適用し、特徴マップを生成する。フィルターは画像の局所的なパターンを検出するために使用される。	33
図 E トランスフォーマーの基本構造。エンコーダーとデコーダーの 2 つの主要なコンポーネントで構成されており、エンコーダーは入力シーケンスを処理し、デコーダーは出力シーケンスを生成する。自己注意メカニズムを用いて入力シーケンス内の異なる位置の情報を動的に重み付けする (Vaswani et al., 2017, p.3, Figure 1);。	34

図 F 事前学習済み ConvNeXt-Base の各層における情報表現。各線は、各属性(主観的価値、健康度、栄養価、色)の予測と実際の評価との相関係数を示す。ファインチューニング後のモデル(図 4)と比較して、全体的に同様の傾向が見られるが、主観的価値の予測精度は後期層でも約 0.55 程度に留まる。 42

図 G 各層固有成分の ROI 効果量。\*は有意(SVC FWE 補正  $p < 0.05$ )。 43

図 H 各層固有成分・共有成分・Global 成分の ROI 効果量。共有(初-中)は初期-中間層間、共有(中-後)は中間-後期層間、共有(後-最)は後期-最終層間の共有成分。\*は有意(SVC FWE 補正  $p < 0.05$ )。 43

図 I 各層+共有+Global 成分の ROI 効果量。各層固有成分に関連する共有成分および Global 成分を加えた効果量。\*は有意(SVC FWE 補正  $p < 0.05$ )。 43

# 1 はじめに

摂食行動は人間の生存に不可欠であり、「何を食べるか」の選択は日常生活における中心的な関心事である。こうした決定は、利用可能な選択肢の主観的価値を比較することで下されると考えられている(Rangel et al., 2008)。偏った食物選好は肥満や摂食障害と関連しており(Foerde et al., 2015; Spinelli & Monteleone, 2021)、2021年には、食生活リスクに起因する死亡は約800万人に上る(Vaduganathan et al., 2022)。これらの知見から、食の意思決定の偏りによって健康被害が生じ、死亡に繋がっている可能性がある。そのため、主観的価値の計算方法を解明することが不適切な食事選択の解決策となり、公衆衛生の向上に寄与する可能性がある。

食品の主観的価値は、栄養成分、味、健康への影響、見た目、価格などさまざまな要因によって影響を受ける(Hare et al., 2009; Motoki & Suzuki, 2020; Suzuki et al., 2017; Tang et al., 2014)。最終的な統合された主観的価値は腹内側前頭前野(vmPFC)で表出されると考えられている(Chib et al., 2009)。さらに、脳内における主観的価値の計算は抽象化の度合いに応じて階層的かつ分散的に情報を処理するという仮説がある。しかし、感覚表象から主観的価値を計算する脳の中間過程は、それが階層的であるか否かを含め、依然としてほとんど解明されていない。

計算神経科学は、脳のメカニズムに関する仮説を数学的モデルやアルゴリズムとして定式化し、脳を情報処理システムとして理解しようとするアプローチである。これらのモデルは、行動データやfMRIなどの脳活動データと比較を行うことで、複雑な脳の計算過程を解明することにつながる。認知科学における従来のトップダウンアプローチでは、理論をモデル化し実験で検証することで単一ニューロンや小規模な固定回路の解明してきたが、可塑性を持つ神経回路の詳細な計算過程をモデル化するには、複雑なモデルが必要である(Kriegeskorte & Douglas, 2018; Richards et al., 2019)。このため近年、人工ニューラルネットワーク(ANN)を用いたボトムアップアプローチが注目を集めている。ANNは脳の構造と機能に着想を得ており、層状に情報を処理する相互接続されたノード(ニューロン)で構成される。大規模データセットでANNを学習させることで、人間レベルの認知タスクを遂行できるようになり、脳の計算過程をモデル化する有望な手段となっている。

最近の研究では、深層畳み込みニューラルネットワーク(DCNN)や大規模言語モデル(LLM)の内部表現を人間の行動や脳活動と比較することで、知覚・認知・評価のメカニズムの解明が始まっている(Doerig et al., 2025; Iigaya et al., 2021; 2023; Kriegeskorte, 2015; Lahner et al., 2024; Yamins et al., 2014)。DCNNは視覚処理の階層構造を模倣し、画像認識において人間レベルの性能を達成するため、神経処理の有望な計算モデルとして機能する。例えば、飯ヶ谷らは美術作品の主観的価値を予測するDCNNを構築し、低次画像特徴が初期層で捕捉される一方、高次抽象特徴が後続層で出現することを示した(Iigaya et al., 2021)。一方、LLMは人間の脳とは異なる複雑な構造を持つと考えられ、LLMと脳の内部構造を厳密に比較する方法はほとんどない。それでも、DoerigらはLLMの最終埋め込みを脳と

比較し、LLM が高次視覚野に類似していることを示した(Doerig et al., 2025)。これらの研究は、深層ニューラルネットワーク(DNN:DCNN や LLM を含む)と脳が計算プロセスにおいて比較可能かつ部分的に類似していることを示している。

本研究では、DNN のモデリングと解析を通じて、食品画像から主観的価値(好み)の計算プロセスを解明することを目的とする。まず、食品画像から主観的価値を予測する DNN を訓練し、層ごとの活性化パターンを分析することで、どの段階(層)でどの種類の情報(低次画像特徴、栄養成分、美味しさ、健康性)が表現されているかを検証する。次に、fMRI データと DNN の最終または中間表現を比較し、脳と DNN が類似する領域を分析する。このアプローチにより、脳が視覚情報を処理・統合して最終的な主観的価値の計算過程に関する知見が得られる。

## 2 関連研究

### 2.1 主観的価値

主観的価値とは、意思決定の際に用いられる「共通通貨」である。主観的価値は、経済学の期待効用理論の効用から生まれた、神経科学、心理学、経済学の共通の概念的枠組みである (Glimcher & Rustichini, 2004)。Levy らは、主観的価値が異なる報酬カテゴリーにわたって内側前頭前皮質 (MPFC) で表現されることを示した (Levy et al., 2011)。さらに、Gross らは、前頭前野の価値信号が異なる報酬カテゴリー (活動とスナック菓子) にわたって個人の嗜好を予測できることを示した (Gross et al., 2014)。このため、主観的価値という共通価値を表現する脳領域は共通であると考えられている。

具体的な神経プロセスとして、主観的価値は脳内の前頭前野、線条体などの複数の領域にわたって表現され、最終的に vmPFC で主観的価値が表現される。Hare らは、食品画像と fMRI を使って、主観的価値の予測は内側眼窩前頭皮質、実際の主観的価値は中心眼窩前頭皮質の活動と相関し、主観的価値の予測誤差は腹側線条体の活動と相関していることを示した (Hare et al., 2008)。Samejima らは、サルの線条体における行動特異的な報酬価値の表現が行動選択を導くことを示唆した (Samejima et al., 2005)。主観的価値の最終的な計算表現は明らかになっている一方で、感覚情報から主観的価値を計算する中間過程は依然として不明な部分が多い。

食品の主観的価値は、味や見た目以外のさまざまな要因も考慮して計算されることがわかっている。Suzuki らは、食品の主観的価値が栄養属性 (タンパク質、脂質、炭水化物、ビタミン)に基づいて計算されることを示した (Suzuki et al., 2017)。Motoki らは、食品の主観的価値がラベル、ブランド、価格などの外的要因によっても影響を受けることを示した (Motoki & Suzuki, 2020)。これらの研究は、食品の主観的価値が多次元的な情報から計算されることを示している。しかし、視覚情報からこれらの属性を抽出し、最終的な主観的価値を計算する中間過程は依然として不明である。

主観的価値の計算を理解するために、学習理論やプロスペクト理論などの理論を計算モデルに落とし込み、計算過程を比較する試みがなされている。例えば、強化学習モデルを用いて線条体、腹内側前頭前皮質 (vmPFC) の活動をモデリングし、線条体や vmPFC が強化学習モデルと近い動きをしていることを明らかにした (Hampton et al., 2006; Rutledge et al., 2014)。Suzuki らは、意思決定の線形モデルを構築し、fMRI データを比較し、意思決定の要因が、自身の選好と過去の選択に関する情報、他者の選択肢の評価であり、それらが異なる脳領域で計算していることを示した (Suzuki et al., 2015)。Yamins らは、DCNN を用いて視覚野の神経応答を予測し、DCNN の高次層が下側頭葉皮質の神経応答を予測することを示した (Yamins et al., 2014)。これらの知見は、計算モデルを用いて主観的価値の計算過程を理解する有望なアプローチであることを示している。

## 2.2 深層ニューラルネットワーク(DNN)

DNN は、脳の構造と機能に着想を得ており、層状に情報を処理する相互接続されたノード(ニューロン)で構成される。その中でも、DCNN は視覚処理の階層構造を模倣し、画像認識において人間レベルの性能を達成するため、視覚情報処理の有望な計算モデルとして機能する (Kriegeskorte, 2015; Yamins et al., 2014)。DCNN は、畳み込み層、プーリング層、全結合層などの複数の層で構成され、各層は前の層からの入力を受け取り、特徴マップを生成する。初期層はエッジやテクスチャなどの低次特徴を抽出し、後続層はオブジェクトの形状やカテゴリなどの高次特徴を抽出する。DCNN は、ImageNet などの大規模データセットで学習され、人間レベルの画像認識性能を達成している (Deng et al., 2009; Krizhevsky et al., 2012)。そのため、DCNN は視覚情報処理の計算モデルとして利用可能であり、脳の視覚野の神経応答と比較することで、視覚情報処理のメカニズムの解明に寄与する可能性がある。

LLM は、自然言語処理タスクにおいて人間レベルの性能を示す。多くの LLM は、トランスフォーマーアーキテクチャに基づいており、自己注意メカニズムを用いて入力シーケンス内の異なる位置の情報を動的に重み付けする。LLM は、大規模なテキストデータセットで学習され、文の生成、質問応答、翻訳などのタスクで優れた性能を示している (Brown et al., 2020; OpenAI, 2024; Vaswani et al., 2017)。LLM は、文脈情報を考慮した単語の意味表現を学習し、文の意味理解に寄与している。そのため、LLM を脳の言語処理に関する領域と比較することで、言語処理のメカニズムの解明に寄与する可能性がある (Caucheteux et al., 2023; Schrimpf et al., 2021)。

## 2.3 DNN と脳の比較

DNN と脳を比較することで、モデルのどの部分が脳領域に類似しているのか調べる試みがある。Yamins らは、DCNN の高次層が下側頭葉皮質の神経応答を予測することを示した (Yamins et al., 2014)。DCNN の層ごとの特徴表現は、脳の視覚野の階層的な情報処理と類似していることが示されている。Iigaya らは、DCNN を用いて美術作品の主観的価値を予測し、低次画像特徴が初期層で捕捉される一方、高次抽象特徴が後期層で出現することを示した (Iigaya et al., 2021)。Doerig らは、LLM の最終埋め込みを脳と比較し、LLM が高次視覚野に類似していることを示した (Doerig et al., 2025)。これらの研究は、脳とモデルの構造を比較することで、脳のどの領域でどの情報が使われているか明らかにすることを示唆している。そのため、主観的価値の計算においても、LLM を含めた DNN と脳が比較することで、主観的価値の計算過程に関する新たな知見を提供する可能性がある。

### 3 目的

本研究では、DNN のモデリングと解析を通じて、食品の主観的価値(好み)の計算プロセスを解明することを目的とする。具体的には、以下の研究目的で検証を行う。

- **研究目的 1:DNN は食品画像から主観的価値を予測できるか**

食品画像の視覚特徴から主観的価値を予測する DNN モデルを構築し、その予測精度を評価する。特に、視覚情報のみを扱う DCNN(ConvNeXt、ResNet、VGG)と、視覚-言語情報を統合する CLIP を比較することで、言語的・意味的情報の寄与を検討する。

- **研究目的 2:DNN の各層でどのような情報が表現されているか**

DNN の層ごとの活性化パターンを分析し、異なる種類の情報(低次画像特徴、栄養属性、美味しさ、健康性、主観的価値)がどの層で表現されているかを検証する。先行研究(Iigaya et al., 2021)に基づき、低次特徴は初期層、高次属性は後期層で表現されるという階層的処理仮説を検証する。

- **研究目的 3:DNN は脳の表現構造をどの程度説明できるか**

表現類似性解析(RSA)を用いて DNN と脳の表現類似度を定量化し、ノイズ上限値を基準としてモデルの説明力を評価する。ノイズ上限値は被験者間の一致度から推定される理論的上限であり、これに対するモデルの到達度を検証することで、DNN が脳の表現をどの程度捉えているかを明らかにする。

- **研究目的 4:DNN の層構造は脳の情報処理階層と対応するか**

fMRI データと DNN の層別活性化パターンを比較し、以下の仮説を検証する：

- **仮説 4a:** 視覚モデル(DCNN)の初期層は一次視覚野、後期層は高次視覚野および価値関連領域(vmPFC)と対応する

- **仮説 4b:** 視覚-言語モデル(CLIP)は、意味・言語・価値関連領域と対応する

#### 期待される貢献

本研究は、(1)DNN モデルを使って、画像から食品の主観的価値はどれだけ予測可能か、(2)視覚-言語統合が食品の価値評価に与える影響を神経科学的に検証すること、(3)ノイズ上限値を用いて DNN の脳表現をどれくらい説明できるかを定量化すること、(4)新たな手法である階層的 PCA を用いた GLM で、DNN と脳を比較することで、食品の主観的価値の詳細な中間処理過程を明らかにすることを目的とする。

## 4 方法

本研究では、オンライン調査で収集した食品画像に対する主観的価値評価データを使用し、DNN の学習と解析を行った。さらに、fMRI データを用いて、DNN の最終および中間表現と脳活動パターンの比較を行った。

### 4.1 オンライン調査

オンライン調査には、20 歳から 72 歳までの健康な成人男女 200 名(平均年齢 39.08 歳)が参加した。参加者は、Food-Pics (Blechert et al., 2014; 2019) から 896 枚の食品画像に対して、「好み」(主観的価値)、「美味しさ」「健康性」を 8 段階リッカート尺度(1 = 強く不同意、8 = 強く同意)で評価した。データ不一致により 1 名を除外した後、残りの 199 名の評価を各画像ごとに平均化した(図 1 は主観的価値による上位および下位の画像の例を示す)。なお、主観的価値は美味しさと正の相関( $r = 0.89$ )、健康性とは弱い負の相関( $r = -0.28$ )を示し、美味しさと健康性の間には負の相関( $r = -0.52$ )が認められた。これらの相関パターンは、主観的価値が主に美味しさによって駆動される一方、健康性は独立した、あるいは相反する要因として機能していることを示唆している。研究は国立精神・神経医療研究センター倫理委員会の承認を受けて実施された。



図 1: 主観的価値による上位(上)および下位(下)の食品画像の例。

### 4.2 fMRI 実験

#### 4.2.1 参加者

fMRI 実験には、オンライン調査とは別に健康な成人男女 31 名(年齢: 平均 21.29 歳、範囲 18–25 歳、女性 11 名)が参加した。全員が右利きであり、神経学的または精神医学的疾患の既往歴はなかった。参加者は研究の目的と手順について説明を受け、書面によるインフォームドコンセントを提供した。研究は一橋大学倫理委員会の承認を受けて実施された。

## 4.2.2 刺激

刺激には、Food-Pics (Blechert et al., 2014; 2019) から選択された 568 枚の食品画像を使用した。これらの画像は、さまざまな食品カテゴリー(果物、野菜、スナックなど)を網羅している。

## 4.2.3 fMRI 実験手続き

fMRI 実験では、参加者はスキャナー内で食品画像を視覚的に評価した。各試行では、食品画像が 4 秒間表示され、その後、2.5 秒間の評価期間が続いた。参加者は、4 段階リッカート尺度(1 = 強く不同意、4 = 強く同意)を使用して、画像に対する「好み」(主観的価値)を評価した。全体で 568 試行があり、各参加者は 3 日間にわたり 1 日あたり 4 回のセッションに分けて実施した。試行間には 2~12 秒(平均 7 秒)のランダムなインターバルが設けられた。刺激提示とデータ収集は PsychoPy を用いて行われた(図 2, fMRI の基本的な知識については付録 A を参照)。

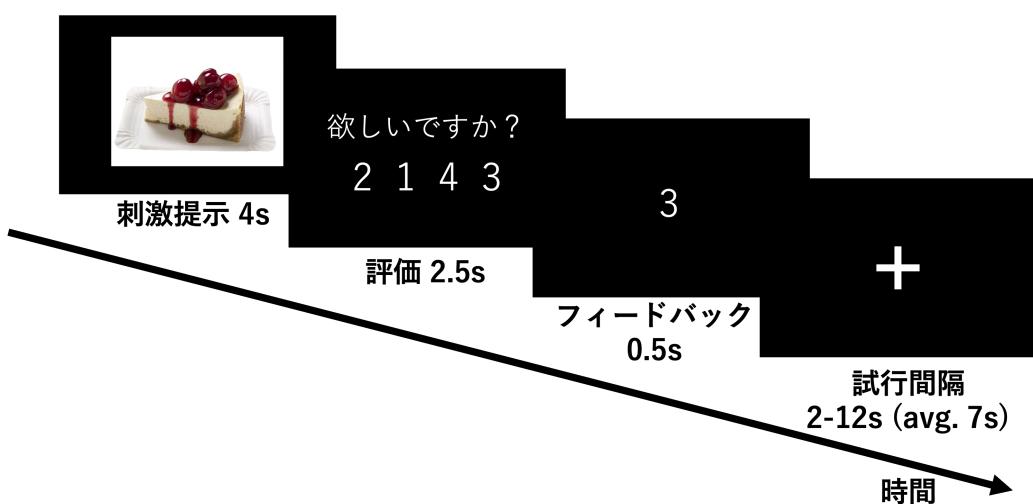


図 2: 食品評価課題における 1 試行のタイムライン。各試行において、参加者は 1 つの食品に対する「どのくらい欲しいか」(すなわち主観的価値)を報告した。この評価は支払意思額(Willingness to Pay; WTP)に類似した概念であり、食品に対する動機づけの強さを反映している。評価フェーズでは、キーと選択肢の順序が試行間でランダム化されている。

## 4.2.4 使用した DNN モデル

本研究では、3 つの異なる DCNN アーキテクチャ(VGG、ResNet、ConvNeXt)を使用して、食品画像から主観的価値を予測するモデルを構築した。さらに、マルチモーダルモデルである CLIP も使用し、画像エンコーダーの最終埋め込みを用いて主観的価値を予測した。(DNN については 付録 B 付録 C を参照)

### 4.2.4.1 VGG

VGG は、Visual Geometry Group によって開発された深層畠み込みニューラルネットワーク(CNN)アーキテクチャであり、画像認識タスクで高い性能を示している (Simonyan & Zisserman, 2015)。VGG は、非常に深い層構造を持ち、 $3 \times 3$  の小さな畠み込みフィルターを使用して特徴抽出を行う。VGG は、16 層(VGG16)および 19 層(VGG19)のバリエーションがあり、各層は畠み込み層、プーリング層、および全結合層で構成されている。

#### 4.2.4.2 ResNet

ResNet(Residual Network)は、Microsoft Researchによって開発された深層畳み込みニューラルネットワーク(CNN)アーキテクチャであり、非常に深いネットワークの学習を可能にするための残差学習フレームワークを導入している(He et al., 2016)。ResNetは、スキップ接続(skip connections)を使用して、層間の情報伝播を容易にし、勾配消失問題を軽減する。各ResNetブロックは、複数の畳み込み層とバッチ正規化層、およびReLU活性化関数で構成されている。

#### 4.2.4.3 ConvNeXt

ConvNeXtは、従来の畳み込みニューラルネットワーク(CNN)アーキテクチャを改良したものであり、Vision Transformer(ViT)の設計原則を取り入れている(Liu et al., 2022)。ConvNeXtは、深い層と広い受容野を持つことで、高次特徴の抽出能力が向上している。また、ConvNeXtは、正規化手法や活性化関数の選択など、最新の技術を採用しており、従来のCNNと比較して性能が向上している。

#### 4.2.4.4 CLIP

CLIP(Contrastive Language-Image Pretraining)は、OpenAIによって開発されたマルチモーダルモデルであり、画像とテキストのペアを用いて学習されている(Radford et al., 2021)。CLIPは、画像エンコーダーとテキストエンコーダーの2つの主要なコンポーネントで構成されており、これらは共通の埋め込み空間にマッピングされる。CLIPは、大規模なインターネットデータセットから収集された画像とテキストのペアを使用して学習されており、ゼロショット学習能力を持つ。つまり、CLIPは事前に見たことのないクラスの画像に対しても、高い分類性能を示すことができる。CLIPは、画像キャプション生成、画像検索、視覚質問応答などのさまざまなタスクで優れた性能を示している。CLIPは視覚情報処理と自然言語処理の計算モデルとして、脳の視覚野および言語処理に関与する領域の神経応答と比較することで、これらの情報処理メカニズムの解明に寄与する可能性がある。

### 4.3 DNNモデルの訓練

#### 4.3.1 交差検証とデータ拡張

データリークを防ぐため、まず896枚の元画像を6分割し、各フォールドで5分割を訓練用、1分割を検証用とした。データ拡張は訓練フォールドのみに適用し、検証フォールドには元画像のみを使用した。これにより、同一元画像由来の拡張版が訓練・検証間で重複することを防いだ。

訓練フォールドに適用したデータ拡張手法は以下の通りである：リサイズ( $240 \times 240$ )、センタークロップ( $224 \times 224$ )、水平反転、アフィン変換( $\pm 20\%$ 、上下左右への平行移動 $20\%$ 、スケール $70\text{--}120\%$ )、ガウシアンブラー(カーネル $5 \times 5$ 、 $\sigma = 0.01\text{--}4.0$ )、カラージッター(明るさ $7.5\%$ 、色相 $3\%$ 、彩度 $3\%$ )、正規化(訓練セットの平均と標準偏差に基づく)。各元画像に対して5つの拡張バージョンを生成し、訓練セットを6倍に拡大した。

### 4.3.2 モデル訓練

事前学習済みの ConvNeXt-Base、ResNet152、および VGG16 モデルを使用し、最終の全結合層を主観的価値スコア(1–8)を出力するように置き換え、全層をファインチューニングした。損失関数には Huber 損失( $\delta = 1$ )を使用し、AdamW オプティマイザを用いて学習を行った。学習率は 1e-4、ミニバッチサイズは 373(GPU メモリ制約による)、エポック数は 250 とした。6 分割交差検証における各検証フォールドからの予測を連結し、実際の評価との相関を計算した。計算環境は、Intel Xeon w5-2465X プロセッサと RTX 4000 Ada GPU を搭載し、Python 3.12.7 および PyTorch 2.5.0(CUDA 対応)を使用した。

## 4.4 DNN モデル評価

### 4.4.1 事前学習モデルの評価

事前学習済みモデル(ファインチューニングなし)の予測性能を評価するため、各モデルの最終層手前から特徴量を抽出した:VGG16 は avgpool 後(25,088 次元)、ConvNeXt-Base は avgpool 後(1,024 次元)、ResNet152 は avgpool 後(2,048 次元)、CLIP は encode\_image 出力(768 次元)。抽出した特徴量は PCA により 512 次元に統一し、リッジ回帰( $\alpha = 1.0$ )を用いて主観的価値を予測した。8 分割交差検証により性能を評価し、予測と実際の評価とのピアソン相関係数を算出した。

### 4.4.2 ファインチューニングモデルの評価

DCNN モデル(ConvNeXt、ResNet、VGG)の性能は、6 分割交差検証における予測評価と実際の評価との相関係数(Pearson の相関係数)で評価した。各モデルの予測精度を比較し、最も高い相関を示したモデルを特定した。

## 4.5 DNN のデコーディング解析

DNN の各層の活性化パターンを分析し、異なる種類の情報(低次画像特徴、栄養属性、美味しさ、健康性、主観的価値)がどの層で表現されているかを調査した。各層について、 $896 \times d$ (画像 × ユニット)の活性化マトリックスを構築し、主成分分析(PCA)を適用して累積説明分散が 80% に達するまで次元削減を行い、得られた主成分をリッジ回帰の予測子として使用して各属性(主観的価値、美味しさ、健康性、色、栄養)を推定した。正則化パラメータは 8 分割交差検証で最適化し、性能は予測と実際の評価との相関係数(Pearson の相関係数)で評価した(付録 D を参照)。

## 4.6 fMRI データ収集

fMRI 画像は、一橋大学に設置された 3 テスラの MRI スキャナー(Siemens MAGNETOM Prisma)を使用して収集された。機能的画像は、T2\*強調マルチバンドエコーブラナーイメージング(MB-EPI)シーケンスで取得された。主な取得パラメータは以下の通りである:繰り返し時間(TR)800 ms、エコー時間(TE)34.4 ms、フリップ角 52 度、マルチバンドファクター 6、ボクセルサイズ  $2.4 \times 2.4 \times 2.4$  mm、マトリッ

クスサイズ  $86 \times 86$ 。各セッションで 700 ボリュームが取得された(スキャン時間:約 560 秒)。受信には 32 チャンネルヘッドコイルを使用した。高解像度の T1 強調構造画像も取得された(MPRAGE シーケンス)。fMRI データは、MATLAB R2025a および MacBook Pro (14 インチ, M4 Pro, 2024, Mac OS X 15.1) 上で SPM25 を用いて解析した。データ収集および解析は、実験条件を盲検化せずに実施した。

## 4.7 fMRI データ前処理

fMRI データの前処理は、fMRIprep 23.2.1 (Esteban et al., 2018) を使用して行われた。前処理手順には、以下が含まれる:スライスタイミング補正、モーション補正、コレジストレーション、空間正規化 (MNI152NLin2009cAsym テンプレートへの変換)。この処理の後に、空間平滑化(6 mm FWHM ガウスカーネル)が SPM25 で実施された。前処理後のデータは、統計解析および DCNN との比較のために使用された。

## 4.8 fMRI データ解析

### 4.8.1 GLM 分析

一般線形モデル(GLM)に基づいて行われた(付録 E を参照)。各参加者の前処理済み fMRI データに対して、以下の条件を含む GLM を構築した:

1. **Image 条件:** 食物画像提示期間(画像提示から質問提示まで)。この条件には、以下の 8 つのパラメトリックモジュレーター(parametric modulator)を含めた:
  - 主観的価値(Value): 参加者が評定した食物の好ましさ
  - 色情報(R, G, B): 画像の赤・緑・青チャンネルの平均値
  - 栄養成分(Protein, Fat, Carbs, Kcal): 食品のタンパク質、脂質、炭水化物、カロリー
2. **Question 条件:** 質問提示期間(質問提示から評定開始まで)
3. **Response 条件:** 評定応答時点(持続時間 0 秒)
4. **Feedback 条件:** フィードバック提示期間(0.5 秒)
5. **Miss 条件:** 応答がなかった試行全体

パラメトリックモジュレーターは直交化せずにモデルに投入し( $\text{orth} = 0$ )、各変数の独立した効果を推定した。さらに、6 つの頭部運動パラメータを共変量として含めることで、頭部運動の影響を統計的に除去した。各条件は、試行開始時点でのオンセットと持続時間に基づいてモデル化され、各参加者のデザインマトリックスが構築された。GLM は SPM25 で実装され、各参加者のベータ画像が推定された。

### 4.8.2 集団レベル解析

各参加者の個別レベルの GLM 解析結果を用いて、グループレベルのランダム効果解析(one-sample t-test)を実施した(付録 F を参照)。主観的価値に関連する脳領域を特定するために、クラスターレベル FWE 補正を適用した(クラスター形成閾値: $p < 0.001$  uncorrected、クラスターレベル FWE: $p < 0.05$ ; 付

録 G を参照)。クラスターレベル FWE 補正は SPM のランダム場理論 (Random Field Theory; RFT)に基づき、多重比較補正を行った。

#### 4.8.3 表現類似性解析(RSA)

DNN の内部表現と脳活動パターンの構造的類似性を評価するため、表現類似性解析 (RSA) (Kriegeskorte et al., 2008) を実施した(付録 H を参照)。RSA では、刺激セット内の各ペア間の神経活動パターンの非類似度を表現非類似度行列 (RDM) として表現し、異なるシステム(脳と DNN)間の RDM を比較することで、表現構造の類似性を定量化する。

##### 4.8.3.1 ROI-RSA

ROI-RSA では、Harvard-Oxford 確率的アトラス(Frazier et al., 2005; Makris et al., 2006) から 31 個の ROI を定義した(25% 以上の確率を示すボクセルを含む)。視覚皮質(V1、初期視覚野、LOC、紡錘状回、IT)、頭頂葉(SPL、IPL、角回、楔前部)、側頭葉(STG、MTG、側頭極、PHC)、前頭葉(IFG、Broca 野、DLPFC、VLPFC、OFC、前頭極)、帯状皮質(ACC、PCC)、島皮質、および皮質下構造(海馬、扁桃体、尾状核、被殻、側坐核、視床、淡蒼球)を含めた。

各 ROIにおいて、Least Squares Separate(LSS)法(Mumford et al., 2012)を用いて各食品画像に対する単一試行ベータ値を推定した。490 枚の共通画像について、各被験者の ROI 内のベータパターンから表現類似度行列 (RDM) を構築した(1 - Pearson の相関係数)。DNN の各層についても、同じ 490 枚の画像に対する活性化パターンから RDM を算出した。脳 RDM と DNN RDM の類似度はスピアマン相関で評価した。

モデルの説明力の上限を評価するため、Leave-One-Out 法によるノイズ上限値(NC)を算出した。NC 上限は各被験者の RDM と全被験者平均 RDM との相関の平均値として、NC 下限は各被験者の RDM と当該被験者を除いた残りの被験者の平均 RDM との相関の平均値として定義した。NC 上限比(モデル相関 / NC 上限 × 100)により、理論的に達成可能な最大説明力に対するモデルの到達度を評価した。

#### 4.8.4 エンコーディング解析

DNN の各層の活性化パターンと fMRI データの脳活動パターンを比較するために、3 レベル階層的 PCA(主成分分析)を用いた(付録 I を参照)。この手法は、DNN の活性化パターンを直交化された階層構造に分解し、fMRI データの予測子として使用する。

まず、DNN の層を階層的な情報処理段階に基づいて 4 つのグループに分類した:(1) **Initial 層**: 低次視覚特徴(エッジ、テクスチャ)を抽出する初期層、(2) **Middle 層**: 中次特徴を抽出する中間層、(3) **Late 層**: 高次特徴(オブジェクトの形状やカテゴリ)を抽出する後期層、(4) **Final 層**: 最終的な分類や埋め込み表現を生成する層。ConvNeXt と CLIP は共に ConvNeXt-Base アーキテクチャを画像エンコーダーとして使用しているため、対応する層グループを同一の基準で抽出した。ConvNeXt では、Initial 層に features\_1 および features\_3 の全プロックと features\_5 の初期プロック(0-3)、Middle 層に features\_5

の中間ブロック(4-13)、Late 層に features\_5 の後期ブロック(14-22)、Final 層に features\_5 の最終ブロック(23-26)と features\_7 の全ブロックおよび classifier\_1 を含めた。CLIP では、Initial 層に stage0-1 の全ブロックと stage2 の初期ブロック(0-3)、Middle 層に stage2 の中間ブロック(4-13)、Late 層に stage2 の後期ブロック(14-22)、Final 層に stage2 の最終ブロック(23-26)と stage3 の全ブロックおよび head を含めた。

層グループ	ConvNeXt	CLIP
Initial	features_1, features_3, features_5[0-3]	stage0-1, stage2[0-3]
Middle	features_5[4-13]	stage2[4-13]
Late	features_5[14-22]	stage2[14-22]
Final	features_5[23-26], features_7, classifier_1	stage2[23-26], stage3, head

表 1: ConvNeXt および CLIP の層グループ分類

次に、3 レベル階層的 PCA を適用した。各画像について各層の活性化を抽出し、以下の 3 レベルの主成分を算出した：

1. **Global PC:** すべての層に共通する分散を捕捉する主成分(累積寄与率 60% に達するまで抽出)
2. **Layer-Shared PC:** 隣接する層グループ間で共有される分散を捕捉する主成分(正準相関分析(CCA)を用いて抽出、各隣接ペアについて 2 成分ずつ、計 6 成分)
3. **Layer-Specific PC:** 各層グループに固有の分散を捕捉する主成分(Global および Shared 成分を除去した残差に対して、累積寄与率 50% に達するまで PCA を適用)

累積寄与率の閾値は、情報保持とモデルの簡潔さのバランスを考慮して事前に設定した。Global PC では全層に共通する主要な分散構造を捕捉するため比較的高い閾値(60%)を、Layer-Specific PC では共有成分除去後の残差から層固有の情報を抽出するためより低い閾値(50%)を採用した。各レベルの主成分は、QR 分解を用いて相互に直交化し、共線性を排除することで各レベルの独立した説明力を評価可能とした。

GLM では、これらの主成分を画像提示時のパラメトリックモジュレータとして使用した。パラメトリックモジュレータは日ごとにまとめ(3 セッション)、画像提示時の定数項はランごとに設定した。SPM の自動直交化は無効化し(orth = 0)、事前に直交化された主成分構造を保持した。6 つの頭部運動パラメータもランごとに共変量として含めた。統計的検定のためのコントラストはセッション間で平均化した。

統計検定では、各レベルおよび各層グループについて F 検定を用いて説明力を評価した：

1. Global\_F: すべての Global PC の説明力
2. 各層グループの F 検定(Initial\_F, Middle\_F, Late\_F, Final\_F): 各層グループ固有の Layer-Specific PC の説明力
3. 各層グループ+関連 Shared(例: Initial\_withShared\_F): Layer-Specific PC と関連する Layer-Shared PC を組み合わせた説明力

#### 4. 各 shared ペアの F 検定(例:Initial-Middle\_Shared\_F):隣接する層グループ間で共有される Layer-Shared PC の説明力

有意な脳領域を同定するため、クラスターレベル FWE 補正を適用した(クラスター形成閾値: $p < 0.001$  uncorrected、クラスターレベル FWE: $p < 0.05$ )。本研究では 2 モデル(ConvNeXt, CLIP)×10 コントラストの計 20 回の統計検定を実施した。コントラスト間の Bonferroni 補正は以下の理由から適用しなかった:(1) Global PC、Layer-Shared PC、Layer-Specific PC は 3 レベル階層的 PCA により直交化されており、各成分は統計的に独立した分散を捉えている、(2) 各層グループ(Initial, Middle, Late, Final)は視覚処理の異なる階層段階に対応し、先行研究に基づく事前仮説を検証している(Kriegeskorte, 2015; Yamins et al., 2014)、(3) 2 つのモデル(ConvNeXt, CLIP)は異なるアーキテクチャと学習目標を持ち、独立した計算仮説を表現している。各コントラスト内ではクラスターレベル FWE 補正により偽陽性率を厳密に制御した。さらに、グループレベルのランダム効果解析を行い、各レベルおよび各層グループと有意に相関する脳領域を同定した(クラスターレベル FWE 補正、 $p < 0.05$ )。

##### 4.8.4.1 ROI 解析

仮説駆動型の解析として、先行研究に基づいて定義した 16 個の関心領域(ROI)における効果を検証した。これらの 16 ROI は、ROI-RSA で使用した 31 ROI の中から、視覚処理および価値処理に関する先行研究の知見に基づいて選定したサブセットである。ROI は Harvard-Oxford 確率的アトラス(Frazier et al., 2005; Makris et al., 2006) を用いて定義し、25% 以上の確率を示すボクセルを含めた。

視覚処理に関連する ROI として、一次視覚野(V1)、外側後頭皮質(LOC)、紡錘状回(Fusiform)、下側頭回(IT)を定義した。価値処理に関連する ROI として、腹内側前頭前皮質(vmPFC)、眼窩前頭皮質(OFC)、前帯状皮質(ACC)、側坐核(NAcc)、線条体(Striatum:尾状核+被殻)を定義した。その他の ROI として、島皮質(Insula)、扁桃体(Amygdala)、海馬(Hippocampus)、海馬傍回(PHC)、上頭頂小葉(SPL)、楔前部(Precuneus)、角回(Angular Gyrus)を含めた。

各 ROI において、Small Volume Correction(SVC)を適用し、ROI 内での FWE 補正を行った( $p < 0.05$ )。SVC はランダム場理論に基づき、各 ROI のボクセル数と平滑度推定値(FWHM)から算出した RESEL(resolution element:空間平滑化後のデータにおける独立な解像度要素)数を用いて p 値を補正した。効果量は  $\beta$  値の RMS(二乗平均平方根)として算出した。

## 5 結果

### 5.1 DNN モデルによる主観的価値予測

3つのDCNNモデル(ConvNeXt、ResNet、VGG)とCLIPを用いて、食品画像から主観的価値を予測した(図3)。事前学習済みモデルの比較では、CLIPモデルが最も高い予測精度を示し、予測と実際の評価との相関係数は  $r = 0.78$  であった。一方、事前学習のみのDCNNモデルはそれぞれConvNeXt  $r = 0.63$ 、ResNet  $r = 0.44$ 、VGG  $r = 0.19$  と低い精度を示した。食品評価データでファインチューニングを行った結果、DCNNモデルの予測精度は大幅に向上し、ConvNeXt  $r = 0.69$ 、ResNet  $r = 0.61$ 、VGG  $r = 0.58$  となった。これらの結果は、CLIPが事前学習のみで高い予測精度を達成する一方、DCNNはファインチューニングにより性能が向上することを示している。

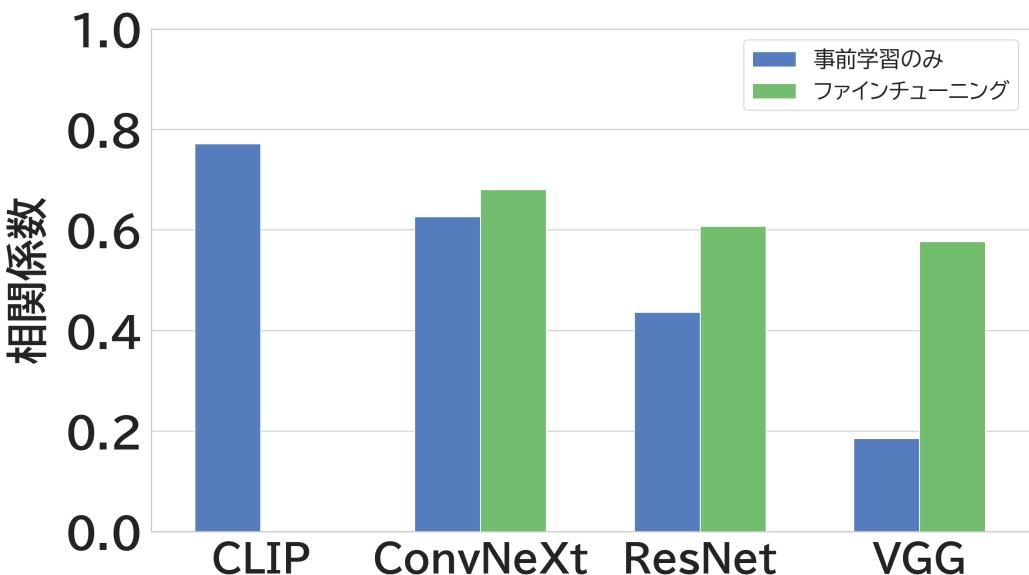


図3: DNNモデルによる主観的価値予測の精度(予測と実際の評価との相関係数)。青:事前学習のみ、緑:ファインチューニング後。CLIPは事前学習済みモデルのみ使用。

### 5.2 DNN の層別情報表現

ファインチューニング後のConvNeXtモデルおよび事前学習済みCLIPモデルの各層における情報表現を分析し、異なる種類の情報(低次画像特徴、栄養属性、美味しさ、健康性、主観的価値)がどの層で表現されているかを調査した。各層について、主成分分析(PCA)を適用し、得られた主成分をリッジ回帰の予測子として使用して各属性を推定した。層ごとの説明力(予測と実際の評価との相関係数)を計算し、図4に示すように、各属性の層別表現を評価した。高次属性(主観的価値、美味しさ、健康性)は初期層での説明力が低く、後期層で増加する傾向が見られた。一方、低次の色情報(赤、緑、青)は初期層で高い説明力を示し、後期層でも情報が保持されていた。栄養属性はConvNeXtでは全体的に弱い表現を示したが、CLIPでは後期層で説明力が増加した。

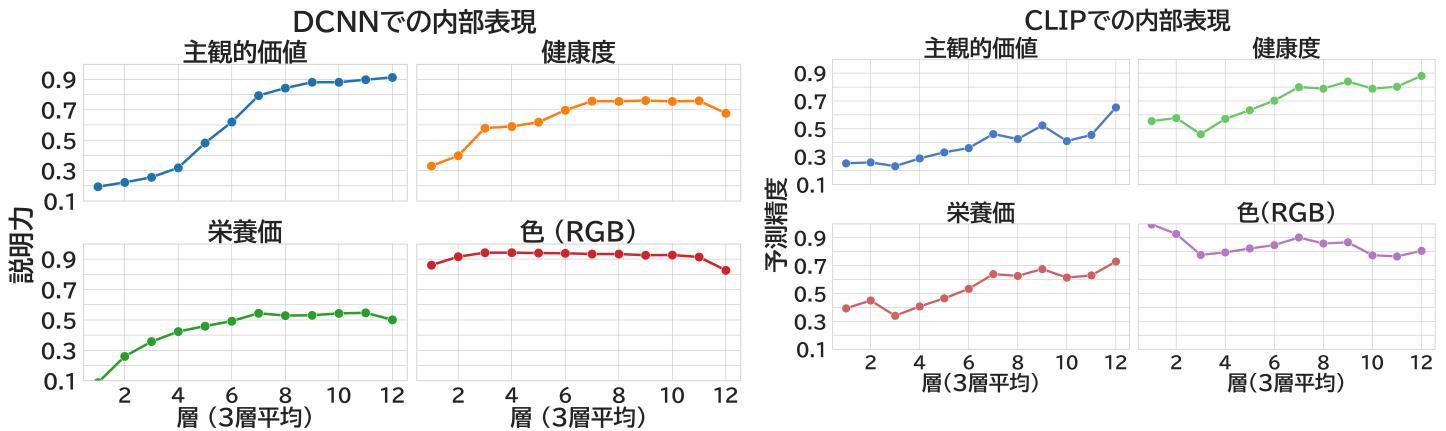


図 4: DNN モデルの各層における情報表現。左: フайнチューニング後の ConvNeXt、右: 事前学習済み CLIP。各線は、各属性の予測と実際の評価との相関係数を示す。

これらの結果は、ファインチューニング後の ConvNeXt および事前学習済み CLIP モデルにおいて、高次属性が後期層で主に表現され、低次の色情報が初期層で強く表現される傾向があることを示している。なお、事前学習済み(ファインチューニングなし)の ConvNeXt モデルについても同様の分析を行った結果を Section J に示す。

## 5.3 fMRI 解析結果

### 5.3.1 主観的価値に関連する脳活動

GLM 解析により、食品画像に対する主観的価値評価と有意に相關する脳領域を同定した。主観的価値のパラメトリックモジュレーター (Image  $\times$  Value) について、クラスターレベル FWE 補正 ( $p < 0.05$ 、クラスター形成閾値  $p < 0.001$  uncorrected) を適用した結果、87 個の有意なクラスター (総ボクセル数 11,697、ピーク T 値 6.80) が検出された(図 5)。

活性化は広範な脳領域で観察され、特に後頭葉の視覚関連領域(一次視覚野、外側後頭皮質、紡錘状回)、側頭葉(下側頭回)、前頭葉(腹内側前頭前皮質、眼窩前頭皮質)、および頭頂葉で顕著であった。これらの結果は、食品の主観的価値評価が視覚処理から価値表現に至る広範なネットワークを動員することを示唆している。

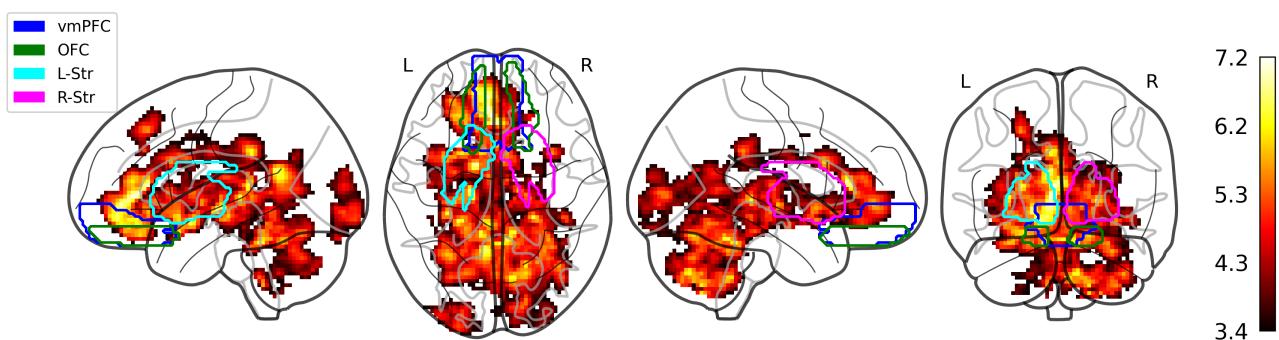


図 5: 主観的価値 (Image  $\times$  Value) に関する脳活動。クラスターレベル FWE 補正 ( $p < 0.05$ 、クラスター形成閾値  $p < 0.001$  uncorrected)。カラーバーは T 値を示す。色付きの輪郭線は関心領域 (ROI) を示す: vmPFC (青)、OFC (緑)、左線条体 (シアン)、右線条体 (マゼンタ)。

### 5.3.2 ROI 解析

先行研究に基づいて定義した関心領域(ROI)における主観的価値の効果を検証した。Small Volume Correction(SVC)を適用したROI解析の結果、価値処理に関連する複数の領域で有意な効果が認められた(図6)。

腹内側前頭前皮質(vmPFC)では有意な効果が観察され、主観的価値の表出に関与することが確認された。内側眼窩前頭皮質(mOFC)でも有意な効果が認められた。さらに、左右の線条体でも有意な効果が観察され、報酬処理への関与が示唆された。

島皮質および扁桃体でも統計的に有意な効果が認められた。これらの結果は、食品の主観的価値がvmPFC、mOFC、線条体に加え、島皮質や扁桃体を含む広範な脳領域で表現されることを示している。

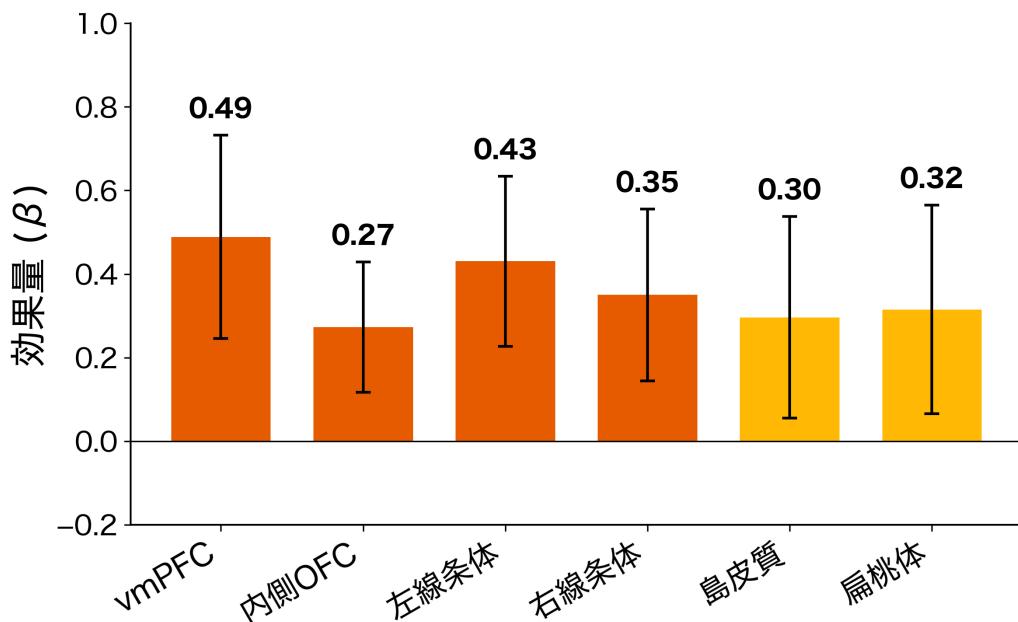


図6: ROI解析における主観的価値の効果量( $\beta$ )。すべてのROIで有意な効果が認められた( $p < 0.05$ )。vmPFC:腹内側前頭前皮質、mOFC:内側眼窓前頭皮質。

### 5.3.3 DNN 比較分析

DNNの内部表現とfMRIデータの脳活動パターンを比較するために、表現類似性解析(RSA)と3レベル階層的PCAを用いたエンコーディング解析を行った。

#### 5.3.3.1 表現類似性解析(RSA)

DNNの内部表現と脳活動パターンの構造的類似性を評価するため、関心領域(ROI)ベースの表現類似性解析(RSA)を実施した。各ROIにおいて、490枚の食品画像に対するベータ値から表現類似度行列(RDM)を構築し、DNNの各層の活性化パターンから同様に算出したRDMとのスピアマン相関を計算した。

モデルの説明力を評価するため、Leave-One-Out法によるノイズ上限値(NC)を算出した。NC上限は各被験者と全被験者平均RDMとの相関、NC下限は各被験者と残りの被験者の平均RDMとの相

関として定義した。NC 上限比(モデル相関 / NC 上限 × 100)により、理論的に達成可能な最大説明力に対するモデルの到達度を評価した。

### 5.3.3.1.1 ROI-RSA 結果

Double-centering を適用した RSA 解析の結果、視覚野において最も高い説明率が観察された。V1 では 3 モデルとも約 65-69% の説明率を達成し、CLIP が 67.0%、ImageNet が 64.3%、Food が 68.7% であった。初期視覚野でも同様に、CLIP が 62.5%、ImageNet が 63.5%、Food が 68.7% の説明率を示した。

全 31 ROI における平均説明率を比較すると、CLIP モデルが 13.9% で最も高く、Food モデル (12.9%)、ImageNet モデル (12.0%) が続いた(右)。統計検定の結果、CLIP と ImageNet の間には有意差が認められ(対応のある t 検定、 $t(30) = 7.05, p < 0.001$ )、Food と ImageNet の間にも有意差が認められた( $t(30) = 2.10, p = 0.045$ )。一方、視覚野のみ(n=5)での比較では、3 モデル間に有意差は認められなかった(CLIP vs ImageNet:  $p = 0.059$ )。

高次視覚野および価値関連領域では説明率が低下した。LOC(CLIP: 36.9%、ImageNet: 32.2%、Food: 29.0%)、紡錘状回(CLIP: 39.1%、ImageNet: 35.2%、Food: 30.7%)では約 30-40% の説明率に留まった。vmPFC(CLIP: 4.7%、ImageNet: 3.9%、Food: 8.0%)ではさらに低い値を示した。

これらの結果は、DNN モデルが初期視覚野の表現構造を高い精度で説明できる一方、高次領域および価値関連領域の表現については未説明の分散が大きいことを示唆している。また、CLIP モデルが全 ROI 平均で最も高い説明率を示したことは、視覚-言語モデルが脳の表現構造をより広範に捉えている可能性を示している。

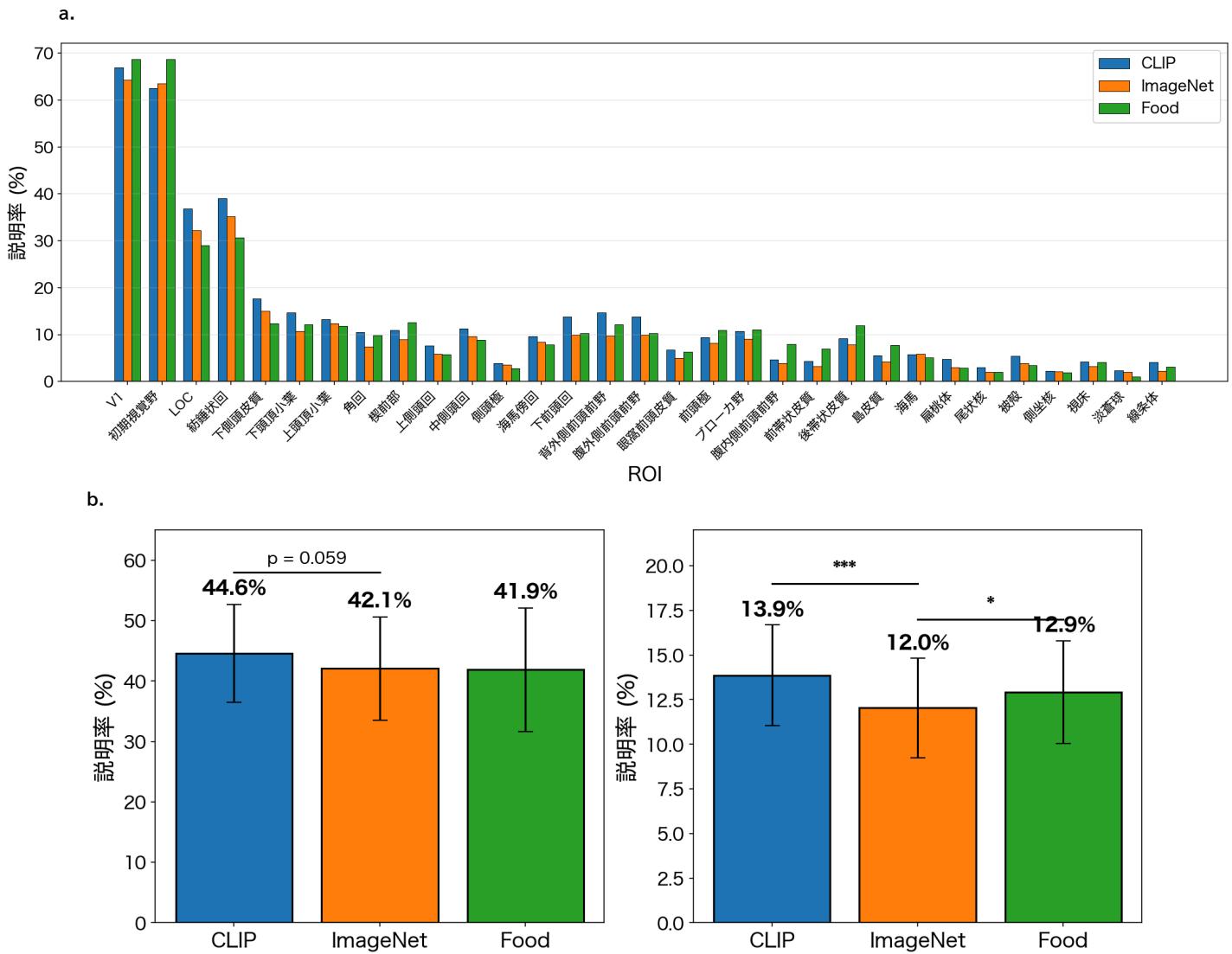


図 7: ROI-RSA 解析結果(Double-centering 適用)。a. 各 ROI におけるモデル別説明率の比較。b. 右は視覚野(V1、初期視覚野、LOC、紡錐状回、IT)における平均説明率。左は全 31 ROI における平均説明率。エラーバーは標準誤差(SEM)。\*\*\*  $p < 0.001$ 、\*  $p < 0.05$ (対応のある t 検定)。

### 5.3.3.2 エンコーディング解析

#### 5.3.3.2.1 ConvNeXt モデル

ConvNeXt モデルの階層的解析結果を図 8a に示す。各層グループに関連する共有成分を加えた解析の結果、初期層は後頭葉の一次視覚野(V1)および外側後頭皮質で有意な効果を示した。中間層は側頭葉および頭頂葉で活性化を示した。後期層は広範な活性化を示し、視覚野から側頭葉、頭頂葉にわたる領域で有意な効果が認められた。最終層は最も広範な活性化を示し、腹内側前頭前皮質(vmPFC)を含む前頭葉領域でも有意な効果が認められた。

これらの結果は、ConvNeXt の階層構造が脳の視覚処理階層と対応しており、初期層が低次視覚野、後期層が高次視覚野および価値関連領域と相関することを示唆している。

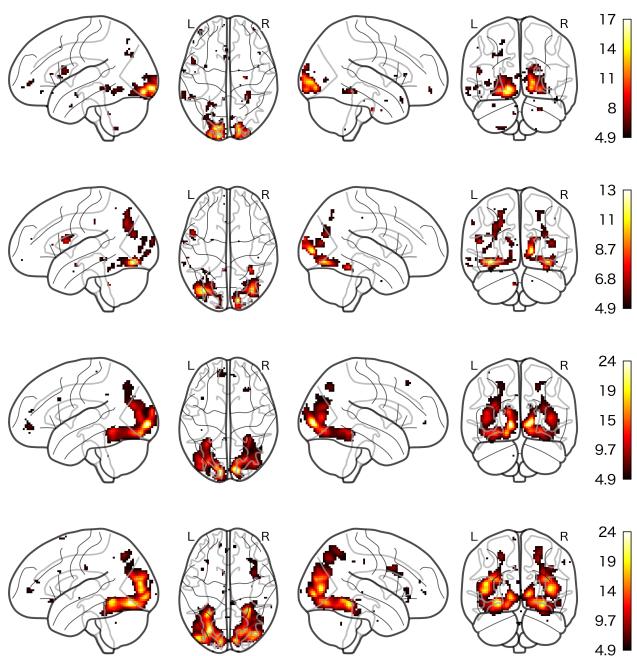
#### 5.3.3.2.2 CLIP モデル

CLIP モデルの階層的解析結果を図 8b に示す。初期層は後頭葉で最も強い活性化を示した。中間層は側頭葉および頭頂葉で広範な活性化を示し、視覚的オブジェクト認識に関連する領域との対応が示唆された。後期層および最終層は、ConvNeXt と比較して限定的な活性化パターンを示した。

CLIP モデルでは、視覚野から側頭葉にかけての領域と強く相関することが示された。初期層から中間層にかけての広範な活性化は、CLIP の学習方法の特性を反映している可能性がある。

a.

**CONVNEXT**



b.

**CLIP**

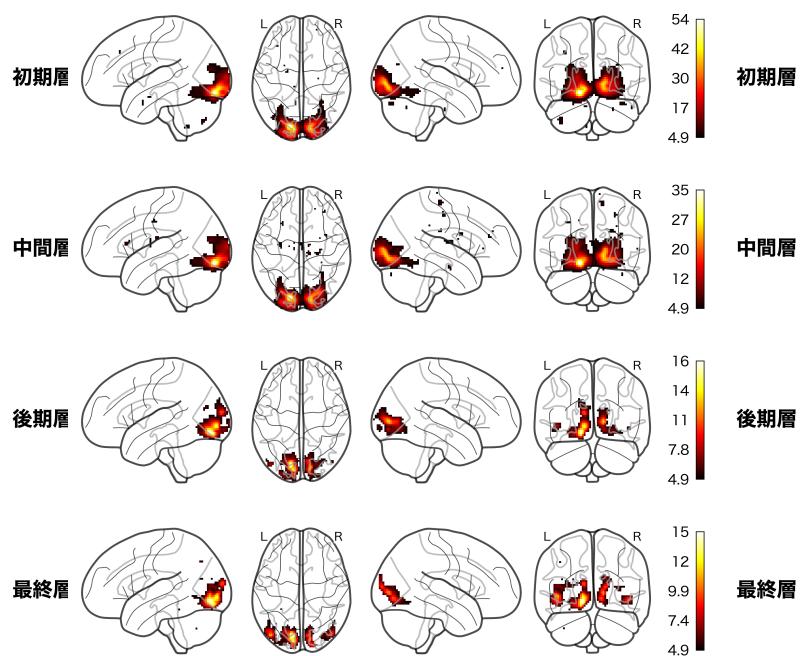


図 8: DNN モデルの層グループ別脳活動(FWE 補正  $p < 0.05$ )。a. ConvNeXt モデル。b. CLIP モデル。各行は異なる層グループ(初期層+共有、中間層+共有、後期層+共有、最終層+共有)と脳活動の対応を示す。

### 5.3.3.2.3 DNN 層グループと ROI の対応

DNN の各層グループと関心領域(ROI)の対応関係を詳細に検討するため、Small Volume Correction (SVC) を用いた ROI 解析を行った(図 9)。

視覚野(V1、初期視覚野、LOC、紡錘状回、IT、海馬傍回)(Grill-Spector & Malach, 2004; Kravitz et al., 2013) では、両モデルとも複数の層グループで有意な対応を示した(SVC FWE  $p < 0.05$ )。

ConvNeXt では V1、初期視覚野、LOC、IT が全層で有意であり、紡錘状回は中間層以降、海馬傍回は後期層以降で有意であった。CLIP では V1 と初期視覚野が全層で有意であり、LOC は初期・中間・最終層、紡錘状回は初期・中間層で有意であった。

空間注意系(SPL)(Corbetta & Shulman, 2002) では、ConvNeXt で SPL が中間層以降で有意であった。CLIP では有意な対応は認められなかった。

記憶系(PCC)では、ConvNeXt で後期層以降、CLIP では中間層で有意な対応が認められた。

言語野(IFG、MTG、STG、側頭極、角回)(Fedorenko et al., 2024) では、CLIP の中間層で IFG および左角回に有意な対応が認められた。ConvNeXt でも後期層以降で左角回に有意な対応が認められた。この左角回の効果は、視覚-言語モデル(CLIP)が視覚皮質活動を説明する際に左角回との接続が重要であることを示した先行研究(Chen et al., 2025) と整合する。

報酬系(NAcc、OFC)(Haber & Knutson, 2010)では、ConvNeXtは後期層以降で NAcc、最終層で OFC と有意な対応を示した。CLIP では中間層で NAcc と OFC、最終層でも NAcc に有意な対応が認められた。

価値判断系(vmPFC、DLPFC)(Kahnt et al., 2010)では、ConvNeXt の後期層以降で vmPFC と DLPFC に有意な対応が認められた。CLIP では有意な対応は認められなかった。

注意選択系(扁桃体、島皮質、ACC)(Morawetz et al., 2017; Seeley et al., 2007)では、ConvNeXt では有意な対応は認められなかった。一方、CLIP では中間層で扁桃体、島皮質、ACC すべてに有意な対応が認められた。

習慣学習系(尾状核、被殼)では、CLIP の中間層でのみ有意な対応が認められた。

これらの結果から、ConvNeXt は視覚処理の階層構造に沿って情報を処理し、後期層で価値判断系、空間注意系、記憶系と対応するのに対し、CLIP では中間層で報酬系、注意選択系、習慣学習系、記憶系との対応が認められた。

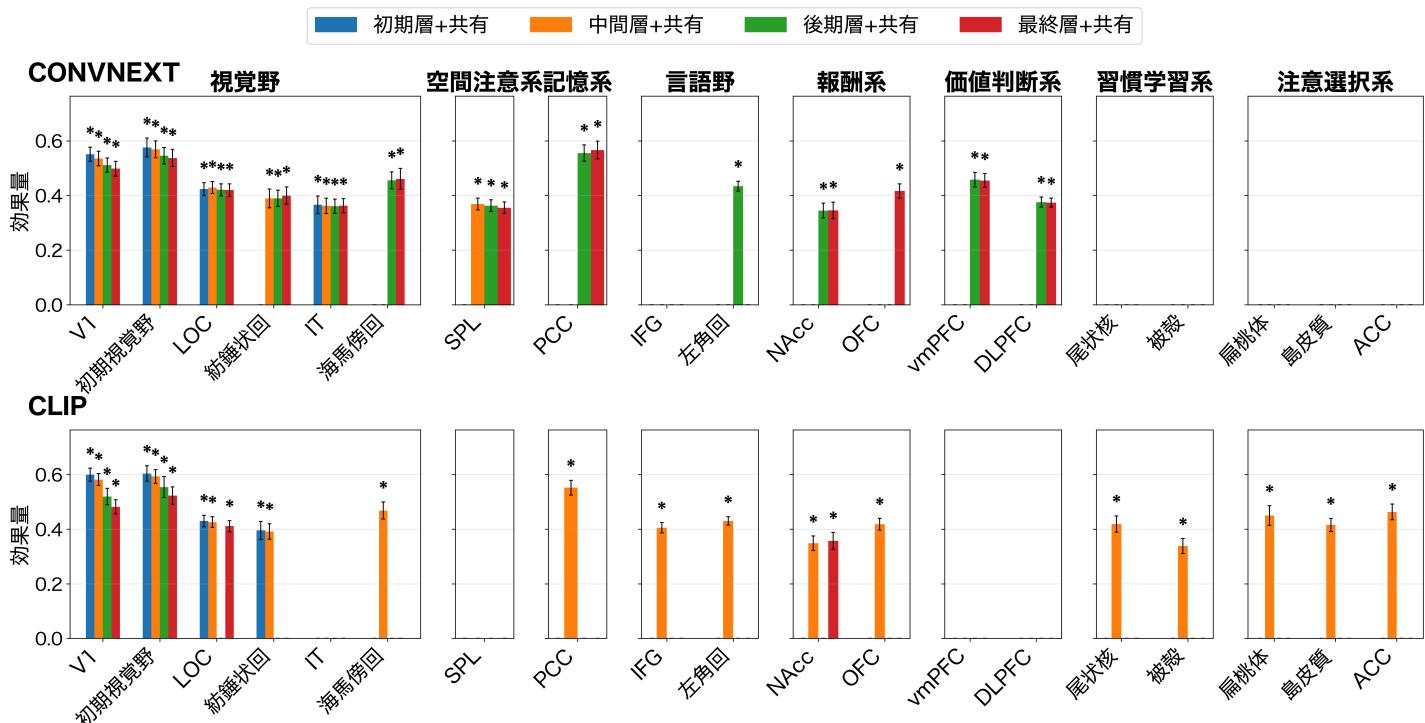


図 9: DNN モデルの層グループ別 ROI 効果量。各層(初期・中間・後期・最終)に関する共有成分を加えた効果量を示す。エラーバーは被験者間の標準誤差(SEM)。有意な効果(SVC FWE  $p < 0.05$ )。追加の効果量図は 補足 K を参照。

## 6 考察

本研究では、DNN を用いて食品画像から主観的価値を予測し、その内部表現と脳活動パターンを比較することで、食品評価の神経計算基盤を検討した。主要な発見は以下の通りである。第一に、CLIP モデルが最も高い予測精度( $r = 0.78$ )を示し、視覚-言語統合が食品の価値予測に重要であることが示された。第二に、DNN の層別解析により、高次属性(美味しさ、健康性、主観的価値)は後期層で、低次の色情報は階層全体を通じて表現されることが明らかになった。第三に、ROI-RSA 解析により、DNN モデルは初期視覚野の表現構造を高い精度で説明できる(V1 で 65-69%)一方、価値関連領域(vmPFC: 4-8%)では説明力が低いことが示された。第四に、エンコーディング解析により、ConvNeXt は視覚処理階層に沿った対応を示し、CLIP はより早期の層で情動・報酬・言語関連領域との対応を示した。

DCNN の活性化における栄養属性の弱い表現は、主観的な栄養情報が主観的価値の予測に寄与することを示した先行研究(Suzuki et al., 2017)とは対照的である。一方で、CLIP では栄養価が高次層で説明力が高くなっていたことから、視覚モデルでは栄養情報が抽出できないが、視覚-言語モデルでは栄養情報が抽出できることが示唆された。これは、CLIP が Transformer を含み膨大な Web 上のデータを学習していることに起因し、意味情報の逆伝播によって画像の視覚情報から栄養情報を読み取ることを可能にしている可能性がある。

fMRI 解析において、主観的価値評価に関連する脳領域として、視覚野(V1、LOC、紡錘状回)、前頭葉(vmPFC、OFC)、および線条体が同定された。ROI 解析では、vmPFC および線条体が主観的価値の表出に重要な役割を果たすことが確認された。これらの結果は、先行研究における価値表現の知見と一致している(Chib et al., 2009; Hare et al., 2008; Suzuki et al., 2017)。

情報表現のパターンは、視覚芸術の価値評価に関する先行研究(Iigaya et al., 2021)の知見とは異なる。先行研究では、色情報は主に初期層で表現され、後期層では減少することが示されている。本研究では、食品画像において色がより持続的かつ影響力のある役割を果たし、階層的処理を通じて後期層まで説明力が維持された。この違いは、食品評価における色の重要性に起因する可能性がある。ヒトの三色型色覚は食物採集能力を向上させるために進化したと考えられており、食品の色は栄養価や鮮度を示すシグナルとして機能する(Foroni et al., 2016)。

RSA 解析の結果、DNN モデルは視覚野との類似度が高い一方、価値関連領域の表現についてはうまく説明できないことが示された。DCNN と脳のマッピングを試みた研究では、視覚領域における説明可能な分散の 60% 以上が説明されており(Dwivedi et al., 2021)、本研究の V1 における結果は先行研究と一致する。一方、vmPFC において説明率が 4-8% と低い値を示したことは、価値計算が視覚情報処理以外の多様な情報(感情、記憶、社会的文脈)を統合する複雑なプロセスであり、現在の DNN モデルではこれらの要素を十分に反映できていない可能性を示唆している。

エンコーディング解析の結果、CLIP モデルでは、より早期の層(中間層)で情動・報酬関連領域との対応を形成した。CLIP の中間層は、感情制御に関わる PCC・扁桃体・島皮質・ACC(Morawetz et al., 2017)、言語処理の中核である IFG(Fedorenko et al., 2024)、および意味処理に関わる左角回と有意な対応を示した。特に左角回との対応は、CLIP が腹側後頭側頭皮質(VOTC)の活動を説明する際に左角回との白質接続が重要であるという先行研究(Chen et al., 2025)と整合し、食品の主観的価値計算における言語的・意味的情報の統合を支持する。これらの結果は、食品評価において経験・言語・感情が価値計算に寄与することを示唆しており、感情制御(Morawetz & others, 2021) や食品ラベル(Grabendorst et al., 2013) が食品評価を変化させるという知見とも一致する。

本研究の解釈にはいくつかの方法論的限界がある。第一に、CLIP モデルは学習データ(LAION-400M vs ImageNet-1K)と学習タスク(対照学習 vs 分類)の両方が ImageNet モデルと異なるため、これらの効果を分離することは困難である。ただし、本研究では同一の ConvNeXt-Base アーキテクチャを使用しており、構造的な交絡は排除されている。また、事前学習データのスケールの違い(128 万 vs 4 億枚)は予測精度の差に影響しうるが、モデル内の階層的な情報表現構造や脳領域との対応関係は、データスケールよりもアーキテクチャに依存することが示されている(St-Yves et al., 2023)。実際、両モデルで類似した階層構造(初期層で低次視覚特徴、後期層で高次属性を表現)が観察された。今後の研究では、対照学習で学習した ImageNet モデルなどを用いて、学習データと学習方法の効果を分離して検討する必要がある。

第二に、DNN の最終層ではなく中間～後期層で脳との対応が最大になる点について、解釈には注意が必要である。この現象は、最終層が分類タスクに過剰に特化していることを反映している可能性がある一方、脳の視覚処理が再帰的な情報統合を含むことを示唆している可能性もある。これらの解釈を区別するには、再帰接続を明示的にモデル化した DNN との比較が必要である。

第三に、本研究は相関解析に基づいており、DNN と脳が類似した計算を行っているという因果的結論を導くことはできない。経頭蓋磁気刺激法(TMS)や動物実験などで、因果関係の検証もしくは、動的因果モデリング(DCM)を用いた情報フローの解析が今後の課題である。

第四に、DCNN は食品に関する意味的・概念的知識を直接エンコードせず、重要な点で人間の視覚と異なる可能性がある(Bowers et al., 2023; Caplette & Turk-Browne, 2024)。今回の RSA 解析では、CLIP モデルが他の DNN モデルよりも高い説明力を示したが、高次領域での説明力は依然として低かった。そのため、将来的には、視覚情報と概念的知識や感情情報を統合した専門領域モデルを組み合わせたマルチモーダルモデルの開発が期待される。

また、本研究では全参加者の平均評価を使用したが、特定の集団(肥満者や特異な食習慣を持つ個人)への解析の拡張は重要な方向性である。先行研究では、肥満や摂食障害において嗜好が異なる可能性が示唆されており(Foerde et al., 2015; Spinelli & Monteleone, 2021)、集団間で視覚特徴の重み付けや価値評価プロセスに違いがある可能性がある。

本研究は、DNN を用いた計算論的アプローチにより、食品画像からの主観的価値計算における視覚情報処理の階層構造を明らかにした。fMRI データと DNN との比較では、初期視覚野(V1)においてノイズ上限値の 65-69% を説明できる一方、vmPFC では 4-8% に留まった。この結果は、現在の DNN モデルが視覚処理は捉えられるが、価値計算の神経基盤は十分に捉えられていないことを示している。ConvNeXt は視覚処理階層に沿った対応(初期層-V1、後期層-vmPFC)を示し、CLIP は中間層で情動・言語関連領域(扁桃体、島皮質、IFG、左角回)との対応を示した。CLIP モデルが最も高い予測精度( $r = 0.78$ )と脳活動説明力を示したことは、視覚-言語統合が食品評価において重要な役割を果たすことを示唆している。これらの知見は、食品選択の神経計算基盤の理解に貢献し、将来的には肥満や摂食障害などの食事関連疾患への応用が期待される。

## 7 付録

### A fMRI の仕組み

#### A.a 機能的磁気共鳴画像法(functional Magnetic Resonance Imaging; fMRI)

fMRI は、脳活動に伴う血流動態の変化を非侵襲的に測定する神経画像法である。fMRI は、強力な磁場を利用して脳の血液酸素レベル依存信号(Blood Oxygen Level-Dependent; BOLD 信号)と呼ばれる血液酸素量から神経活動を間接的に測定する技術である(Scott et al., 2016)。以下、本セクションの技術的説明は同文献に基づく。

#### A.b MRI 信号の仕組み: 核磁気共鳴と緩和現象

MRI は、強力な静磁場(通常 1.5~7 テスラ)中に置かれた陽子(プロトン)の核磁気共鳴させることで、電磁場を発生させそれを MR 信号として受信する。fMRI での静磁場とは、磁場が広い領域均一であり(空間的に均一)、磁場が時間によって変化しない(時間的に均一)である磁場のことを指す。静磁場中でプロトンは磁場方向に整列するが、この状態にラジオ波コイルでラーモア周波数(Larmor frequency)を照射すると、プロトンはエネルギーを吸収して励起状態へ遷移する。ラーモア周波数とは、プロトンを共鳴することができる周波数のことである。RF パルスを停止すると、プロトンは平衡状態(元の状態)に戻る過程で電磁波を放出する。この信号を検出して画像化するのが MRI の基本原理である。

プロトンが平衡状態に戻る過程は緩和と呼ばれ、T1 緩和と T2 緩和の 2 つの独立した過程がある。T1 緩和(縦緩和または spin-lattice relaxation)は、励起されたプロトンが周囲の格子(lattice)にエネルギーを放出して、縦磁化(静磁場方向の磁化)が回復する過程である。T1 緩和時間は組織によって異なり、灰白質では約 1 秒、白質では約 800 ミリ秒程度である。一方、T2 緩和(横緩和または spin-spin relaxation)は、プロトン間の相互作用により位相がずれ、静磁場に垂直な磁化が減衰する過程である。T2 緩和時間は組織の微細構造や水分含有量に依存し、灰白質では約 100 ミリ秒、白質では約 80 ミリ秒程度である。このため、T1 画像は T1 緩和時間の違いを利用して組織コントラストを生成し、脂肪含有量の多い組織(例えば白質)が高信号となる。一方、T2 画像は T2 緩和時間の違いを利用し、水分含有量の多い組織(例えば脳脊髄液)が高信号となる。

さらに、T2\*緩和は T2 緩和に加えて、局所的な磁場不均一性による信号減衰を含む。T2\*緩和時間は T2 緩和時間よりも短く、組織の磁気的性質や血液中の酸素化状態に敏感である。この特性を利用して、fMRI では T2\*強調画像を取得し、BOLD 信号を検出する。

#### A.c BOLD 信号の生成メカニズム

神経活動が生じると、活動領域への血流が増加し、デオキシヘモグロビンがオキシヘモグロビンに置き換わる。オキシヘモグロビン(oxyHb)は反磁性であるのに対し、デオキシヘモグロビン(deoxyHb)は常磁性である。この磁気的性質の違いにより、局所的な磁場の均一性が変化し、T2\*緩和時間に影響を及

ぼす。具体的には、神経活動に伴ってデオキシヘモグロビンの相対濃度が減少すると、局所磁場の均一性が高まり、T<sub>2\*</sub>緩和時間が延長する結果、MR信号強度が増加する。この信号変化が BOLD 信号として検出される。

fMRI データは、脳全体を 3 次元的に分割したボクセル(voxel: 体積要素)ごとに時系列の MR 信号として取得される。EPI シーケンスにより、脳全体のボクセル信号を 1~3 秒ごとに繰り返し取得することで、時間的に変動する BOLD 信号を記録する。

#### A.d 血行動態応答と時間特性

BOLD 信号は神経活動に対して間接的な指標であり、血行動態応答(hemodynamic response)を反映している。神経活動の開始後、BOLD 信号は約 2 秒の遅延を経て上昇し始め、4~6 秒後にピークに達する。その後、信号は徐々に減少し、ベースライン付近で一時的なアンダーシュート(post-stimulus undershoot)を示した後、約 12~20 秒で元のレベルに戻る。この時間経過は血行動態応答関数(hemodynamic response function; HRF)によって、イベントなどをモデル化することができる。この関数から、統計解析を行うことで、BOLD 信号変化に対するイベントなどが与える効果を推定できる。

fMRI の空間解像度は通常 1~3 mm 程度であり、時間解像度は 1~3 秒程度である。この時空間解像度により、fMRI は非侵襲的に脳の特定領域における活動パターンを測定することができるため、ヒトの認知機能や感覚処理、意思決定などの神経基盤を解明するために広く利用されている。

#### B ニューラルネットワークモデルの構造

ニューラルネットワーク(Neural Network; NN)は、生物の神経系に着想を得た計算モデルであり、情報処理やパターン認識に広く用いられている。NN は、複数の層にわたる相互接続されたノード(ニューロン)で構成され、各ノードは入力信号を受け取り、重み付けされた和を計算し、活性化関数を通じて出力信号を生成する。

ニューラルネットワークは、入力層、中間層(隠れ層)、出力層の 3 つの主要な層で構成される。入力層は外部からのデータを受け取り、中間層は複数のニューロンで構成され、入力データの特徴を抽出し、非線形変換を行う。出力層は最終的な予測や分類結果を生成する。各ニューロンは、前の層からの入力信号に対して重みを適用し、バイアス項を加えた後、活性化関数を通じて出力信号を生成する。活性化関数には、シグモイド関数、ReLU 関数、ソフトマックス関数などがある。活性化関数は、各ニューロンの重みを一定の範囲に制限することによって、NN の計算量が増加するのを防ぐ役割を果たす。

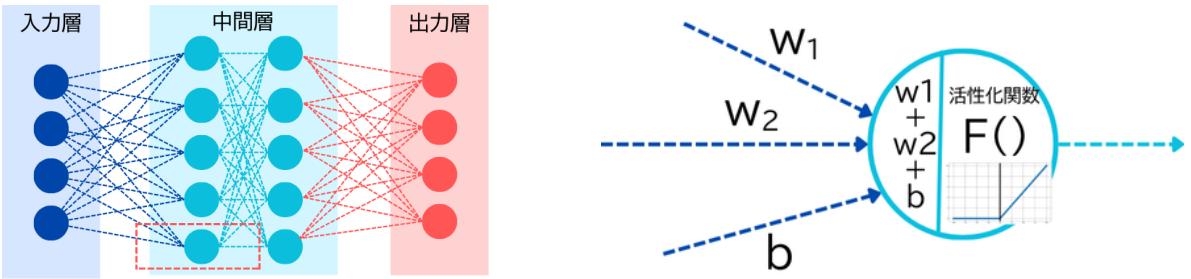


図 A: ニューラルネットワークの基本構造。入力層、中間層(隠れ層)、出力層から構成される。各ニューロンは、前の層からの入力信号に対して重みを適用し、バイアス項を加えた後、活性化関数を通じて出力信号を生成する。

図 B: ニューラルネットワークにおける代表的な活性化関数の例。左からシグモイド関数、ReLU 関数、ソフトマックス関数。各関数は、ニューロンの出力信号を生成するために使用される。

ニューラルネットワークの訓練は、主に教師あり学習に基づいて行われる。訓練データセットを用いて、ネットワークの重みとバイアスを最適化する。一般的な訓練手法として、誤差逆伝播法(Backpropagation)と勾配降下法(Gradient Descent)がある。誤差逆伝播法は、出力層で計算された誤差を中間層および入力層に逆伝播させ、各ニューロンの重みとバイアスの勾配を計算する。勾配降下法は、これらの勾配を用いて重みとバイアスを更新し、損失関数を最小化する。損失関数には、平均二乗誤差(Mean Squared Error; MSE)や交差エントロピー損失(Cross-Entropy Loss)などがある。訓練プロセスは、エポック(Epoch)と呼ばれる複数の反復で行われ、各エポックで全ての訓練データがネットワークに入力される。

## C ディープニューラルネットワークモデル(DNN)

### C.a CNN アーキテクチャ

畳み込みニューラルネットワーク(CNN)は、特に画像データの処理に優れた性能を発揮するディープラーニングモデルである。CNNは、畳み込み層、プーリング層、全結合層から構成されており、画像の特徴を自動的に抽出する能力を持つ。畳み込み層は、入力画像に対して畳み込みフィルターを適用し、特徴マップを生成する。プーリング層は、特徴マップの空間的次元を削減し、計算量を軽減するとともに、位置不变性を持たせる役割を果たす(Kriegeskorte, 2015)。全結合層は、最終的な分類結果を出力するために、抽出された特徴を用いている。また、CNNは多層構造で畳み込みを行う視覚野の階層的処理を模倣しており、初期層はエッジやテクスチャなどの低次特徴を抽出し、後続層はオブジェクトの形状やカテゴリなどの高次特徴を抽出する(Yamins et al., 2014); (Kriegeskorte, 2015)。

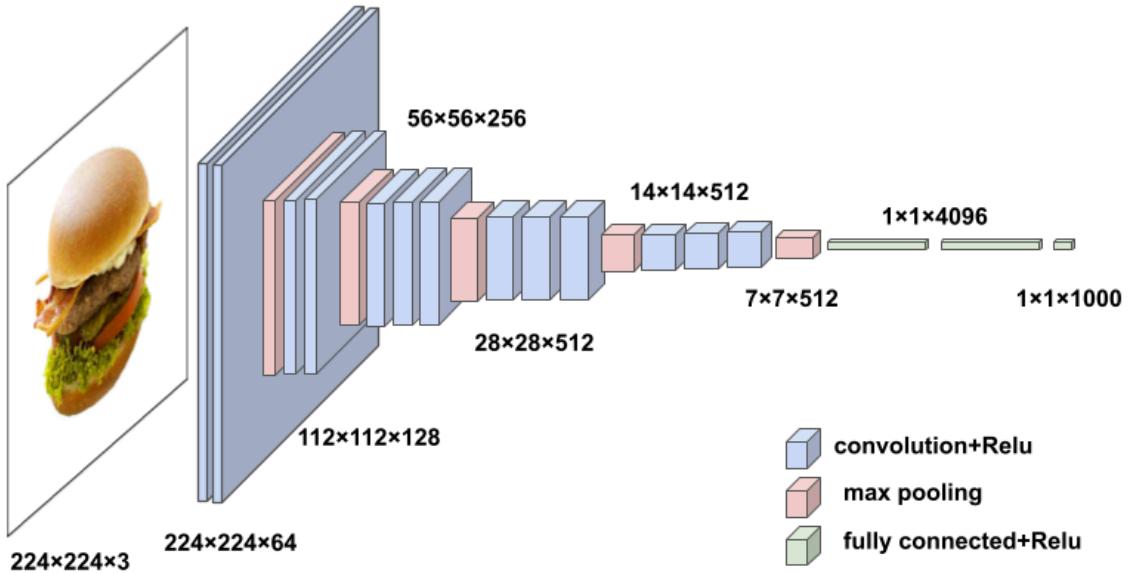


図 C: (Simonyan & Zisserman, 2015); から作成。畳み込みニューラルネットワーク(CNN)の基本構造の例(VGG16)。畳み込み層、プーリング層、全結合層から構成される。畳み込み層は入力画像に対して畳み込みフィルターを適用し、特徴マップを生成する。プーリング層は特徴マップの空間的次元を削減し、計算量を軽減するとともに、位置不变性を持たせる役割を果たす。全結合層は最終的な分類結果を出力するために、抽出された特徴を用いる。

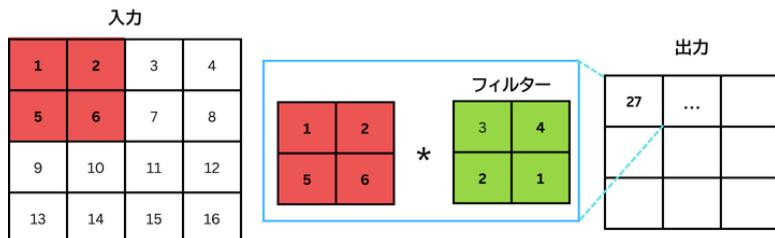


図 D: 畳み込み操作の例。入力画像に対して畳み込みフィルターを適用し、特徴マップを生成する。フィルターは画像の局所的なパターンを検出するために使用される。

### C.b Transformer アーキテクチャ

トランスフォーマー(Transformer)は、自然言語処理タスクにおいて高い性能を示すディープラーニングモデルであり、自己注意メカニズムを用いて入力シーケンス内の異なる位置の情報を動的に重み付けする。トランスフォーマーは、エンコーダーとデコーダーの2つの主要なコンポーネントで構成されており、エンコーダーは入力シーケンスを処理し、デコーダーは出力シーケンスを生成する。トランスフォーマーは、大規模なテキストデータセットで学習され、文の生成、質問応答、翻訳などのタスクで優れた性能を示している。

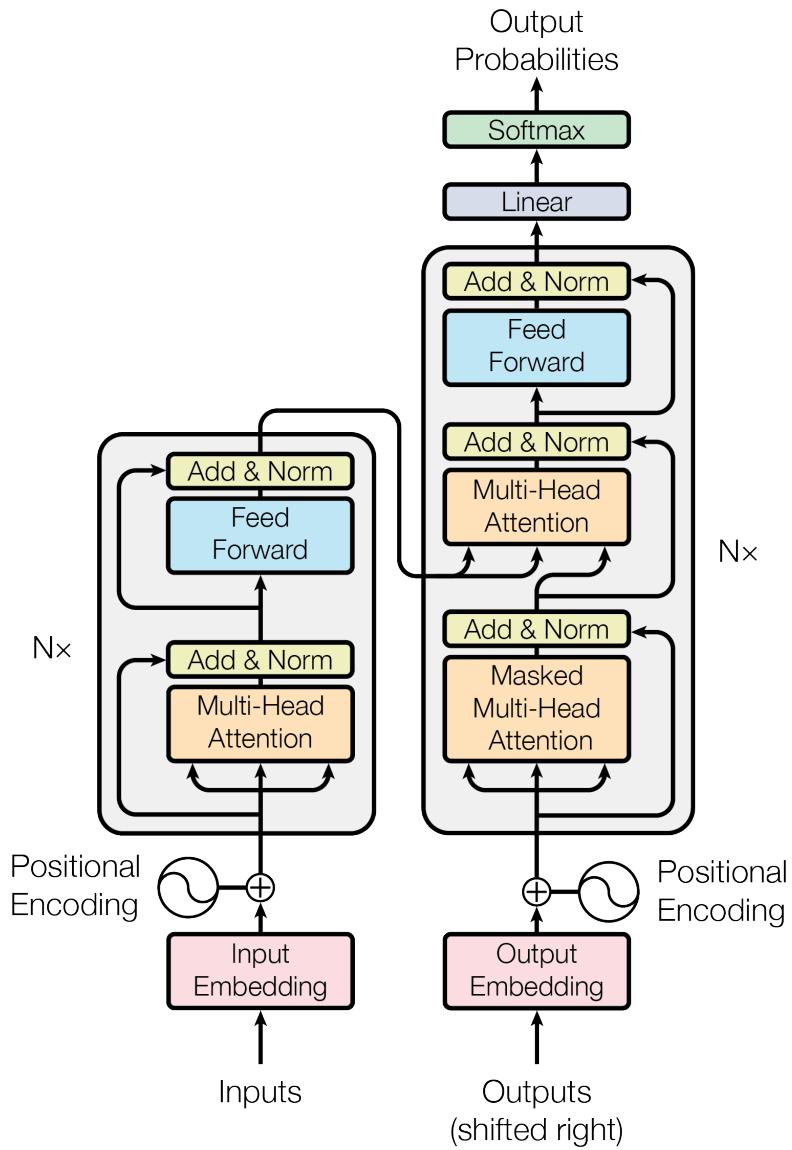


図 E: トランسفォーマーの基本構造。エンコーダーとデコーダーの 2 つの主要なコンポーネントで構成されており、エンコーダーは入力シーケンスを処理し、デコーダーは出力シーケンスを生成する。自己注意メカニズムを用いて入力シーケンス内の異なる位置の情報を動的に重み付けする (Vaswani et al., 2017, p.3, Figure 1);。

マルチヘッド自己注意 (Multi-Head Self-Attention; MHSA) は、トランسفォーマーの主要な構成要素であり、入力シーケンス内の異なる位置の情報を同時に処理する能力を持つ。MHSA は、複数の注意ヘッドを使用して、入力シーケンスの異なる部分に焦点を当てることができる。各注意ヘッドは、クエリ (Query)、キー (Key)、バリュー (Value) の 3 つのベクトルを生成し、これらを用いて注意重みを計算する。注意重みは、入力シーケンス内の各位置に対する重要度を表し、これを用いてバリューを加重平均することで、出力ベクトルを生成する。MHSA は、複数の注意ヘッドからの出力を結合し、線形変換を適用して最終的な出力を得る。例えば、エンコーダーでは、 $X$  を入力とすると、各注意ヘッド  $i$  に対して以下の計算が行われる。

$$Q_i = XW_i^Q, K_i = XW_i^K, V_i = XW_i^V$$

$$\text{head}_i = \text{Attention}(Q_i, K_i, V_i) = \text{softmax} \left( \frac{Q_i K_i^T}{\sqrt{d_k}} \right) V_i$$

ここで、 $W_i^Q, W_i^K, W_i^V$  はそれぞれクエリ、キー、バリューの重み行列であり、 $d_k$  はキーの次元数である。最終的な MHA の出力は以下のように計算される。

$$\text{MHA } (Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O$$

具体例を挙げると、入力シーケンスが「The cat sat on the mat」である場合、 $Q_i$  はその単語自身の情報、 $K_i$  は他の単語の情報、 $V_i$  は単語の実際の意味情報を保持している。そのため、「cat」が「sat」と強く関連していて、 $q_{\text{cat}}$  と  $k_{\text{sat}}$  の内積が大きくなり、 $v_{\text{cat}}$  の意味情報が強調される。これらを複数のヘッドで同時に処理することにより、さまざまな角度からの意味情報を捉えることができる。

## D 各層の活性化パターン分析

層  $\ell$  ( $\ell = 1, 2, \dots, L$ ) に対して、 $i$  番目の画像 ( $i = 1, 2, \dots, N$ ) を入力したときの **フラット化された活性化ベクトル** を

$$\mathbf{h}_\ell^i \in \mathbb{R}^{d_\ell}$$

とする。ここで、 $d_\ell$  は層  $\ell$  のフラット化後の次元数。これらを  $N$  枚分を並べた行列を

$$\mathbf{H}_\ell = \begin{pmatrix} (\mathbf{h}_\ell^1)^T \\ (\mathbf{h}_\ell^2)^T \\ \vdots \\ (\mathbf{h}_\ell^N)^T \end{pmatrix} \in \mathbb{R}^{N \times d_\ell}$$

と定義する。

### D.a PCA による次元削減

各層  $\ell$  ごとに活性化パターン行列  $\mathbf{H}_\ell$  に対して主成分分析(PCA)を実施し、累積寄与率が 80% に達するまで主成分を選択した。選択された主成分の数を  $M_\ell$  とし、対応する射影行列を  $\mathbf{W}_{\ell, \text{PCA}} \in \mathbb{R}^{d_\ell \times M_\ell}$  と定義する。固有値が大きい上位  $M_\ell$  個の主成分を用いて、データを低次元空間へ射影する。各次元の平均を 0 にするために、 $\mathbf{H}_\ell$  の各列から対応する列平均を引いて中心化をする。中心化した行列  $\tilde{\mathbf{H}}_\ell$  の共分散行列  $C$  を 固有値分解あるいは特異値分解(SVD)により対角化する。

$$C = \frac{1}{n} \tilde{\mathbf{H}}_\ell^T \tilde{\mathbf{H}}_\ell = \mathbf{W} \Lambda \mathbf{W}^T,$$

ここで、 $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_{d_\ell})$  は固有値(分散の大きさ)を対角に並べた対角行列、 $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_{d_\ell}]$  の各列  $\mathbf{w}_i$  は対応する固有ベクトル(主成分ベクトル)を意味する。

$$\mathbf{Z}_\ell = \tilde{\mathbf{H}}_\ell \mathbf{W}_{\ell, \text{PCA}} \in \mathbb{R}^{N \times M_\ell}$$

これにより、 $\mathbb{R}^{d_\ell}$  から  $\mathbb{R}^{M_\ell}$  ( $M_\ell \ll d_\ell$ ) へと次元削減が行われる。

### D.b Ridge 回帰による画像特徴予測

各画像属性( $y_i$  の各成分)ごとに 層  $\ell$  の特徴量  $\tilde{\mathbf{H}}_\ell$  を用いて **Ridge 回帰** を行い、回帰係数を求める。すなわち、属性  $k$  に対応する目的変数を

$$\mathbf{y}^k = \begin{pmatrix} y_{1,k} \\ y_{2,k} \\ \vdots \\ y_{N,k} \end{pmatrix} \quad (\in \mathbb{R}^N)$$

とおき、その回帰係数ベクトル(あるいは行列)を  $\mathbf{b}_\ell^k$  とする。すると、属性  $k$  に対する Ridge 回帰の目的関数は

$$\min_{\mathbf{b}_\ell^k} \left\| \mathbf{y}^k - \tilde{\mathbf{H}}_\ell \mathbf{b}_\ell^k \right\|_2^2 + \lambda \left\| \mathbf{b}_\ell^k \right\|_2^2$$

となる( $\|\cdot\|_2$  はユークリッドノルム,  $\lambda \geq 0$  は正則化パラメータ)。画像属性ごとに別々に回帰を行うことで、それぞれの属性に合わせた係数を推定した。

画像を学習データとテストデータに分け、学習データで8分割クロスバリデーションのグリットサーチを行い最適な正則化パラメータを求めた。正則化パラメータで学習したモデルで層ごとに予測精度を算出し、どの層がどの特徴をよく表現しているかを評価した。予測精度はピアソンの相関係数を使用した。栄養属性(protein\_100g, fat\_100g, carbs\_100g, grams\_total)については、一部の画像で値が欠損していたため、これらの属性の予測では欠損値を含む画像を除外して分析を行った。

## E fMRIにおける一般化線型モデル(GLM)分析

fMRI データに対して一般化線型モデル(Generalized Linear Model; GLM)を適用し、各ボクセルの BOLD 信号変動を説明することを提案した。(Friston et al., 1995)

GLM は、以下の形式で表される。

$$\mathbf{y}_i = \mathbf{x}_i \beta + e_i$$

ここで、 $\mathbf{y}_i$  はボクセル  $i$  の BOLD 信号の時系列データ、 $\mathbf{x}_i$  は設計行列、 $\beta$  は回帰係数ベクトル、 $e_i$  は誤差項である。デザイン行列  $\mathbf{x}_i$  には、実験条件や刺激イベントに対応する説明変数が含まれ、各説明変数は血行動態応答関数(Hemodynamic Response Function; HRF)で畳み込まれる。

ここで、自己相関を考えるために

$$e_i \sim N(0, \sigma_i^2 V)$$

と仮定する。 $V$  は自己相関を表す共分散行列であり、 $\sigma_i^2$  はボクセル  $i$  の誤差の分散である。 $\sigma_i^2 V$  を詳しく見ると

$$\sigma^2 V = \begin{pmatrix} \sigma^2 V_{11} & \sigma^2 V_{12} & \dots \\ \sigma^2 V_{21} & \sigma^2 V_{22} & \dots \\ \vdots & \vdots & \ddots \end{pmatrix}$$

のように表され、 $V_{i,j}$  は時点  $i$  と時点  $j$  の自己相関を示す要素である。 $V$  を推定した後、GLM の自己相関を考慮したモデルは  $W = V^{-\frac{1}{2}}$  とすると以下のように変形できる。

$$W \mathbf{y}_i = W \mathbf{x}_i \beta + W e_i$$

これは、独立同一分布を持つ誤差項に変換するために  $W = V^{-\frac{1}{2}}$  をかけたものであり、以下のように誤差項の共分散が単位行列になる。

$$\text{Cov}(We_i) = W \cdot \text{Cov}(e_i) \cdot W^T = V^{-\frac{1}{2}} \cdot (\sigma_i^2 V) \cdot V^{-\frac{1}{2}} = \sigma_i^2 I$$

SPM では  $V$  の推定は、ReML(制限付き最大尤度法)で行われる。GLM のパラメータ推定には、最小二乗法や最大尤度法が用いられる。推定された回帰係数  $\beta$  は、各説明変数が BOLD 信号に与える影響を示し、統計的検定を通じて有意性が評価される。

## F 集団解析におけるランダム効果モデル

### F.a ランダム効果解析(Random Effects Analysis)

集団解析では、被験者間の変動を考慮するためにランダム効果モデル(Random Effects Model; RFX)を適用した。RFX モデルは、各被験者のデータを個別に解析した後、被験者間の変動を考慮して集団レベルでの統計的検定を行う手法である。被験者ごとの回帰係数を  $\beta_i$  とすると、RFX モデルは以下のように表される。

$$\mathbf{y}_i = \mathbf{X}_i \beta_i + \mathbf{e}_i$$

$$\beta_i = \beta_{\text{all}} + \mathbf{u}_i$$

ここで、 $\mathbf{y}_i$  は被験者  $i$  の BOLD 信号の時系列データ、 $\mathbf{X}_i$  は設計行列、 $\beta_i$  は被験者  $i$  の回帰係数、 $\mathbf{e}_i$  は誤差項、 $\beta_{\text{all}}$  は集団レベルの平均回帰係数、 $\mathbf{u}_i$  は被験者間のランダム効果を表す。

### F.b 要約統計量を使ったランダム効果解析

RFX 解析では、各被験者の解析結果から要約統計量を抽出し、集団レベルでの解析に用いる。具体的には、各被験者の回帰係数  $\beta_i$  とその推定誤差の分散  $\sigma_{\beta_i}^2$  を計算する。これらの要約統計量を用いて、集団レベルでの統計的検定を行う。母集団平均  $\beta_{\text{all}}$  は、被験者ごとの要約統計量の平均として計算される

$$\hat{\beta}_{\text{all}} = \frac{1}{N} \sum_{i=1}^N \bar{\beta}_i$$

この推定値の分散  $\text{Var}(\hat{\beta}_{\text{all}})$  は、以下のように導出される。

$$\text{Var}(\hat{\beta}_{\text{all}}) = \frac{\sigma_b^2}{N} + \frac{\sigma_{\beta}^2}{Nn}$$

この結果が、全データを用いた最大尤度(ML)推定による分散と完全に一致することが知られている(Penny & Holmes, 2007)。

## G 多重比較補正

### G.a ファミリー・ワイズ・エラー率(Family-Wise Error Rate; FWE)補正

fMRI データの解析では、多数のボクセルに対して統計的検定を行うため、多重比較問題が生じる。これにより、偽陽性率(Type I error rate)が増加する可能性がある。ファミリー・ワイズ・エラー率(Family-Wise Error Rate; FWE)補正は、全体の誤検出率を制御するための手法であり、Bonferroni 補正やランダムフィールド理論(Random Field Theory; RFT)に基づく方法などがある。

$$FWE = P(\text{全ての検定の中で少なくとも 1 つの偽陽性} \mid \text{帰無仮説が真})$$

すなわち、FWE は帰無仮説が真である場合に、全ての検定の中で少なくとも 1 つの偽陽性が発生する確率を表す。

#### G.a.a ランダムフィールド理論(Random Field Theory; RFT)に基づく FWE 補正

RFT に基づく FWE 補正は、空間的に連続した fMRI データの特性を考慮し、統計マップにおけるピークの分布をモデル化する手法である。RFT は、統計場がガウス過程として近似できることを仮定し、ピークの高さやクラスタサイズに基づいて有意性を評価する(Worsley et al., 1996)。これにより、ボクセルレベルおよびクラスターレベルでの FWE 補正が可能となる。

$$P(\max Z > u) \approx \sum_{d=0}^3 R_{D(V)} \rho_{D(t)}$$

ここで、 $u$  は観測された統計量の閾値、 $R_{D(V)}$  はある領域  $V$  におけるレセル数、 $\rho_{D(t)}$  は期待エウラー標数と呼ばれる。その領域で  $t$  を偶然越えるレセルの期待値密度である。また、レセル数は以下のように計算される。

$$R_{D(V)} = \frac{V}{\text{FWHM}_x * \text{FWHM}_y * \text{FWHM}_z}$$

ここで、FWHM は半値全幅(Full Width at Half Maximum)を表し、空間の滑らかさの程度を示す。

FWE 補正された  $p$  値は、以下のように計算される。

$$p_{\text{FWE}} = P(\max Z > z_{\text{obs}})$$

ここで、 $z_{\text{obs}}$  は観測された統計量の値である。

#### G.a.b ボクセルレベル FWE 補正

ボクセルレベル FWE 補正では、各ボクセルに対して独立に統計的検定を行い、FWE 補正された  $p$  値を計算する。RFT に基づく方法では、観測された統計量が閾値を超える確率を評価し、全体の誤検出率を制御する。

#### G.a.c クラスターレベル FWE 補正

クラスターレベル FWE 補正では、連続したボクセルのクラスタに対して統計的検定を行い、クラスタサイズに基づいて FWE 補正された  $p$  値を計算する。ボクセルレベル FWE 補正では、ボクセル単位での検定を行うのに対し、クラスターレベル FWE 補正では、空間的に連続したボクセル群(クラスタ)に対して検

定を行う点が異なる。そのため、クラスターレベル FWE 補正は、空間的に広がった効果を検出するのに適している。RFTに基づく方法では、クラスタの大きさが偶然に観測される確率を評価し、全体の誤検出率を制御する。

$$p_{\text{FWE, cluster}} = P(\max \text{ cluster size} > k_{\text{obs}})$$

ここで、 $k_{\text{obs}}$  は観測されたクラスタサイズである。

#### **G.a.d Small Volume Correction (SVC)**

Small Volume Correction (SVC)は、特定の関心領域(Region of Interest; ROI)に対して FWE 補正を行う手法である。SVC では、全脳解析に比べて検定の数が減少するため、より厳密な FWE 補正が可能となる。SVC は、事前に定義された ROI に基づいて、統計マップ内のボクセルに対して FWE 補正を適用する。これにより、特定の脳領域における効果の有意性を評価することができる。

$$p_{\text{SVC}} = P(\max Z > z_{\text{obs}} \mid \text{within ROI})$$

## **H 表現類似性解析(RSA)**

表現類似性解析(Representational Similarity Analysis; RSA)は、異なるシステム(例:脳と DNN)間の表現構造を比較するための手法である(Kriegeskorte et al., 2008)。RSA では、刺激セット内の各ペア間の類似度(または非類似度)を行列として表現し、この行列間の相関を計算することで表現構造の類似性を定量化する。

#### **H.a 表現非類似度行列(RDM)**

表現非類似度行列(Representational Dissimilarity Matrix; RDM)は、 $N$ 個の刺激に対する応答パターン間の非類似度を $N \times N$ の対称行列として表現したものである。RDM の各要素 $(i, j)$ は、刺激*i*と刺激*j*に対する応答パターン間の非類似度を示す。

脳活動の場合、各刺激に対するボクセルパターン  $\mathbf{b}_i \in \mathbb{R}^V$  ( $V$ はボクセル数)から、非類似度は以下のように計算される:

$$\text{RDM}_{i,j} = 1 - \text{corr}(\mathbf{b}_i, \mathbf{b}_j)$$

ここで、corrはピアソン相関係数である。同様に、DNN の各層についても、刺激に対する活性化パターンから RDM を算出する。

#### **H.b 二重中心化(Double centering)**

二重中心化は、RDM から行平均・列平均・全体平均を除去する前処理である。RDM を  $D$  とすると、二重中心化された RDM  $\tilde{D}$  は以下のように計算される:

$$\tilde{D}_{i,j} = D_{i,j} - \bar{D}_{i\cdot} - \bar{D}_{\cdot j} + \bar{D}_{..}$$

ここで、 $\bar{D}_{i\cdot}$  は行*i*の平均、 $\bar{D}_{\cdot j}$  は列*j*の平均、 $\bar{D}_{..}$  は全体平均である。

二重中心化により、RDM の行和・列和がすべて 0 となり、絶対的なスケールではなく相対的なパターン構造のみを比較することが可能になる。この処理は、Centered Kernel Alignment(CKA)と等価であり、表現の類似性をより厳密に評価できる(Williams, 2024)。

### H.c ノイズ上限値(Noise Ceiling)

ノイズ上限値は、被験者間の一致度から推定される理論的な説明力の上限である。Leave-One-Out 法によるノイズ上限値は以下のように算出される：

- **NC 上限**: 各被験者の RDM と全被験者平均 RDM との相関の平均値
  - **NC 下限**: 各被験者の RDM と当該被験者を除いた残りの被験者の平均 RDM との相関の平均値
- NC 上限比(モデル相関 / NC 上限 × 100)により、理論的に達成可能な最大説明力に対するモデルの到達度を評価する。NC 上限比が 100% に近いほど、モデルが被験者間で共有される表現構造を完全に説明していることを示す。

## I 3 レベル階層的 PCA 分析

DNN の各層の活性化パターンと fMRI データを比較するため、3 レベル階層的 PCA を用いた手法を開発した。この手法は、DNN の活性化パターンを直交化された階層構造に分解し、各レベルの独立した説明力を評価可能とする。

### I.a 層グループの定義

DNN の層を情報処理の階層段階に基づいて 4 つのグループに分類した：

- **Initial 層**(初期層): 低次視覚特徴(エッジ、テクスチャ)を抽出する初期層
- **Middle 層**(中間層): 中次特徴を抽出する中間層
- **Late 層**(後期層): 高次特徴(オブジェクトの形状やカテゴリ)を抽出する後期層
- **Final 層**(最終層): 最終的な分類や埋め込み表現を生成する層

各層グループ  $g \in \{1, 2, 3, 4\}$  に対して、グループ内の全層の活性化ベクトルを連結した行列を  $\mathbf{H}_g \in \mathbb{R}^{N \times d_g}$  とする( $N$  は画像数、 $d_g$  はグループ  $g$  の総次元数)。

### I.b 3 レベル階層的 PCA

#### レベル 1:Global PC(全層共通成分)

まず、全層グループの活性化を連結した行列  $\mathbf{H}_{\text{all}} = [\mathbf{H}_1, \mathbf{H}_2, \mathbf{H}_3, \mathbf{H}_4]$  に対して PCA を適用し、全層に共通する分散を捕捉する主成分を抽出する。累積寄与率が 60% に達するまでの主成分を選択し、これを Global PC とする。

$$\mathbf{Z}_{\text{Global}} = \tilde{\mathbf{H}}_{\text{all}} \mathbf{W}_{\text{Global}} \in \mathbb{R}^{N \times M_{\text{Global}}}$$

ここで、 $\mathbf{W}_{\text{Global}}$  は Global PC への射影行列、 $M_{\text{Global}}$  は Global PC の次元数である。

#### レベル 2:Layer-Shared PC(隣接層間共有成分)

隣接する層グループ間で共有される分散を捕捉するため、正準相関分析(CCA)を用いる。まず、各層グループの活性化から Global 成分を除去した残差を計算する：

$$\mathbf{R}_g = \tilde{\mathbf{H}}_g - \mathbf{Z}_{\text{Global}} \mathbf{B}_g$$

ここで、 $\mathbf{B}_g$  は層グループ  $g$  の活性化を Global PC で回帰した係数行列である。

次に、隣接する層グループペア  $(g, g+1)$  に対して CCA を適用し、共有成分を抽出する：

$$\max_{\mathbf{a}, \mathbf{b}} \text{corr}(\mathbf{R}_g \mathbf{a}, \mathbf{R}_{g+1} \mathbf{b})$$

各隣接ペアについて上位 2 成分を抽出し、計 6 成分(3 ペア × 2 成分)を Layer-Shared PC とする。

### レベル 3:Layer-Specific PC(層グループ固有成分)

各層グループに固有の分散を捕捉するため、Global 成分と Shared 成分を除去した残差に対して PCA を適用する：

$$\mathbf{R}'_g = \mathbf{R}_g - \sum_{j \in N(g)} \mathbf{Z}_{\text{Shared},j} \mathbf{C}_{g,j}$$

ここで、 $N(g)$  は層グループ  $g$  に関連する Shared 成分の集合、 $\mathbf{C}_{g,j}$  は対応する回帰係数である。残差  $\mathbf{R}'_g$  に対して PCA を適用し、累積寄与率が 50% に達するまでの主成分を各グループについて抽出する。

### I.c 直交化の意義

この 3 レベル階層的 PCA により、以下の直交構造が保証される：

1. Global PC、Layer-Shared PC、Layer-Specific PC は互いに直交
2. 各レベル内の成分も互いに直交
3. 異なる層グループの Layer-Specific PC 同士も直交

この直交化により、GLM において各レベル・各層グループの独立した説明力を評価可能となる。従来の PCA のみを用いた場合、層間で共有される分散と層固有の分散が混在し、各層の独自の寄与を分離できない問題があった。

### I.d GLM への適用

これらの主成分を画像提示時のパラメトリックモジュレータとして GLM に投入する。パラメトリックモジュレータは日ごとにまとめ(3 セッション)、画像提示時の定数項はランごとに設定した。SPM の自動直交化は無効化し( $\text{orth} = 0$ )、事前に直交化された主成分構造を保持する。

重要な点として、DNN の活性化から抽出した主成分スコアは被験者全体で共通の変数として使用した。これは、同一の食品画像に対する DNN の活性化パターンは被験者間で同一であり、各画像に対して一意の主成分スコアが割り当てられるためである。すなわち、各被験者の複数セッションにわたって提示された同一画像には同一の主成分スコアが適用される。

頭部運動パラメータ(6 パラメータ)はランごとに共変量として含めた。統計的検定のためのコントラストはセッション間で平均化した。

統計検定では、各レベルおよび各層グループについて F 検定を用いて説明力を評価する：

- **Global\_F**: 全ての Global PC の説明力
  - **Shared\_F**: 全ての Layer-Shared PC の説明力
  - **Initial\_F, Middle\_F, Late\_F, Final\_F**: 各層グループ固有の Layer-Specific PC の説明力
- 有意な脳領域を同定するため、クラスターレベル FWE 補正(クラスター形成閾値: $p < 0.001$  uncorrected、クラスターレベル FWE: $p < 0.05$ )を適用する。

## J 補足:事前学習済み ConvNeXt の層別情報表現

事前学習済み(ファインチューニングなし)の ConvNeXt-Base モデルにおける各層の情報表現を分析した結果を示す。本文中の図 4 ではファインチューニング後の ConvNeXt モデルを示しているが、ここでは比較のため事前学習済みモデルの結果を示す。

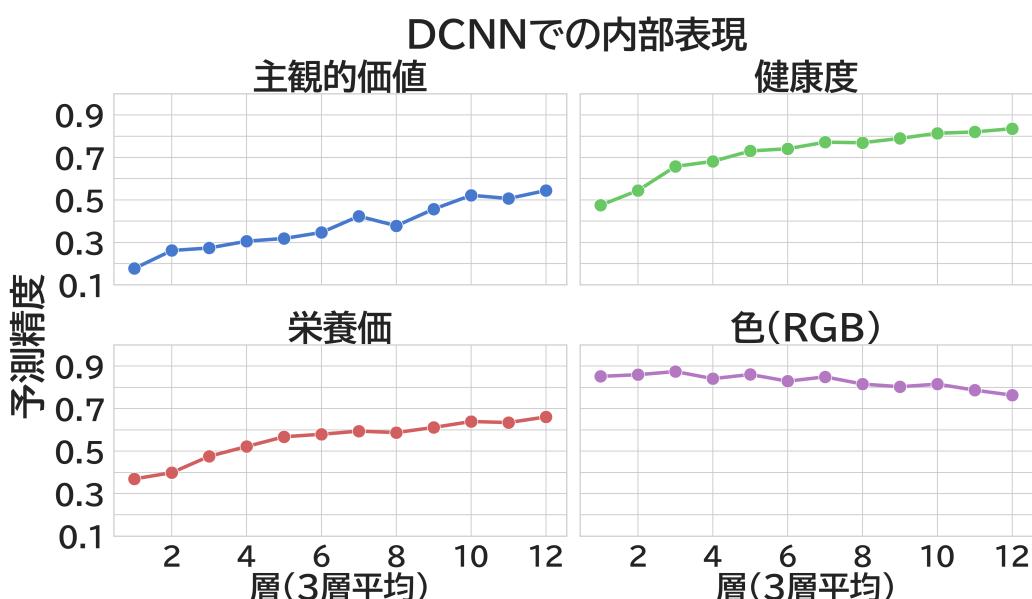


図 F: 事前学習済み ConvNeXt-Base の各層における情報表現。各線は、各属性(主観的価値、健康度、栄養価、色)の予測と実際の評価との相関係数を示す。ファインチューニング後のモデル(図 4)と比較して、全体的に同様の傾向が見られるが、主観的価値の予測精度は後期層でも約 0.55 程度に留まる。

## K 補足:DNN 比較分析 ROI 効果量

効果量を RMS( $\beta$  値の二乗平均平方根)で示す。エラーバーは被験者間の標準誤差(SEM)。

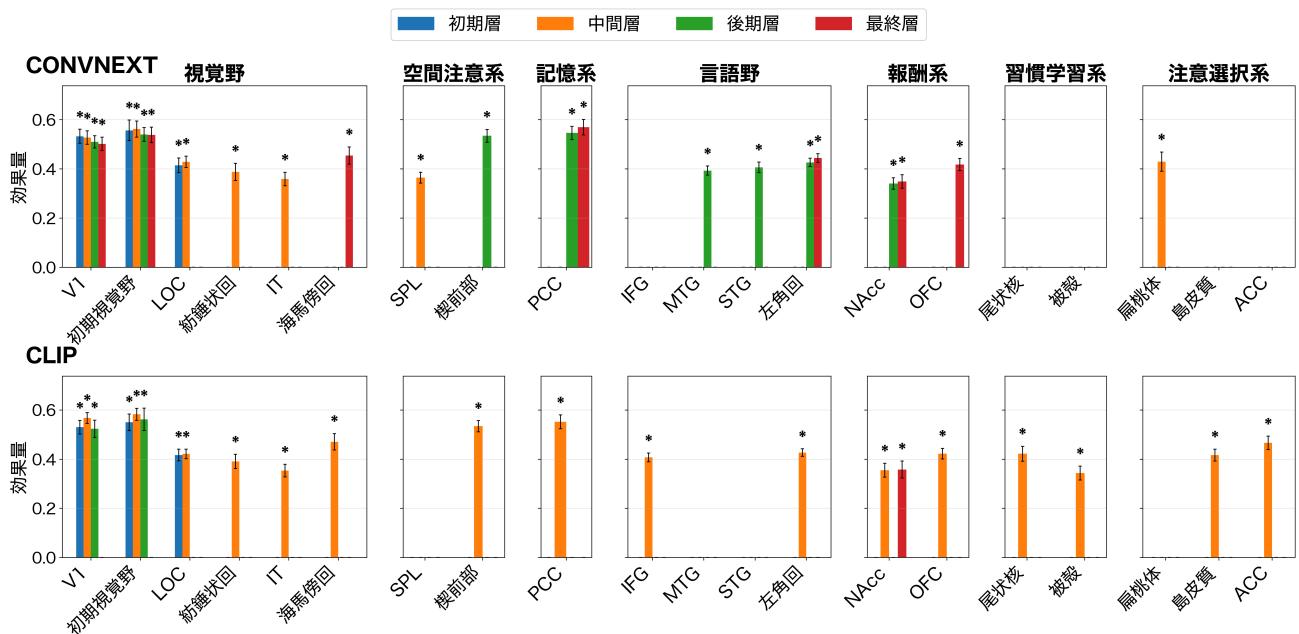


図 G: 各層固有成分の ROI 効果量。\*は有意(SVC FWE 補正  $p < 0.05$ )。

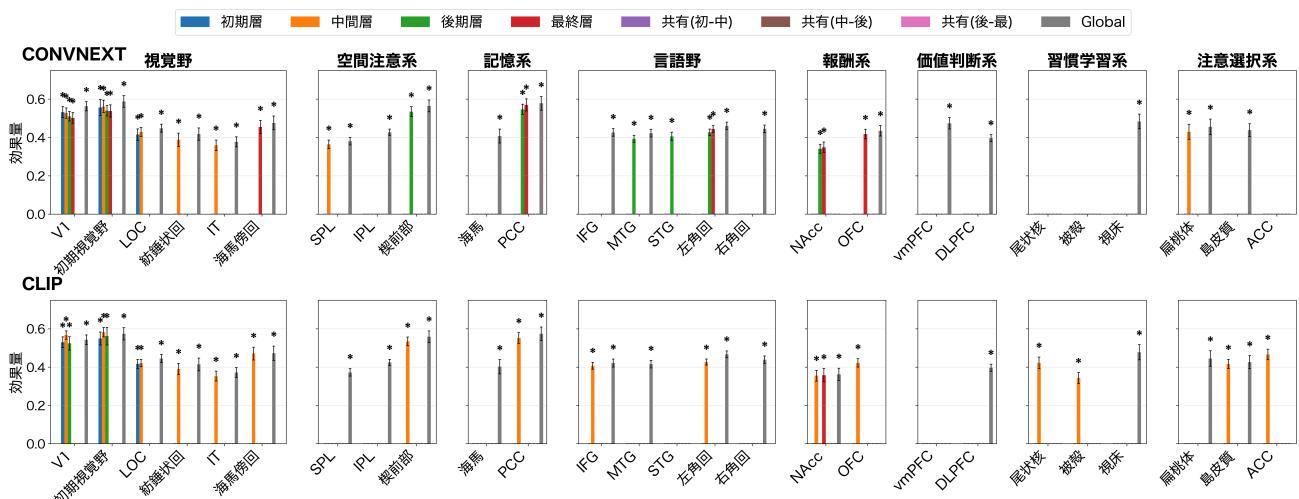


図 H: 各層固有成分・共有成分・Global 成分の ROI 効果量。共有(初-中)は初期-中間層間、共有(中-後)は中間-後期層間、共有(後-最)は後期-最終層間の共有成分。\*は有意(SVC FWE 補正  $p < 0.05$ )。

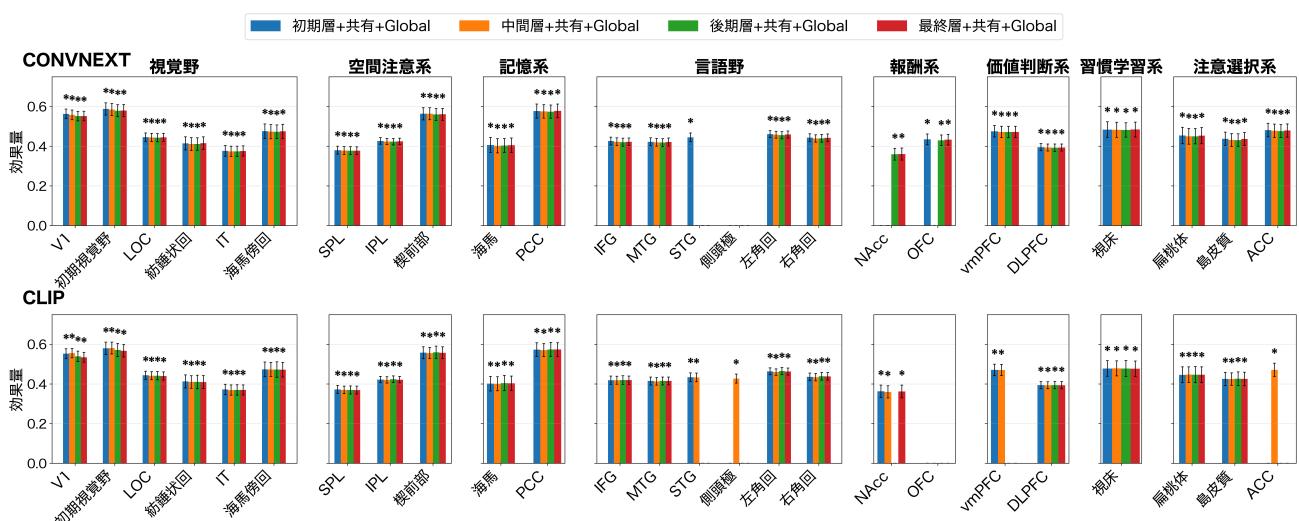


図 I: 各層+共有+Global 成分の ROI 効果量。各層固有成分に関する共有成分および Global 成分を加えた効果量。\*は有意(SVC FWE 補正  $p < 0.05$ )。

# 引用文献

- Blechert, J., Lender, A., Polk, S., Busch, N. A., & Ohla, K. (2019). Food-Pics\_Extended—An Image Database for Experimental Research on Eating and Appetite: Additional Images, Normative Ratings and an Updated Review. *Frontiers in Psychology*, 10. <https://doi.org/10.3389/fpsyg.2019.00307>
- Blechert, J., Meule, A., Busch, N. A., & Ohla, K. (2014). Food-pics: an image database for experimental research on eating and appetite. *Frontiers in Psychology*, 5. <https://doi.org/10.3389/fpsyg.2014.00617>
- Bowers, J. S., Malhotra, G., Dujmović, M., Llera Montero, M., Tsvetkov, C., Biscione, V., Puebla, G., Adolfi, F., Hummel, J. E., Heaton, R. F., Evans, B. D., Mitchell, J., & Blything, R. (2023). Deep problems with neural network models of human vision. *Behavioral and Brain Sciences*, 46, e385. <https://doi.org/10.1017/S0140525X22002813>
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., & others. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.
- Caplette, L., & Turk-Browne, N. B. (2024). Computational reconstruction of mental representations using human behavior. *Nature Communications*, 15(1), 4183. <https://doi.org/10.1038/s41467-024-48114-6>
- Caucheteux, C., Gramfort, A., & King, J.-R. (2023). Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nat Hum Behav*, 7(3), 430–441.
- Chen, H., Liu, B., Wang, S., & Bi, Y. (2025). Combined evidence from artificial neural networks and human brain-lesion models reveals that language modulates vision in human perception. *Nature Human Behaviour*. <https://doi.org/10.1038/s41562-025-02357-5>
- Chib, V. S., Rangel, A., Shimojo, S., & O'Doherty, J. P. (2009). Evidence for a common representation of decision values for dissimilar goods in human ventromedial prefrontal cortex. *J Neurosci*, 29(39), 12315–12320.
- Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, 3(3), 201–215. <https://doi.org/10.1038/nrn755>
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, , 248–255. <https://doi.org/10.1109/CVPR.2009.5206848>
- Doerig, A., Kietzmann, T. C., Allen, E., Wu, Y., Naselaris, T., Kay, K., & Charest, I. (2025). High-level visual representations in the human brain are aligned with large language models. *Nature Machine Intelligence*, 7(8), 1220–1234. <https://doi.org/10.1038/s42256-025-01072-0>

- Dwivedi, K., Bonner, M. F., Cichy, R. M., & Roig, G. (2021). Unveiling functions of the visual cortex using task-specific deep neural networks. *PLOS Computational Biology*, 17(8), e1009267. <https://doi.org/10.1371/journal.pcbi.1009267>
- Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., Kent, J. D., Goncalves, M., DuPre, E., Snyder, M., Oya, H., Ghosh, S. S., Wright, J., Durnez, J., Poldrack, R. A., & Gorgolewski, K. J. (2018). fMRIprep: a robust preprocessing pipeline for functional MRI. *Nat Methods*, 16(1), 111–116.
- Fedorenko, E., Ivanova, A. A., & Regev, T. I. (2024). The language network as a natural kind within the broader landscape of the human brain. *Nature Reviews Neuroscience*, 25(5), 289–312. <https://doi.org/10.1038/s41583-024-00802-4>
- Foerde, K., Steinglass, J. E., Shohamy, D., & Walsh, B. T. (2015). Neural mechanisms supporting maladaptive food choices in anorexia nervosa. *Nature Neuroscience*, 18(11), 1571–1573. <https://doi.org/10.1038/nn.4136>
- Foroni, F., Pergola, G., & Rumiati, R. I. (2016). Food color is in the eye of the beholder: the role of human trichromatic vision in food evaluation. *Scientific Reports*, 6, 37034. <https://doi.org/10.1038/srep37034>
- Frazier, J. A., Chiu, S., Breeze, J. L., Makris, N., Lange, N., Kennedy, D. N., Herbert, M. R., Bent, E. K., Koneru, V. K., Dieterich, M. E., Hodge, S. M., Rauch, S. L., Grant, P. E., Cohen, B. M., Seidman, L. J., Caviness, V. S., & Biederman, J. (2005). Structural brain magnetic resonance imaging of limbic and thalamic volumes in pediatric bipolar disorder. *American Journal of Psychiatry*, 162(7), 1256–1265. <https://doi.org/10.1176/appi.ajp.162.7.1256>
- Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J.-B., Frith, C. D., & Frackowiak, R. S. J. (1995). Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping*, 2(4), 189–210. <https://doi.org/10.1002/hbm.460020402>
- Glimcher, P. W., & Rustichini, A. (2004). Neuroeconomics: The Consilience of Brain and Decision. *Science*, 306(5695), 447–452. <https://doi.org/10.1126/science.1102566>
- Grabenhorst, F., Schulte, F. P., Maderwald, S., & Brand, M. (2013). Food labels promote healthy choices by a decision bias in the amygdala. *Neuroimage*, 74, 152–163. <https://doi.org/10.1016/j.neuroimage.2013.02.012>
- Grill-Spector, K., & Malach, R. (2004). The human visual cortex. *Annual Review of Neuroscience*, 27, 649–677. <https://doi.org/10.1146/annurev.neuro.27.070203.144220>
- Gross, J., Woelbert, E., Zimmermann, J., Okamoto-Barth, S., Riedl, A., & Goebel, R. (2014). Value signals in the prefrontal cortex predict individual preferences across reward categories. *Journal of Neuroscience*, 34(22), 7580–7586. <https://doi.org/10.1523/JNEUROSCI.5082-13.2014>

- Haber, S. N., & Knutson, B. (2010). The reward circuit: linking primate anatomy and human imaging. *Neuropsychopharmacology*, 35(1), 4–26. <https://doi.org/10.1038/npp.2009.129>
- Hampton, A. N., Bossaerts, P., & O'Doherty, J. P. (2006). The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. *Journal of Neuroscience*, 26(32), 8360–8367. <https://doi.org/10.1523/JNEUROSCI.1010-06.2006>
- Hare, T. A., Camerer, C. F., & Rangel, A. (2009). Self-control in decision-making involves modulation of the vmPFC valuation system. *Science*, 324(5927), 646–648. <https://doi.org/10.1126/science.1168450>
- Hare, T. A., O'Doherty, J., Camerer, C. F., Schultz, W., & Rangel, A. (2008). Dissociating the role of the orbitofrontal cortex and the striatum in the computation of goal values and prediction errors. *Journal of Neuroscience*, 28(22), 5623–5630. <https://doi.org/10.1523/JNEUROSCI.1309-08.2008>
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.
- Iigaya, K., Yi, S., Wahle, I. A., Tanwisuth, K., & O'Doherty, J. P. (2021). Aesthetic preference for art can be predicted from a mixture of low- and high-level visual features. *Nature Human Behaviour*, 5(6), 743–755. <https://doi.org/10.1038/s41562-021-01124-6>
- Iigaya, K., Yi, S., Wahle, I. A., Tanwisuth, S., Cross, L., & O'Doherty, J. P. (2023). Neural mechanisms underlying the hierarchical construction of perceived aesthetic value. *Nature Communications*, 14(1), 127. <https://doi.org/10.1038/s41467-022-35654-y>
- Kahnt, T., Heinze, J., Park, S. Q., & Haynes, J.-D. (2010). Decoding different roles for vmPFC and dlPFC in multi-attribute decision making. *Neuroimage*, 52(2), 506–514. <https://doi.org/10.1016/j.neuroimage.2010.04.229>
- Kravitz, D. J., Saleem, K. S., Baker, C. I., Ungerleider, L. G., & Mishkin, M. (2013). The ventral visual pathway: an expanded neural framework for the processing of object quality. *Trends in Cognitive Sciences*, 17(1), 26–49. <https://doi.org/10.1016/j.tics.2012.10.011>
- Kriegeskorte, N. (2015). Deep neural networks: A new framework for modeling biological vision and brain information processing [Journal Article]. *Annual Review of Vision Science*, 1(Volume 1, 2015), 417–446. <https://doi.org/10.1146/annurev-vision-082114-035447>
- Kriegeskorte, N., & Douglas, P. K. (2018). Cognitive computational neuroscience. *Nature Neuroscience*, 21(9), 1148–1160. <https://doi.org/10.1038/s41593-018-0210-5>
- 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>
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, & K. Weinberger (Eds.),

- Advances in Neural Information Processing Systems: Vol. 25. Advances in Neural Information Processing Systems.* [https://proceedings.neurips.cc/paper\\_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf)
- Lahner, B., Dwivedi, K., Iamshchinina, P., Graumann, M., Lascelles, A., Roig, G., Gifford, A. T., Pan, B., Jin, S., Ratan Murty, N. A., Kay, K., Oliva, A., & Cichy, R. (2024). Modeling short visual events through the BOLD moments video fMRI dataset and metadata. *Nature Communications*, 15(1), 6241. <https://doi.org/10.1038/s41467-024-50310-3>
- Levy, I., Lazzaro, S. C., Rutledge, R. B., & Glimcher, P. W. (2011). Choice from non-choice: Predicting consumer preferences from blood oxygenation level-dependent signals obtained during passive viewing. *Journal of Neuroscience*, 31(1), 118–125. <https://doi.org/10.1523/JNEUROSCI.3214-10.2011>
- Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A ConvNet for the 2020s. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11976–11986.
- Makris, N., Goldstein, J. M., Kennedy, D., Hodge, S. M., Caviness, V. S., Faraone, S. V., Tsuang, M. T., & Seidman, L. J. (2006). Decreased volume of left and total anterior insular lobule in schizophrenia. *Schizophrenia Research*, 83(2–3), 155–171. <https://doi.org/10.1016/j.schres.2005.11.020>
- Morawetz, C., & others. (2021). Emotion regulation modulates dietary decision-making via activity in the prefrontal-striatal valuation system. *Cerebral Cortex*. <https://doi.org/10.1093/cercor/bhaa398>
- Morawetz, C., Bode, S., Derntl, B., & Heekeren, H. R. (2017). The effect of strategies, goals and stimulus material on the neural mechanisms of emotion regulation: A meta-analysis of fMRI studies. *Neuroscience & Biobehavioral Reviews*, 72, 111–128. <https://doi.org/https://doi.org/10.1016/j.neubiorev.2016.11.014>
- Motoki, K., & Suzuki, S. (2020). Extrinsic Factors Underlying Food Valuation in the Human Brain. *Front Behav Neurosci*, 14, 131.
- Mumford, J. A., Turner, B. O., Ashby, F. G., & Poldrack, R. A. (2012). Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *Neuroimage*, 59(3), 2636–2643. <https://doi.org/10.1016/j.neuroimage.2011.08.076>
- OpenAI. (2024). GPT-4 Technical Report. *Arxiv Preprint Arxiv:2303.08774*.
- Penny, W. D., & Holmes, A. P. (2007). *Random effects analysis* (K. J. Friston, J. T. Ashburner, S. J. Kiebel, T. E. Nichols, & W. D. Penny, Eds.; pp. 156–165). Academic Press. <https://doi.org/10.1016/B978-012372560-8/50012-7>

- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. *International Conference on Machine Learning*, 8748–8763.
- Rangel, A., Camerer, C., & Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. *Nature Reviews Neuroscience*, 9(7), 545–556. <https://doi.org/10.1038/nrn2357>
- Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R. P., Berker, A. de, Ganguli, S., Gillon, C. J., Hafner, D., Kepecs, A., Kriegeskorte, N., Latham, P., Lindsay, G. W., Miller, K. D., Naud, R., Pack, C. C., ... Kording, K. P. (2019). A deep learning framework for neuroscience. *Nature Neuroscience*, 22(11), 1761–1770. <https://doi.org/10.1038/s41593-019-0520-2>
- Rutledge, R. B., Skandali, N., Dayan, P., & Dolan, R. J. (2014). A computational and neural model of momentary subjective well-being. *Proceedings of the National Academy of Sciences*, 111(33), 12252–12257. <https://doi.org/10.1073/pnas.1407535111>
- Samejima, K., Ueda, Y., Doya, K., & Kimura, M. (2005). Representation of action-specific reward values in the striatum. *Science*, 310(5752), 1337–1340. <https://doi.org/10.1126/science.1115270>
- Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45), e2105646118. <https://doi.org/10.1073/pnas.2105646118>
- Scott, A. H., Allen, W. S., & Gregory, M. (2016). *fMRI - 原理と実践 - メディカル・サイエンス・イン ターナショナル*.
- Seeley, W. W., Menon, V., Schatzberg, A. F., Keller, J., Glover, G. H., Kenna, H., Reiss, A. L., & Greicius, M. D. (2007). Dissociable intrinsic connectivity networks for salience processing and executive control. *Journal of Neuroscience*, 27(9), 2349–2356. <https://doi.org/10.1523/JNEUROSCI.5587-06.2007>
- Simonyan, K., & Zisserman, A. (2015, ). Very Deep Convolutional Networks for Large-Scale Image Recognition. *International Conference on Learning Representations*.
- Spinelli, S., & Monteleone, E. (2021). Food Preferences and Obesity. *Endocrinol Metab (Seoul)*, 36(2), 209–219.
- St-Yves, G., Allen, E. J., Wu, Y., Kay, K., & Naselaris, T. (2023). Brain-optimized deep neural network models of human visual areas learn non-hierarchical representations. *Nature Communications*, 14, 3329. <https://doi.org/10.1038/s41467-023-38674-4>
- Suzuki, S., Adachi, R., Dunne, S., Bossaerts, P., & O'Doherty, J. P. (2015). Neural mechanisms underlying human consensus decision-making. *Neuron*, 86(2), 591–602.

- Suzuki, S., Cross, L., & O'Doherty, J. P. (2017). Elucidating the underlying components of food valuation in the human orbitofrontal cortex. *Nat Neurosci*, 20(12), 1780–1786.
- Tang, D. W., Fellows, L. K., & Dagher, A. (2014). Behavioral and neural valuation of foods is driven by implicit knowledge of caloric content. *Psychological Science*, 25(12), 2168–2176. <http://www.jstor.org/stable/24543633>
- Vaduganathan, M., Mensah, G. A., Turco, J. V., Fuster, V., & Roth, G. A. (2022). The global burden of cardiovascular diseases and risk: A compass for future health. *Journal of the American College of Cardiology*, 80(25), 2361–2371. <https://doi.org/https://doi.org/10.1016/j.jacc.2022.11.005>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems*, 30.
- Williams, A. H. (2024). Equivalence between representational similarity analysis, centered kernel alignment, and canonical correlations analysis. *Proceedings of Unireps: The Second Edition of the Workshop on Unifying Representations in Neural Models*, 285, 10–23.
- Worsley, K. J., Marrett, S., Neelin, P., Vandal, A. C., Friston, K. J., & Evans, A. C. (1996). A unified statistical approach for determining significant signals in images of cerebral activation. *Human Brain Mapping*, 4(1), 58–73. [https://doi.org/10.1002/\(SICI\)1097-0193\(1996\)4:1<58::AID-HBM4>3.0.CO;2-O](https://doi.org/10.1002/(SICI)1097-0193(1996)4:1<58::AID-HBM4>3.0.CO;2-O)
- Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc Natl Acad Sci U S a*, 111(23), 8619–8624.