（報道発表資料）

2025.11.17

NTT株式会社

心に思い浮かべた映像を言葉に変換する
脳解読技術「マインド・キャプショニング」を実現

～言葉を使わずに考えを伝える新たなコミュニケーション手段を開拓～

発表のポイント：

- ◆ 脳情報解読技術と言語 AI モデルを組み合わせ、ヒトが「見た」さらには「思い浮かべた」映像の視覚内容に関するテキスト記述を脳活動から生成する技術「マインド・キャプショニング」を実現しました。
- ◆ 脳の言語野を介さずに高精度なテキスト生成が可能であることを実証し、非言語的思考に関わる脳内情報を言語へ翻訳するという脳情報解読の新たな可能性を拓きました。
- ◆ 本成果は、視覚内容の複雑かつ構造化された意味に関する脳内表現を解明する手がかりを提供するとともに、将来的には、視覚以外の感覚イメージや感情、概念的思考など、多様な非言語的思考を言語へ翻訳するための汎用技術として発展させることで、発話困難者の意思伝達支援などに貢献することが期待されます。

NTT 株式会社（本社：東京都千代田区、代表取締役社長：島田　明、以下「NTT」）は、脳活動からヒトが見ている内容に関するテキスト記述を生成する新たな技術（マインド・キャプショニング）を考案しました。さらに本技術を、あらかじめ記憶した映像を想起している時の脳活動に適用することで、想起された動画の視覚内容に関するテキスト記述を脳活動から生成することに、世界で初めて成功しました（図 1）。

本技術は、脳活動のパターン解析により心や身体の状態を解読する「脳情報デコーディング」[※1] に人工知能(AI)モデルを組み合わせた「脳-AI 統合型デコーディング」[※2] を、言語 AI モデルの導入によって拡張したアプローチです。言語 AI モデルの高い表現力と生成力を活用することで、脳内の意味情報と整合性の高いテキストの生成を可能にしました。さらに、本技術は脳の言語野の活動を用いずに、ヒトが「見た」さらには「想起した」映像に関する非言語的内容を、脳活動から言語へ変換できることを実証しました。これは、言語的思考を再構成するのではなく、非言語的思考を言語へ翻訳して解釈可能にするという、新たな脳情報デコーディングの可能性を切り拓く試みです。

本成果は、複雑な視覚的意味情報の脳内表現に関する科学的理解を促進するとともに、発話困難者の意思伝達支援や、言葉を用いずに感情や意図を伝達する新たなコミュニケーション手段の実現に寄与することが期待されます。

本研究成果は、2025 年 11 月 5 日（米国東部時間）に米国科学誌「Science Advances」のオンライン版に掲載されました。また、本研究成果の一部は、2025 年 11 月 19 日～26 日に開催される NTT R&D FORUM 2025 IOWN∴Quantum Leap[※3] に展示予定です。

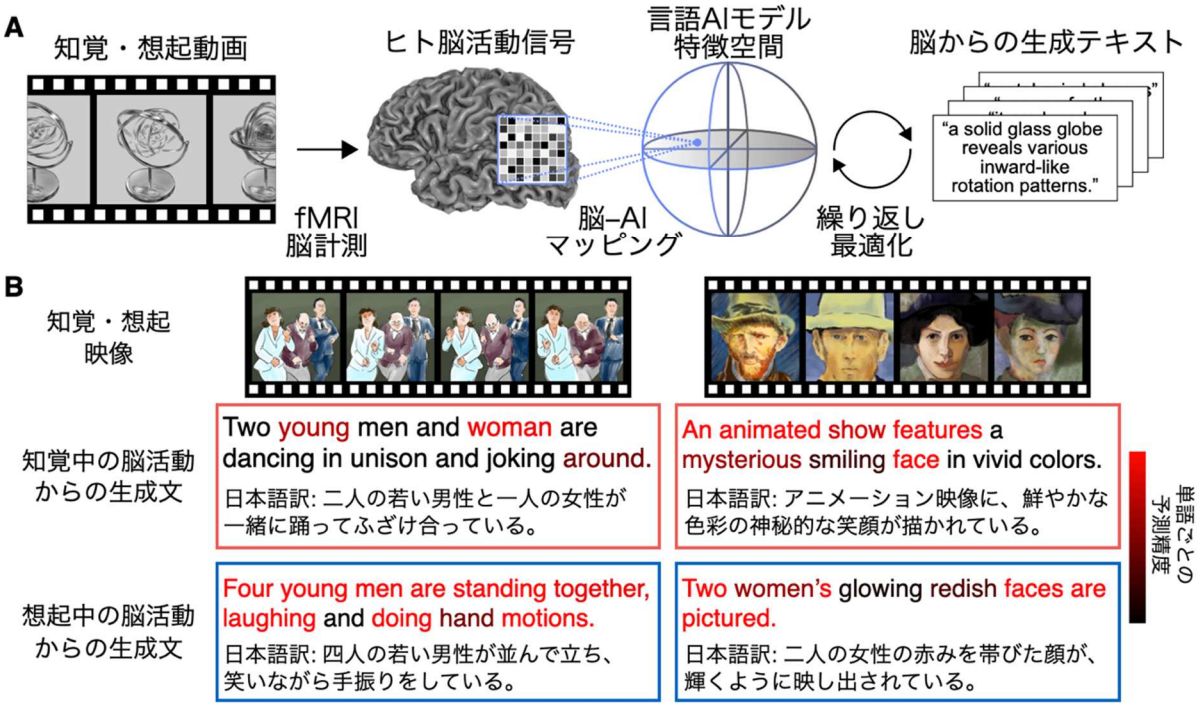

図 1:本研究の概要と主要結果。(A) 本研究で考案した技術「マインド・キャプショニング」の概要図。ヒトが動画を知覚または想起しているときの fMRI 脳活動信号※4を、機械学習の方法で言語 AI モデル(深層言語モデル※5)の特徴空間にマッピングし、その脳活動に対応する特徴とより類似度が高い特徴を持つテキストを探索しながら、テキストを繰り返し最適化することで、知覚・想起した映像の意味内容を反映した文章を脳活動から生成する技術を考案しました。(B) 実験で用いた映像(上段)と、それらを知覚あるいは想起している時の脳活動から生成されたテキスト(下段)の例。生成文中のトークン(図中および以下、「単語」と表記)のうち、人手で作成した動画内容の説明文(参照文)に意味が似た単語が含まれていたものをより赤い色で示しています。知覚条件と想起条件のいずれにおいても、知覚・想起している映像内容を反映し、参照文中の単語と意味の似た単語を含む構造化されたテキストが生成されており、脳から生成したテキストのみを用いて、どの映像を知覚・想起しているかを識別できることが確認できます。

1. 研究の背景

ヒトは視覚を通して世界を認識・記憶し、その内容を言葉で表現することができます。近年、脳から直接「言語的な情報」を解読する技術が発展していますが、視覚的イメージのような「非言語的な思考内容」を文章として解読することは依然として困難です。もし非言語的思考に関する多様な脳内情報を脳活動からテキストへ翻訳できれば、脳活動から心の状態をより柔軟に解釈でき、神経科学研究や応用研究に新たな可能性が拓かれると考えられます。

そこで本研究では、脳活動解析により心身の状態を解読する「脳情報デコーディング」※1 と人工知能(AI)モデルを組み合わせた「脳–AI 統合型デコーディング」※2 に、言語 AI モデル(深層言語モデル※5)を導入し、「知覚・想像した視覚内容に関するテキスト記述をヒト脳活動から生成する技術」(マインド・キャプショニング)を考案しました(図 1)。

2. 研究の成果

本研究で考案した「マインド・キャプショニング」では、(1) 脳活動から深層言語モデルの意味特徴を予測するデコーディングモデル(デコーダ)の訓練、(2) 訓練済みデコーダで予測した特徴に基づくテキスト記述の繰り返し最適化、という二段階の処理によって、機能的磁気共鳴画像法(fMRI)※4 で計測したヒト脳活動から、知覚・想像した視覚内容のテキスト記述生成に成功しました(図 2)。

本手法では二段階構成で脳活動からテキストの生成を行います。Stage 1 では、まず fMRI で計測した動画観察中のヒト脳活動データと、クラウドソーシングによる動画の視覚内容の記述文データを収集します。次に、各動画の記述文から言語 AI モデル(深層言語モデル; DeBERTa-large)を用いて抽出した意味特徴を、対応する動画を見ている時の脳活動から予測(変換)するようデコーダを学習させます。Stage 2 では、新たな動画を見たり想起したりしている時の脳活動を、学習済みデコーダで意味特徴に変換し(デコード特徴)、これをターゲットとして単語レベルの繰り返し最適化によりテキストを生成します。最適化の過程では、任意の単語や文を初期値として(本研究では<unk>から開始)、その一部をランダムにマスク単語([MASK])で置換・挿入して候補文を複数作成し、マスク言語モデル(RoBERTa-large)※6 で補完します。得られた新規候補の中から、デコード特徴との類似度が高いものを選び、この手続きを繰り返すことで、脳情報と整合したテキストを漸進的に生成します。

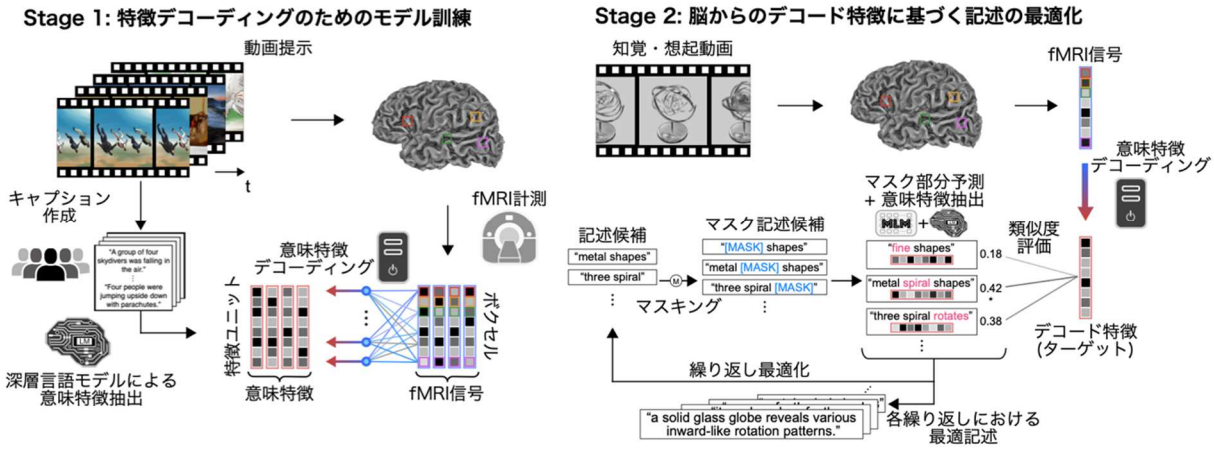

図 2:マインド・キャプショニングによる脳活動からのテキスト生成。

本手法の有効性を検証するため、まず動画を見ている時の脳活動を用いたテキスト生成を行いました。その結果、何の事前情報も持たない<unk>という単語から始めても、最適化を繰り返すことで徐々に見ている映像に含まれる内容を反映した単語が生成され、100 回後には動画全体を的確に説明する記述が得られました(図 3A)。さらに、記憶した動画を想起している時の脳活動に同じ手法を適用することで、想起内容を記述するテキストの生成にも成功しました(図 1B 下段)。

生成テキストの精度を評価するため、各動画について脳からの生成文と人手による参照文の類似度スコアを算出し、その値が無関係な動画の参照文との類似度より高いか、すなわち類似度スコアに基づいて知覚・想起していた動画を複数の候補動画の中から正しく同定できるかを検証しました。その結果、脳全体の活動を用いた場合、100 本の候補動画の中から、観察(知覚)時は約 50%、想起時でも約 30%の精度で正しい動画を同定できました(チャンス水準※7 = 1%、参加者 6 名の平均)(図 3B)。

さらに注目すべきは、言語処理に関与する脳部位(前頭葉から側頭葉にわたる言語ネットワーク、

図 3B の赤色の脳部位)の活動を除外した解析でも、全脳を用いた場合と同様に高品質なテキスト生成と高精度な動画同定が可能であったことです。この結果は、本手法が脳内の言語情報を再構成しているのではなく、非言語的な情報を言語として解釈可能にしていることを示唆しています。

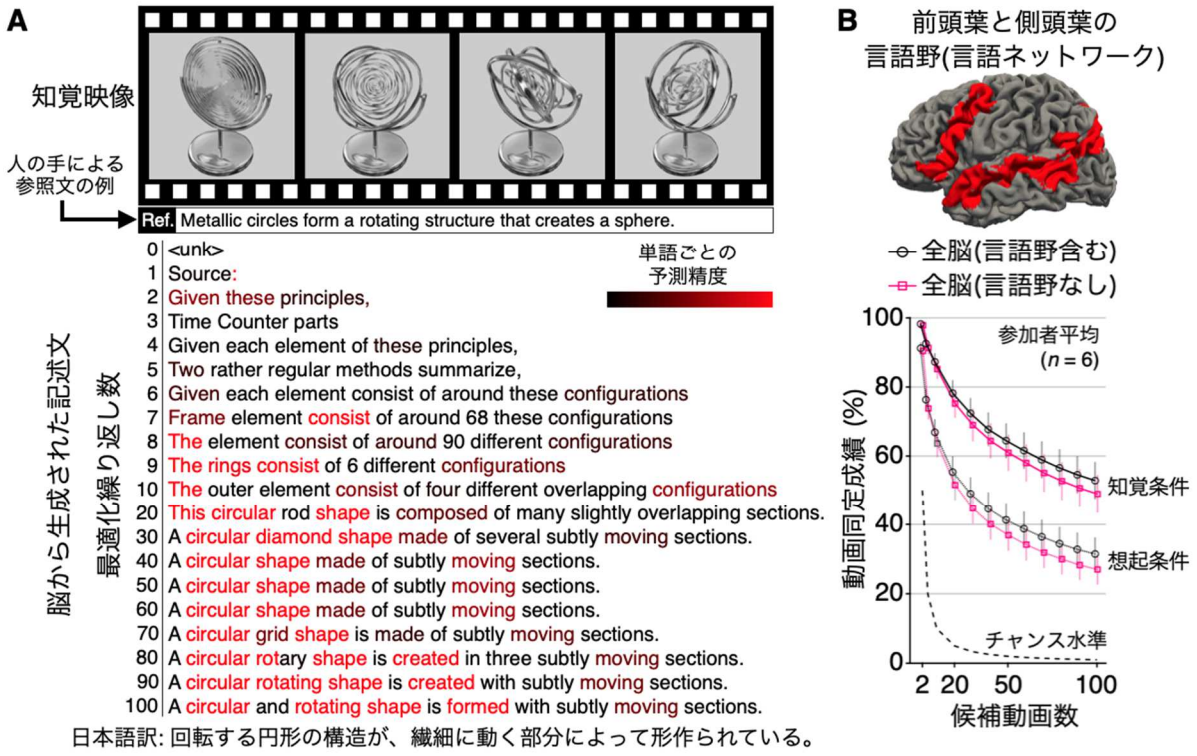

図 3:知覚・想起内容に関する脳活動からのテキスト生成結果。(A)繰り返し最適化の過程において脳活動から生成されたテキストの例(知覚条件)。(B)言語野を解析に含めた場合と除外した場合の動画同定成績。各動画の生成文と参照文の間の類似度スコア(意味特徴の相関)に基づき、候補動画数を 2～100 まで変化させて知覚・想起した動画を正しく同定できるかを評価しました。

3. 今後の展開

本研究では、脳活動から複雑に構造化された視覚内容の意味情報を解読することに成功しました。この成果は、複雑な視覚的意味が脳内でどのように表現されるかを探る新たなツールとして本手法を活用できる可能性を示しています。ただし、本研究ではウェブから収集した自然な動画を実験に用いたため、例えば「人が犬を噛む」といった非典型的な場面への汎化性能は検証されていません。今後は、モデルや訓練データに内在するバイアスの影響を精査するとともに、本技術が、脳内に実在する情報をどの程度正確に捉えているかについて、より詳細な検証が求められます。こうした検証を重ねることで、将来的には、乳幼児や動物など、言葉を話せない対象の脳における構造化された思考の発達過程を解明する研究にも応用できると期待されます。

一方で、脳活動から心に思い浮かべた内容を記述するテキストを生成できるという本研究は、個人の心的プライバシーを侵害するおそれを内包しています。本研究では、実験参加者からの明示的な同意のもと、1 人あたり延べ 17 時間におよぶ脳活動計測を複数日にわたり実施しました。現時点では、脳内情報を高精度に解析するためには、このような大規模かつ長時間のデータ取得が必要不可欠であり、本技術は実験参加者の協力によって初めて成立するものです。しかし将来的に、

脳計測や解析の技術がさらに発展すれば、より少ないデータから個人の思考内容を解読できる可能性が生じます。これは科学的・技術的には大きな前進である一方で、本人の意思に反して、まだ言語化されていない思考内容が推定されるリスクも伴います。さらに、モデルやデータに内在するバイアスやノイズが、生成されるテキストの内容を歪めることで、解釈の方向性やニュアンスを意図せず変化させ、脳活動から得られる情報の正確性を損なう可能性があることについても併せて考慮する必要があります。

こうした課題を踏まえ、本分野の進展にあたっては、研究の健全な発展と個人の精神的自律の両立を図ることが極めて重要となります。今後は、技術を適用する文脈を的確に見極め、その際に求められる精度水準を科学的かつ実践的な観点から慎重に検討するとともに、本人が「どの思考を自身のものとして表現するか」を主体的に選択できる権利を尊重し、技術の精度や信頼性と心的プライバシーの保護を両立させるための技術的・倫理的な議論をさらに深めていく必要があります。そのために私たちは、社会的信頼の確保に向け、透明性の高い研究手法とデータ運用のルールづくりを推進していきます。

このように、人間の情報処理メカニズムの深い理解に基づく技術開発を通して、私たちは人間の思考や心の仕組みを解き明かす科学的探究を進めるとともに、その成果を活かして次世代のコミュニケーション技術を切り拓いていきます。

論文情報
雑誌名：「Science Advances」（オンライン版：11 月 5 日）
論文タイトル：Mind Captioning: Evolving descriptive text of mental content from human brain activity
著者：Tomoyasu Horikawa
DOI：https://doi.org/10.1126/sciadv.adw1464
URL：https://www.science.org/doi/10.1126/sciadv.adw1464

【用語解説】

※1. 脳情報デコーディング： fMRI などで計測された脳活動信号を、機械学習などのパターン解析を用いて解析し、身体や心の状態を脳活動から予測・解読する技術。

※2. 脳-AI 統合型デコーディング： 脳活動パターンを機械学習の方法で AI 特徴空間に写像（マッピング）することで、AI 技術を活用した脳情報解析を可能にするアプローチ。本研究では、fMRI で計測した動画観察中のヒト脳活動を言語 AI モデル（深層言語モデル）の意味特徴空間にマッピングし、その特徴に基づいてテキストを最適化することで、知覚・想起した動画の意味内容を反映した記述文の生成を実現。

※3. NTT R&D FORUM 2025 IOWN∴Quantum Leap」公式サイト https://www.rd.ntt/forum/2025/
出展情報： 展示 No. D14、（展示カテゴリ）生成 AI

※4. 機能的磁気共鳴画像法（functional magnetic resonance imaging; fMRI）：MRI 装置を用いて、脳活動を非侵襲的に計測する代表的手法の一つ。神経活動そのものではなく、活動に伴う血流や血中酸素濃度の変化を反映した BOLD (blood-oxygen-level dependent)信号を捉え、脳活動の指標とする。ヒトを対象とした脳計測技術の中では比較的高い時空間解像度を持ち、本研究では、全脳を 2 mm 角・1 秒間隔で計測。

※5. 深層言語モデル： 大量のテキストデータに基づく学習により、単語の意味や文脈上の関係を内部ベクトル表現に埋め込むことで、多様な自然言語処理課題において、高い性能を発揮している言語 AI モデルの総称。本研究では、その内部表現を「意味特徴」として、fMRI 計測データからの予測対象やテキスト生成の最適化に利用。

※6. マスク言語モデル（masked language modeling model; MLM model）：深層言語モデルの一種。学習時に入力文中の一部をマスク単語([MASK])に置き換え、その部分を予測する課題を通じて訓練される。代表例は BERT（Bidirectional Encoder Representations from Transformers)で、双方向性の注意機構により前後の文脈を踏まえた意味理解を可能にする。

※7. チャンス水準 （chance level）： ランダムに選んだ場合に正答する確率。候補動画が N 本の場合は 1/N で、本研究では 2～100 本の条件に応じて異なる （e.g., 100 本の場合は 1%）。

■本件に関する報道機関からのお問い合わせ先
NTT 株式会社
先端技術総合研究所
広報担当
問い合わせフォームへ