Representações de Conceitos de Objetos Semelhantes a Humanos em LLMs e MLLMs

Fonte: Excertos de "Human-like object concept representations_LLM.pdf"
Data: Outubro de 2024
Autores: C.D., H.H. et al.

Visão Geral

Este estudo investiga a capacidade de Large Language Models (LLMs) e Multimodal Large Language Models (MLLMs)[1] de desenvolver representações de conceitos de objetos semelhantes às humanas a partir de dados linguísticos e multimodais. Combinando análises comportamentais e de neuroimagem, os pesquisadores coletaram 4,7 milhões de julgamentos de tripletos[2] de LLMs (ChatGPT-3.5) e MLLMs (Gemini Pro Vision 1.0) para derivar embeddings de baixa dimensão[3]. Os resultados revelam que esses modelos, especialmente os MLLMs, desenvolvem representações conceituais de objetos que compartilham semelhanças fundamentais com o conhecimento conceitual humano e alinham-se com padrões de atividade neural em regiões cerebrais seletivas por categoria. O estudo sugere que os modelos de IA podem não apenas emular aspectos da cognição humana, mas também fornecer informações sobre a estrutura das representações mentais.

Principais Temas e Ideias

1. Formação de Representações de Objetos Semelhantes a Humanos em LLMs/MLLMs

---

[1] Os LLMs processam e aprendem primariamente a partir de **dados textuais**, enquanto os MLLMs integram e aprendem a partir de **dados textuais e visuais**. Essa diferença na modalidade de entrada de treinamento permite que os MLLMs desenvolvam representações conceituais de objetos que se alinham mais fortemente com as representações humanas e a atividade neural em regiões cerebrais seletivas para categorias, como EBA, PPA, RSC e FFA, em comparação com os LLMs puramente textuais.

[2] No contexto dos estudos mencionados, "tripletos" referem-se a **conjuntos de três objetos** que são apresentados a participantes (sejam humanos, Large Language Models - LLMs, ou Multimodal LLMs - MLLMs) em uma tarefa específica.
A tarefa em questão é denominada **"triplet odd-one-out"** (algo como "o elemento estranho do trio" ou "o diferente do trio"). O objetivo dessa tarefa é que o participante **identifique qual dos três objetos é o mais dissimilar** dos outros dois.
Para ilustrar:
• Os LLMs (como o ChatGPT-3.5) recebiam prompts textuais com descrições dos três objetos, por exemplo: "Dado um trio de objetos {'[Objeto_A]', '[Objeto_B]', '[Objeto_C]'}, qual deles é o diferente?".
• Os MLLMs (como o Gemini Pro Vision) recebiam três imagens de objetos lado a lado e eram solicitados a relatar qual imagem era a menos similar às outras duas.
Esta metodologia é amplamente utilizada em psicologia cognitiva para modelar as dimensões mentais humanas. No estudo em questão, foram coletados **4,7 milhões de julgamentos de tripletos** de LLMs e MLLMs para criar representações de conceitos de objetos. Essa abordagem permite explorar como os modelos conceituam e categorizam objetos naturais.

[3] "Embeddings de baixa dimensão" (ou *low-dimensional embeddings* em inglês) são **representações compactas e eficientes de dados complexos**, neste caso, de objetos naturais. Em vez de representar um objeto por todas as suas características percebidas individualmente (o que seria uma "alta dimensão"), um *embedding* de baixa dimensão condensa essa informação em um número menor de características ou "dimensões" mais essenciais. No contexto dos estudos mencionados, esses *embeddings* foram criados para capturar a estrutura de similaridade de 1.854 objetos naturais.

- **Fundamento da Cognição Humana:** A categorização e conceituação de objetos são "o alicerce da cognição humana, influenciando tudo, desde a percepção até a tomada de decisões." (Introdução)
- **Capacidade Emergente em LLMs:** O estudo aborda a questão crítica de "até que ponto representações psicológicas complexas e de tarefas gerais podem surgir sem treinamento explícito específico de tarefas". (Introdução) LLMs e MLLMs, treinados em vastos corpora, mostram "capacidades impressionantes em tarefas como identificação de objetos, categorização de informações, comunicação de conceitos e inferência." (Introdução)
- **Metodologia de Julgamento de Tripletos:** O estudo empregou uma "tarefa de triplos odd-one-out, um paradigma bem estabelecido em psicologia cognitiva", coletando 4,7 milhões de julgamentos de similaridade de tripletos de 1.854 objetos para LLMs e MLLMs, espelhando dados comportamentais humanos. (Introdução, Resultados)
- **Embeddings de Baixa Dimensão:** Usando o método Sparse Positive Similarity Embedding (SPoSE), foram identificadas 66 dimensões esparsas e não negativas que subjazem aos julgamentos de similaridade dos modelos, levando a "excelentes previsões de comportamento de ensaio único e pontuações de similaridade entre pares de objetos." (Introdução) Essas 66 dimensões foram escolhidas para "alinhar com as 66 dimensões centrais dos julgamentos de similaridade humana". (Resultados)

2. Interpretabilidade e Estrutura das Dimensões

- **Dimensões Interpretáveis:** As dimensões subjacentes aos embeddings são "interpretáveis, exibindo agrupamento semântico espontâneo e caracterizando a estrutura em larga escala das representações mentais de objetos naturais dos LLMs". (Introdução)
- **Tipos de Dimensões:Categorias Semânticas:** Algumas dimensões representam categorias semânticas como "alimentos, animais, veículos". (Resultados)
- **Características Perceptivas:** Outras capturam características perceptivas como "dureza, valor, temperatura ou textura". (Resultados)
- **Propriedades Espaciais/Visuais (MLLMs):** MLLMs exibem dimensões que refletem "propriedades espaciais globais (por exemplo, lotado), enquanto algumas transmitem forma (achatamento, alongamento) e cor." (Resultados)
- **Outros Traços:** Dimensões também distinguem "especificidade do usuário (crianças vs. adultos), composição física (madeira, cerâmica, metal) e traços relacionados ao ambiente (terra vs. mar, interno vs. externo)." (Resultados)
- **Diferenças de Granularidade:** "LLM e MLLM tendem a formar categorias mais específicas (por exemplo, frutas, vegetais, chapéus) do que as categorizações mais amplas dos humanos." (Resultados) Por exemplo, LLM distingue "doces congelados" e "bebidas quentes", ou "animais selvagens" vs. "gado". (Resultados)

3. Alinhamento com a Cognição Humana e Atividade Neural

- **Consistência Modelo-Humano:** O estudo encontrou "forte alinhamento entre os embeddings do modelo e os padrões de atividade neural em regiões cerebrais como EBA, PPA, RSC e FFA." (Resumo, Introdução) Isso fornece "evidências convincentes de que as representações de objetos em LLMs, embora não idênticas às humanas, compartilham semelhanças fundamentais que refletem aspectos chave do conhecimento conceitual humano." (Resumo)
- **Desempenho Preditivo:** Os embeddings derivados do SPoSE "obtêm até 87,1%, 85,9% e 95,4% da acurácia preditiva ideal para LLM, MLLM e humanos, respectivamente" na tarefa odd-one-out. (Resultados)
- **Alinhamento de Dimensões Essenciais:** 31 das 66 dimensões do LLM e 42 das 66 dimensões do MLLM "correlacionam-se fortemente com as dimensões humanas ($r > 0.4$), indicando alinhamento substancial." (Resultados) No geral, "38 das 66 dimensões são compartilhadas entre os três sistemas." (Resultados)

- **Diferenças Persistentes:** Apesar do alinhamento, "diferenças notáveis permanecem entre LLMs e humanos." (Resultados) Por exemplo, "o humano pode fazer escolhas com base na cor (como 'vermelho'), enquanto o LLM só faz escolhas com base na semântica (como 'protetor')." (Resultados) MLLM "ainda carece de dimensões específicas relacionadas à cor", mas se alinha mais com humanos em dimensões como "forma" e "características espaciais". (Resultados)
- **Correlação com o Cérebro:** MLLM e os embeddings humanos "alinham-se mais de perto com a maioria das regiões cerebrais do que LLM e CLIP." (Resultados) LLM e MLLM atingem cerca de "60% e 85% do desempenho humano em RSA searchlight, respectivamente. Na codificação voxel-wise, LLM atinge 90% do desempenho humano, enquanto MLLM quase iguala os níveis humanos." (Resultados)

4. Implicações e Aplicações Futuras

- **Sistemas Cognitivos Semelhantes a Humanos:** As representações mentais de baixa dimensão podem "informar o desenvolvimento de sistemas cognitivos artificiais mais semelhantes aos humanos, melhorando sua interação natural com os humanos." (Discussão)
- **Melhoria do Alinhamento:** "Adaptando os prompts para enfatizar atributos específicos (por exemplo, 'vermelho' ou 'artificial'), acreditamos que os modelos poderiam fazer escolhas mais consistentes com os julgamentos humanos". (Discussão, ver também Suplementar Fig. 2-3)
- **Conjuntos de Dados Comportamentais:** Os extensos conjuntos de dados comportamentais de máquinas coletados oferecem um "valioso benchmark para avaliar as representações do modelo de IA." (Discussão)
- **Limitações e Direções Futuras:** O estudo se concentrou em ChatGPT-3.5 e Gemini Pro Vision (v1.0), mas a metodologia é "extensível a outros LLMs de ponta como GPT-4V". (Discussão) O uso de "anotações em nível de imagem" (capturando atributos visuais como cor e textura) para LLMs, semelhante ao que os MLLMs fazem, pode "tornar o LLM mais consistente com os julgamentos humanos". (Discussão)

Resumo das Descobertas Chave

- LLMs e MLLMs podem desenvolver **representações de conceitos de objetos de baixa dimensão (66 dimensões)** que são estáveis e preditivas.
- Essas dimensões são **interpretáveis** e mostram **agrupamento semântico**, refletindo categorias conceituais (ex: animais, comida) e características perceptivas (ex: textura, forma).
- Existe um **forte alinhamento entre os embeddings do modelo (especialmente MLLM) e as representações cerebrais humanas**, particularmente em regiões seletivas por categoria (EBA, PPA, RSC, FFA).
- **MLLMs (que incorporam dados visuais)** demonstram maior alinhamento com a cognição humana e a atividade cerebral do que LLMs puramente linguísticos, especialmente em características visuais e categorias mais detalhadas.
- Embora as representações não sejam idênticas, há um número significativo de **dimensões compartilhadas (38 de 66)** entre humanos e modelos.
- As discrepâncias podem ser reduzidas com **prompts guiados** que direcionam a atenção do modelo para dimensões priorizadas por humanos.

Este resumo destaca o avanço significativo que este estudo representa para a compreensão da inteligência de máquina e seu potencial para informar o desenvolvimento de sistemas cognitivos artificiais mais semelhantes aos humanos.