

From Textual Entailment in BERT to Visual Entailment in GPT-4V

-テキスト含意関係から
画像の含意関係へ-

2023年10月20日(金)

@羽田空港

太田博三



本LTでは1)画像の含意関係認識と2)画像の常識推論を取り上げます。

目次:

1.テキストの含意関係認識とは

2.画像の含意関係認識とは

3.マーケティングへの応用 : お店画像とクチコミの相性

4.考察/まとめ

5.補足: 画像の常識推論 ("Visual Commonsense reasoning")

The Dawn of LLMs: Preliminary Explorations with GPT-4V(ision)

Zhengyuan Yang*, Linjie Li*, Kevin Lin*, Jianfeng Wang*, Chung-Ching Lin*,
Zicheng Liu, Lijuan Wang*
Microsoft Corporation

* Core Contributor ♦ Project Lead

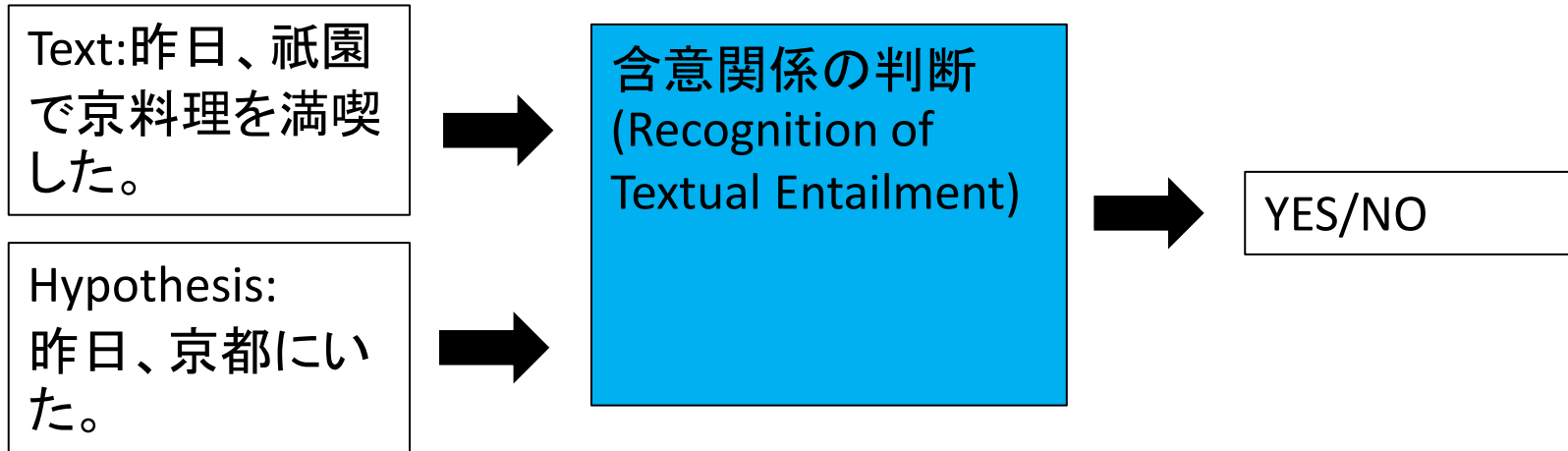
Abstract

Large multimodal models (LMMs) extend large language models (LLMs) with multi-sensory skills, such as visual understanding, to achieve stronger generic intelligence. In this paper, we analyze the latest model, GPT-4V(ision) [9]–[10], [1], to deepen the understanding of LLMs. The analysis focuses on the intriguing tasks that GPT-4V can perform, containing test samples to probe the quality and generality of GPT-4V’s capabilities, its supported inputs and working modes, and the effective ways to prompt the model. In our approach to exploring GPT-4V, we curate and organize a collection of carefully designed qualitative samples spanning a variety of domains and tasks. Observations from these samples demonstrate that GPT-4V’s unprecedented ability in processing arbitrarily interleaved multimodal inputs and the generality of its capabilities together make GPT-4V a powerful multimodal generalist system. Furthermore, GPT-4V’s unique capability of understanding visual markers drawn on input images can give rise to new human-computer interaction methods such as visual referring prompting. We conclude the report with in-depth discussions on the emerging application scenarios and the future research directions for GPT-4V-based systems. We hope that this preliminary exploration will inspire future research on the next-generation multimodal task formulation, new ways to exploit and enhance LLMs to solve real-world problems, and gaining better understanding of multimodal foundation models. Finally, we acknowledge that the model under our study is solely the product of OpenAI’s innovative work, and they should be fully credited for its development. Please see the GPT-4V contributions paper [10] for the authorship and credit attribution: <https://cdn.openai.com/contributions/gpt-4v.pdf>.

4.3 Multimodal
Knowledge and
Commonsense

6 Temporal and Video
Understanding を参照

1. テキストの含意関係認識とは



OT

以下のTextからHypothesisが推論されればYES、そうでなければNOと教えてください。

Text:

昨日、祇園で京料理を満喫した。

Hypothesis:

昨日、京都にいた。



Hypothesis: YES

与えられたテキストから、昨日祇園で京料理を楽しんだという情報が提供されており、そのため昨日は京都にいたと推論されます。

2.画像の含意関係認識とは

前提(Premise)
画像

+

仮説
(Hypothesis)
文

=

結果:

- 1)含意である
- 2)中立である
- 3)反駁している



Premise

+

- Two woman are holding packages.
- The sisters are hugging goodbye while holding to go packages after just eating lunch.
- The men are fighting outside a deli.

Hypothesis

=

- Entailment
- Neutral
- Contradiction

Answer

2.画像の含意関係認識とは

前提(Premise)
画像

+

仮説
(Hypothesis)
文

=

結果:

- 1)含意である
- 2)中立である
- 3)反駁している



+

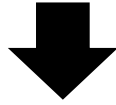
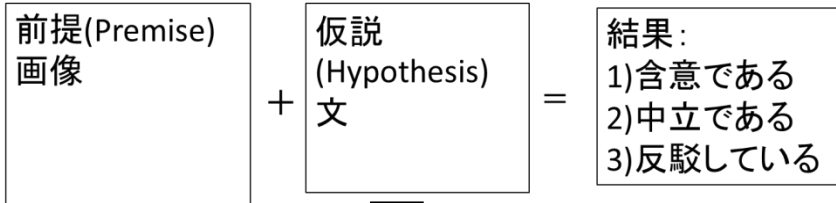
仮説:
2人の女性が
荷物を持っ
ています。

=

結果:

このHypothesisは
画像との関係で「1.
Entailment(含
意)」に当たります。

2.画像の含意関係認識とは



+

仮説:
2人の女性が
荷物を持っ
ています。

=

結果:
このHypothesisは
画像との関係で「1.
Entailment(含
意)」に当たります。

仮説:
市内の路上で心からの抱擁を
交わす2人の女性。1人は白い
容器を持ち、もう1人はフォル
ダーを握り締めている。その後
ろでは建設機械やバイクなど
の街の喧騒が続いている。

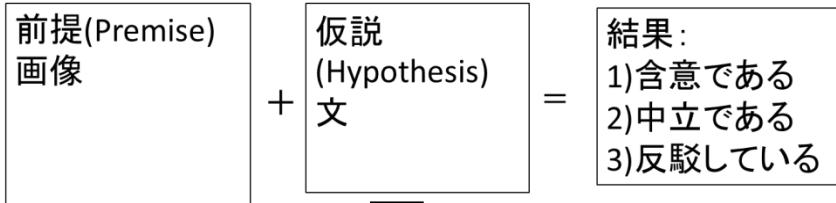
+

仮説:
2人の女性
が荷物を
持っていま
す。

=

結果:
このHypothesisは
画像との関係で「1.
Entailment(含
意)」に当たります。

2.画像の含意関係認識とは



+

仮説:
2人の女性が
荷物を持っ
ています。

=

結果:
このHypothesisは
画像との関係で「1.
Entailment(含
意)」に当たります。

仮説:
市内の路上で心からの抱擁を
交わす2人の女性。1人は白い
容器を持ち、もう1人はフォル
ダーを握り締めている。その後
ろでは建設機械やバイクなど
の街の喧騒が続いている。

+

仮説:
2人の女性
が荷物を
持っていま
す。

=

結果:
このHypothesisは
画像との関係で「1.
Entailment(含
意)」に当たります。

2.画像の含意関係認識とは

前提(Premise)
画像

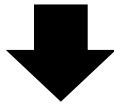
+

仮説
(Hypothesis)
文

=

結果:

- 1)含意である
- 2)中立である
- 3)反駁している



理由: 画像には2人の女性が写っており、少なくとも1人の女性が赤い袋を持っていて、もう1人の女性も何かの容器を持っていることが確認できる。この情報に基づいて、提供されたHypothesisは画像の内容と一致しており、画像がHypothesisを含意していると言えます。

仮説:

市内の路上で心からの抱擁を交わす2人の女性。1人は白い容器を持ち、もう1人はフォルダーを握り締めている。その後ろでは建設機械やバイクなどの街の喧騒が続いている。

+

仮説:

2人の女性が荷物を持っています。

=

結果:

このHypothesisは画像との関係で「1. Entailment(含意)」に当たります。

2.画像の含意関係認識とは

前提(Premise)
画像

+

仮説
(Hypothesis)
文

=

結果:
1)含意である
2)中立である
3)反駁している



+

姉妹たちは昼食を
食べたばかりで、
持ち帰り用の荷物を
持ちながらハグを
しながら別れを告
げている。

=

結果:
このHypothesisは
画像との関係で「2.
Neutral(中立)」に
当たります。

2.画像の含意関係認識とは

前提(Premise)
画像

+

仮説
(Hypothesis)
文

=

結果:
1)含意である
2)中立である
3)反駁している



+

姉妹たちは昼食を食ったばかりで、持ち帰り用の荷物を持ちながらハグをしながら別れを告げている。

=

結果:
このHypothesisは画像との関係で「2. Neutral(中立)」に当たります。

理由: 画像から確認できるのは、2人の女性が何かの容器や袋を持ってハグをしていることです。彼女たちが姉妹であるか、昼食を食ったばかりであるか、また持ち帰り用の荷物を持っている理由は画像からは分かりません。したがって、Hypothesisが真であるかどうかを画像の情報だけで確定することはできません。このため、中立のカテゴリに分類されます。

2.画像の含意関係認識とは

前提(Premise)
画像

+

仮説
(Hypothesis)
文

=

結果:
1)含意である
2)中立である
3)反駁している



+

男たちはデリの
外で争っている。

=

結果:
このHypothesisは
画像との関係で
「3. Contradiction
(矛盾)」に当たり
ます。

2.画像の含意関係認識とは

前提(Premise)
画像

+

仮説
(Hypothesis)
文

=

結果:
1)含意である
2)中立である
3)反駁している



+

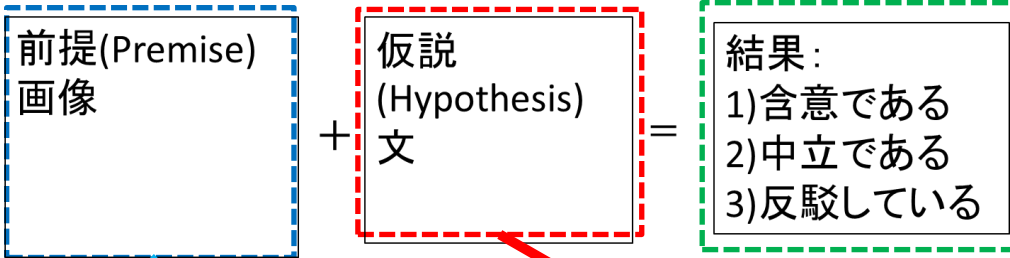
男たちはデリの
外で争っている。

=

結果:
このHypothesisは
画像との関係で
「3. Contradiction
(矛盾)」に当たり
ます。

理由:画像には2人の女性が写っており、彼女たちはハグをしている様子です。画像に男性は写っておらず、また争っている様子も見受けられません。また、画像の背景にはデリ(食料品店)のような場所も見受けられません。このため、提供されたHypothesisは画像の内容と矛盾していると言えます。

3.マーケティングへの応用:お店画像とクチコミの相性



お店が提供するイメージ
(画像)



お客様のクチコミ(文章)

- お店が提供するイメージとおお客様の受取る印象が一致しているとよいと考えます。

3.マーケティングへの応用: お店画像とクチコミの相性



画像をPremiseとした場合、下記のHypothesisを与えると、1.Entailment、2. Neutral、3. Contradictionのどれに当たるかその理由も含めて教えてください。

お店が提供するイメージ
(画像)

お客様のクチコミ(文章)

Hypothesis:

祇園は観光客が増え、とても賑やかな場所になりましたが、こちらのお店は外国人客も少なく、静かでとても良かったです。人気のお店なので予約はされた方が良いでしょう。



1. Neutral

理由: 画像には、日本の伝統的な料理や内装が映し出されていますが、祇園が観光客が増えて賑やかな場所になっているか、そのお店が静かであるか、外国人客が少ないかどうか、そして予約が必要かどうかに関する具体的な情報は含まれていません。したがって、Hypothesisの内容が画像から直接導き出されるわけではないので、Neutralと判断できます。

- お店が提供するイメージとお客様の受取る印象が一致しているとよいと考えます。

3.マーケティングへの応用：お店画像とクチコミの相性



1. Neutral

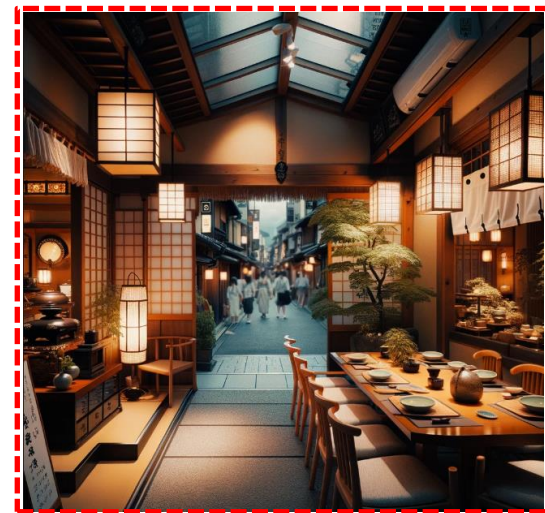
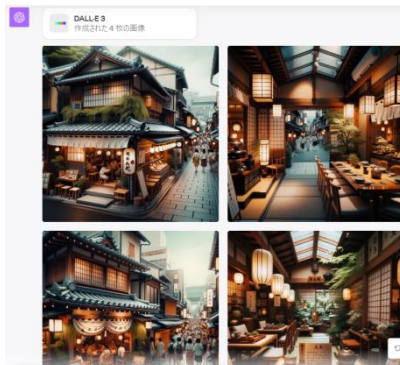
理由: 画像には、日本の伝統的な料理や内装が映し出されていますが、祇園が観光客が増え賑やかな場所になっているか、そのお店が静かであるか、外国人客が少ないかどうか、そして予約が必要かどうかに関する具体的な情報は含まれていません。したがって、Hypothesisの内容が画像から直接導き出されるわけではないので、Neutralと判断できます。

OT

以下は京料理のクチコミです。
このクチコミに見合う画像を生成してください。

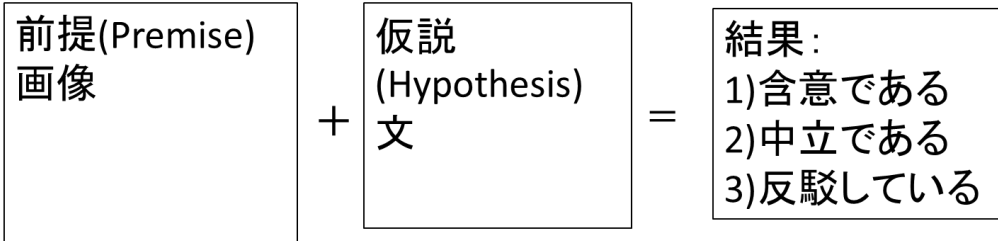
#

祇園は観光客が増え、とても賑やかな場所になりましたが、こちらのお店は外国人客も少なく、静かです。良かったです。人気のお店なので予約はされた方が良いでしょう。



- お店が提供するイメージとお客様の受取る印象が一致しているとよいと考えます。

4. 考察/まとめ



- 画像と文章の含意関係のタスク(Visual Entailment(VE))のフレームワークを応用すると、例えば、飲食店のクチコミの評価に適用できそうです。
- 画像のキャプションも、前後の文脈が示されないと、網羅的な回答になりそうです。
→ 網羅的な回答は常識的に考えておかしいと言えることがありました。
- [課題]”Visual Commonsense reasoning”等で、除外や修正できると実用的に有用になるかもです。

補足：画像の常識推論 (“Visual Commonsense reasoning”)



+

姉妹たちは昼食を
食べたばかりで、
持ち帰り用の荷物を
持ちながらハグを
しながら別れを告
げている。

=

結果：
このHypothesisは
画像との関係で「2.
Neutral(中立)」に
当たります。

Hypothesis(仮説)

- Hypothesis自体のキャプション文章は日常的にありうると考えられる。

補足：画像の常識推論 (“Visual Commonsense reasoning”)

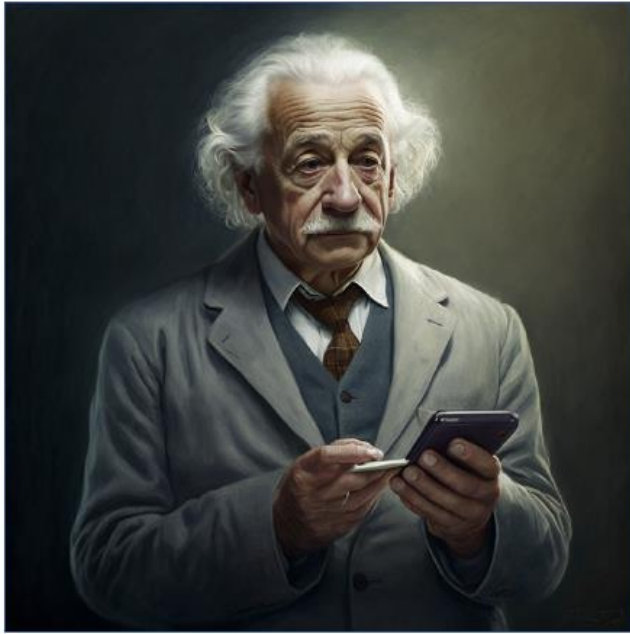


市内の路上で心からの抱擁を交わす2人の女性。1人は白い容器を持ち、もう1人はフォルダーを握り締めている。その後ろでは建設機械やバイクなどの街の喧騒が続いている。

Hypothesis(仮説)

- Hypothesis自体のキャプション文章はありうると考えられる。

補足: 画像の常識推論 ("Visual Commonsense reasoning")



**Albert Einstein holding
a smartphone**



**Einstein's death (1955)
was before the modern
smartphone was invented
(2007).**

アインシュタインがスマートフォンを持っている

⇔

アインシュタインは1955年になくなっている。

スマートフォンは2007年に発明された。

※ ["Breaking Common Sense: WHOOPS!"](#)より引用

参考文献・URL一覧

1. Breaking Common Sense: WHOOPS! A Vision-and-Language Benchmark of Synthetic and Compositional Images

<https://arxiv.org/abs/2303.07274>

2. 祇園 京料理 花咲 口コミ - 楽天ぐるなび <https://r.gnavi.co.jp/k290600/review/>

3. The Dawn of LMMs:

Preliminary Explorations with GPT-4V(ision)

<https://arxiv.org/pdf/2309.17421.pdf>

4. Visual Entailment: A Novel Task for Fine-Grained Image Understanding

<https://arxiv.org/pdf/1901.06706.pdf>

5. The Dawn of LMMs:

Preliminary Explorations with GPT-4V(ision)

<https://arxiv.org/pdf/2309.17421.pdf>

※1 [ChatGPT部特別企画] GPT-4Vのポテンシャルを探る！

LT大会@東京・羽田空港会場 - connpass

<https://chatgptclub.connpass.com/event/298703/>

※2 github otanet

https://github.com/otanet/LT_GPT-4V_VE_20231021

