

画像キャプションを利用した IconQA タスクへのアプローチ

塩野大輝¹ 宮脇峻平^{1,2} 長澤春希¹ 鈴木潤^{1,3}

¹ 東北大学 ² 株式会社キーウォーカー ³ 理化学研究所

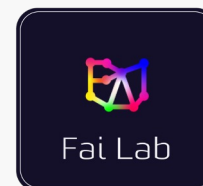
✉ daiki.shiono.s1@dc.tohoku.ac.jp

🐦 @onely7_deep

🌐 <https://github.com/Onely7-nlp>

言語処理学会第29回年次大会, 2023-03-15

TOHOKU
NLP
東北大学
自然言語処理
研究グループ
GROUP



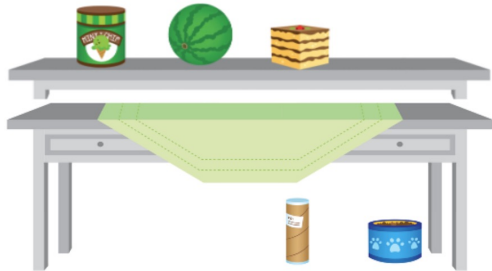
keywalker



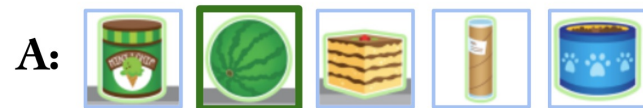
IconQA: ダイアグラム画像に対する質問応答 [Lu+'21]

- 抽象的なダイアグラム画像に対する Visual Question Answering タスク

画像候補選択



Q: Which object is next to the one shaped like a cube?



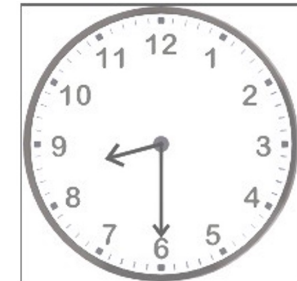
テキスト候補選択



Q: Which picture shows the pizza inside the oven?

A: (A) left one (B) right one

穴埋め

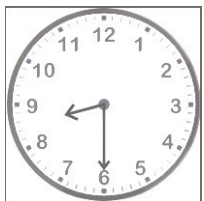


Q: What time is shown?
Answer by typing a time word, not a number. It is () past eight.

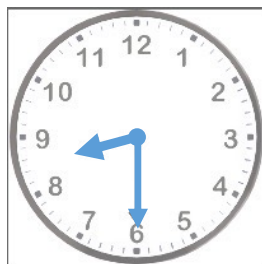
A: half

IconQA タスクに必要な能力

1. 視覚と言語の意味表現を関連付ける



Q: What time is shown?
Answer by typing a time word,
not a number. It is () past eight.



2. 質問を正しく読解して推論する

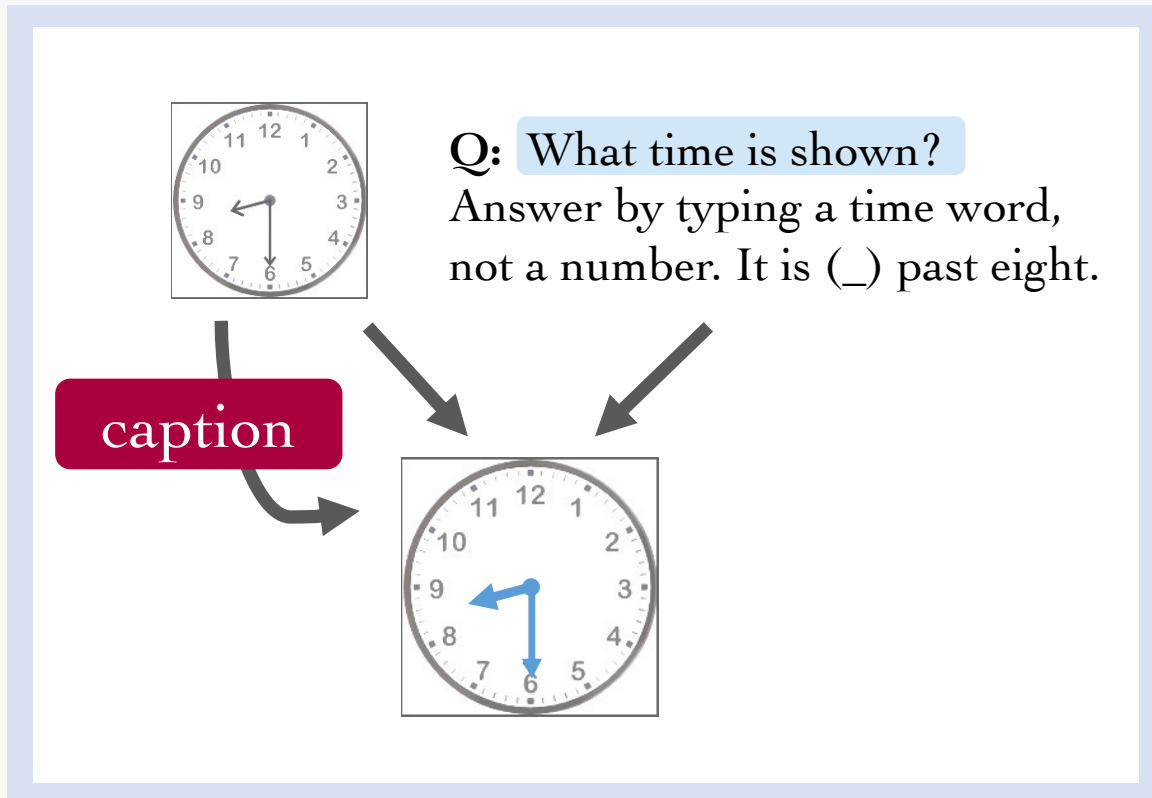
Q. What time is shown?
Answer by typing a time word,
not a number.
It is () past eight.



A. half

IconQA タスクに必要な能力

1. 視覚と言語の意味表現を関連付ける



ダイアグラム画像において
言語による視覚情報の拡張が
読解性能に効果的か？

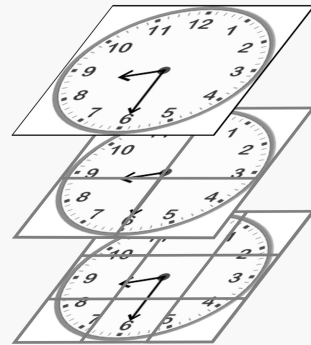
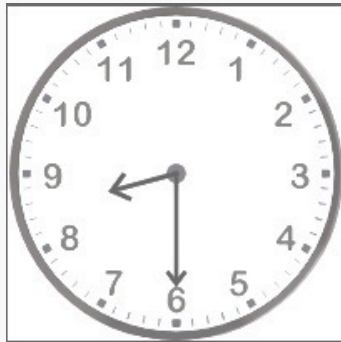
画像を言語で記述することでモデルに解答の手がかりを提供する

質問

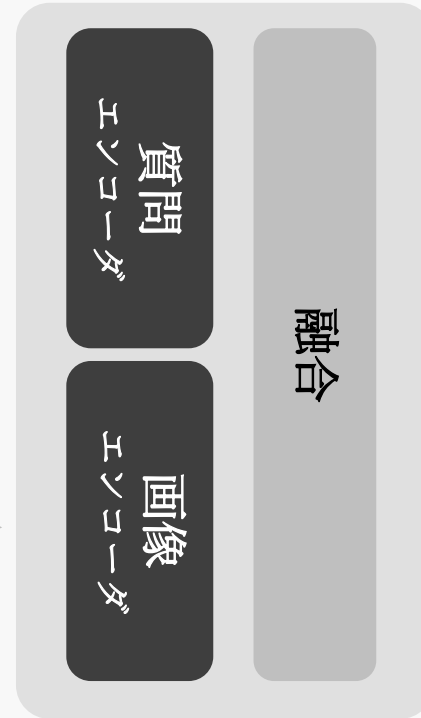
Q: What time is shown?

Answer by typing a time word, not a number.
It is () past eight.

ダイアグラム画像



画像分割



Patch-TRM
[Lu+'21]



解答
A: half

画像を言語で記述することでモデルに解答の手がかりを提供する

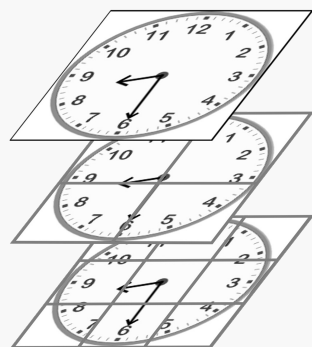
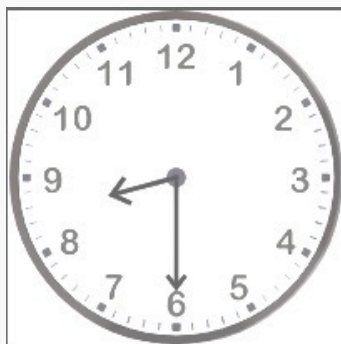
質問

Q: What time is shown?

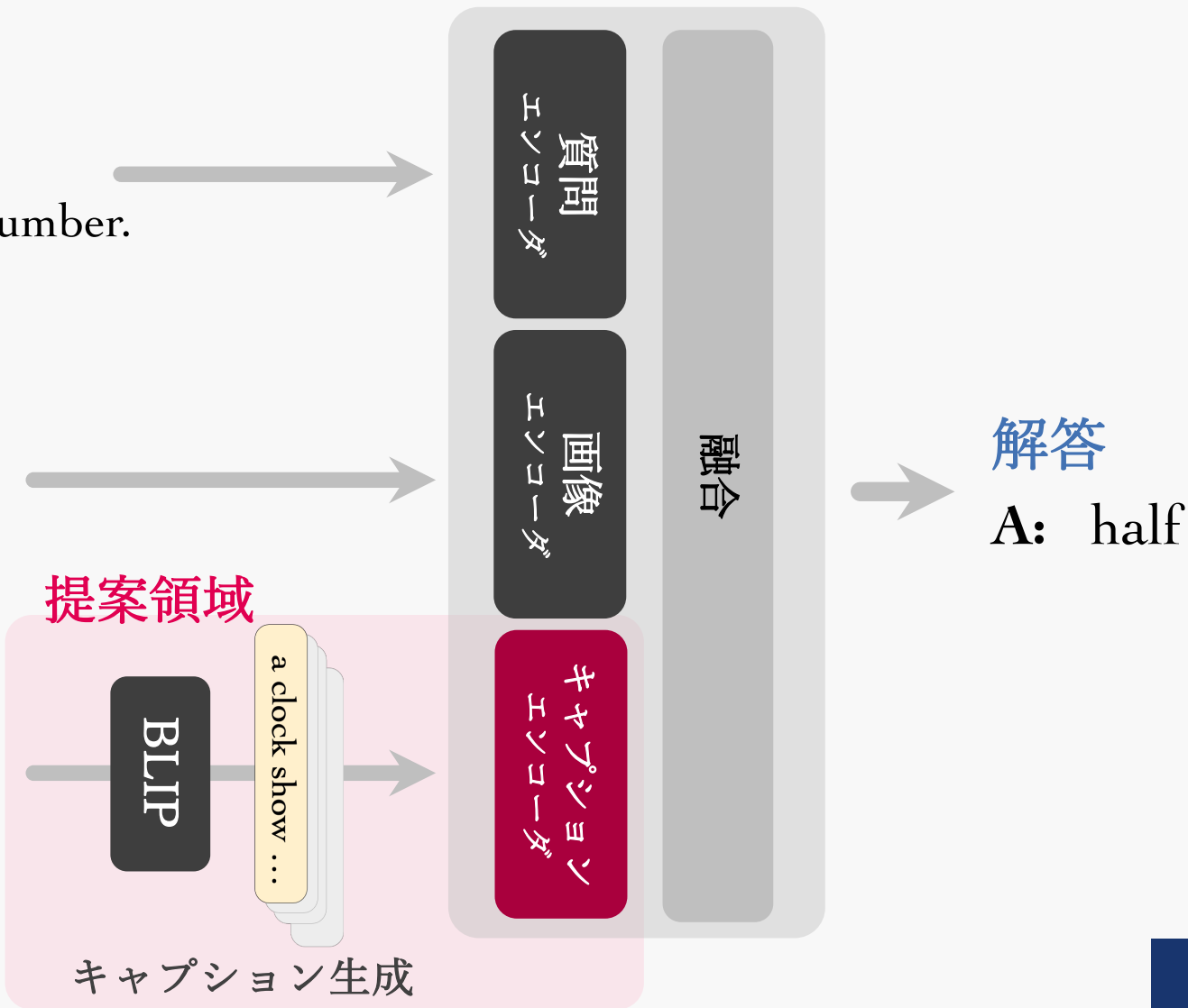
Answer by typing a time word, not a number.

It is () past eight.

ダイアグラム画像

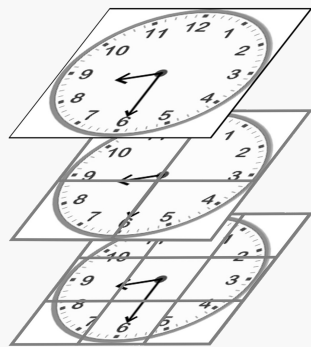


画像分割

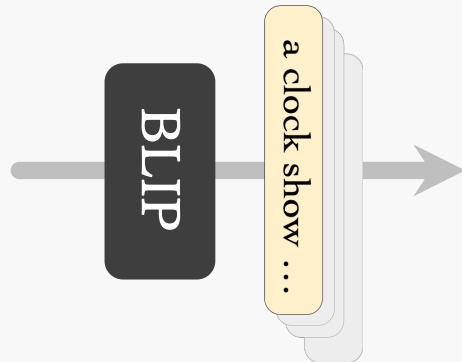


視覚情報の拡張としてキャプションを生成する

視覚と言語の意味関係を学習した BLIP [Li+'22] でキャプションを生成



画像分割



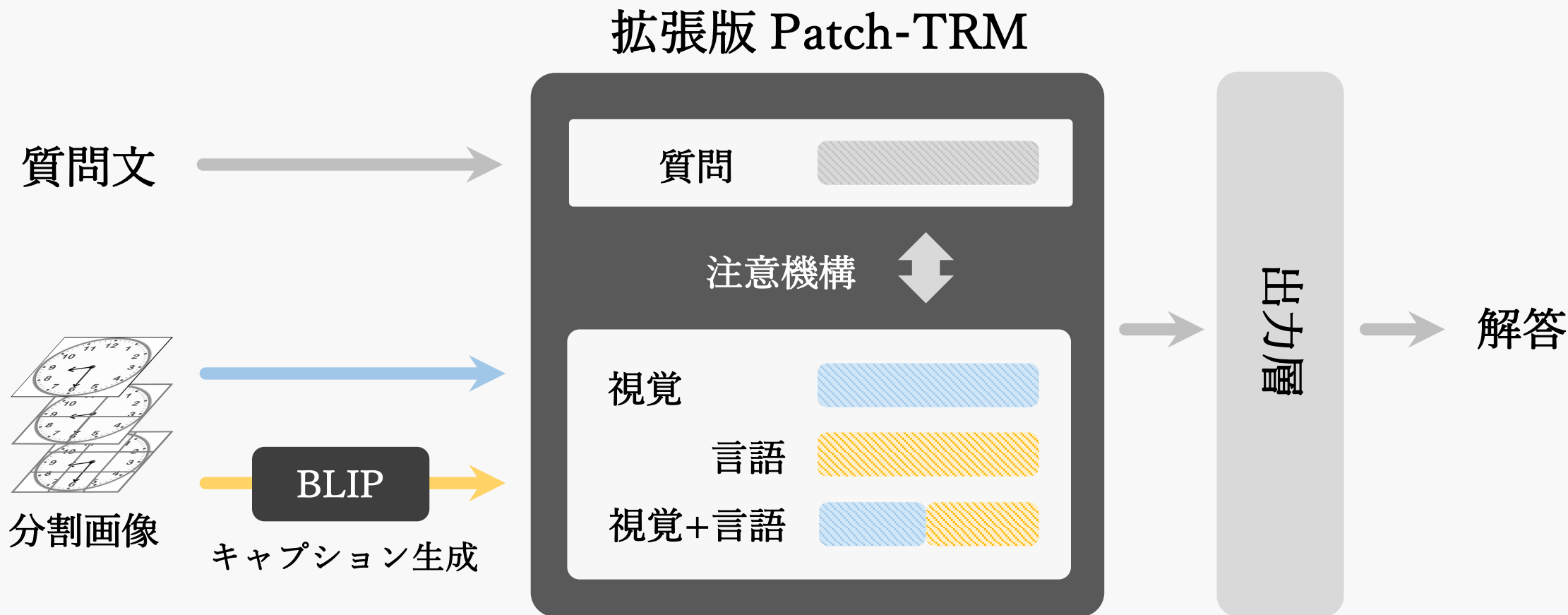
キャプション生成



e.g. BLIP キャプション

A library building has two doors on the front

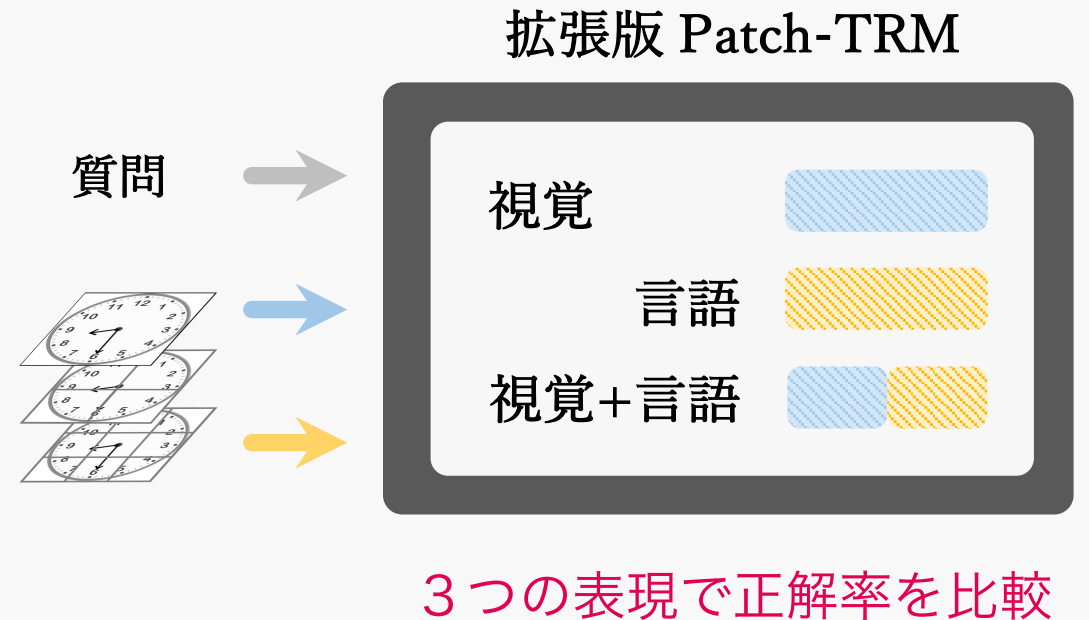
キャプション情報を組み込んだ拡張版 Patch-TRM



3つの表現で正解率を比較




拡張版 Patch-TRM の性能評価

データセット 訓練 / 開発 / 評価	IconQA [Lu+'21] 18,189 / 10,913 / 3,638
画像エンコーダ	ResNet-101 (Icon645 で学習) ViT [Dosovitskiy+'21]
言語エンコーダ	BERT [Devlin+'19]
キャプション生成	BLIP [Li+'22]
パッチサイズ	79 ($1^2 + 2^2 + 3^2 + 4^2 + 7^2$)
キャプション数	14 ($1^2 + 2^2 + 3^2$)



ダイアグラム画像において視覚と言語の融合が読解性能に効果的か？ (1/2)




表2. IconQA サブタスク別の正解率

モデル	入力情報	画像選択	テキスト選択	穴埋め問題
		11535	6316	3638
Patch-TRM	視覚 	78.71	65.92	86.44
	言語 	69.07	62.40	47.16
	両方 	79.60	64.90	87.39
Human		95.69	93.91	93.56

視覚と言語の融合により 画像選択・穴埋め問題にて正解率が向上

ダイアグラム画像において視覚と言語の融合が読解性能に効果的か？ (2/2)

表2. IconQA 推論スキル別の正解率

モデル	入力情報	Geo.	Cou.	Com.	Spa.	Scce.	Pat.	Tim.	Fra.	Est.	Alg.	Mea.	Sen.	Pro.
		8575	7493	2976	2143	2013	1827	1803	1567	1530	1456	1287	1158	1077
Patch-TRM	視覚 	80.9	76.6	74.7	53.5	59.9	56.1	69.9	81.4	97.0	61.0	96.4	79.4	76.3
	言語 	72.1	52.4	68.2	50.0	60.0	54.9	66.9	44.3	65.4	38.9	60.1	82.1	83.9
	両方 	81.1	76.7	74.9	54.7	62.4	55.7	67.7	81.3	99.0	60.9	98.8	78.0	75.2
Human		94.6	97.6	94.4	93.3	92.7	95.7	97.9	97.5	87.5	96.3	86.6	97.1	85.7

言語情報の拡張により 7/13 の推論スキルで正解率が向上

キャプション情報による拡張がダイアグラムの画像読解において有効である可能性を示唆

IconQA タスクに必要な能力

画像をキャプションに変換し、
大規模言語モデルの推論能力を活用して
VQAを解くことができるのか

[Yang+'22]

2. 質問を正しく読解して推論する

Q. What time is shown?

Answer by typing a time word, not a number.

It is () past eight.

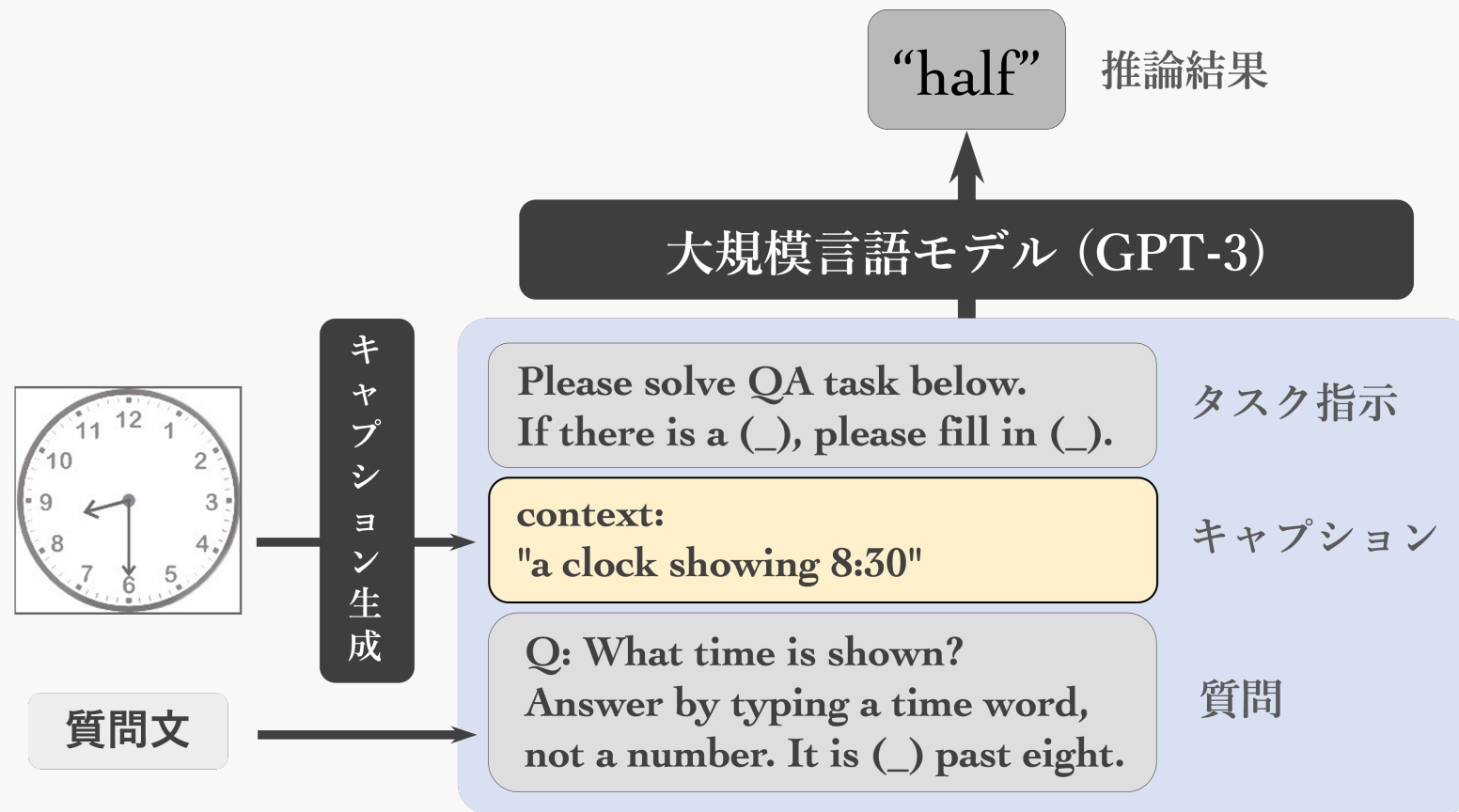


a clock showing 8:30

大規模言語モデル

A. half

GPT-3 によるゼロショット推論

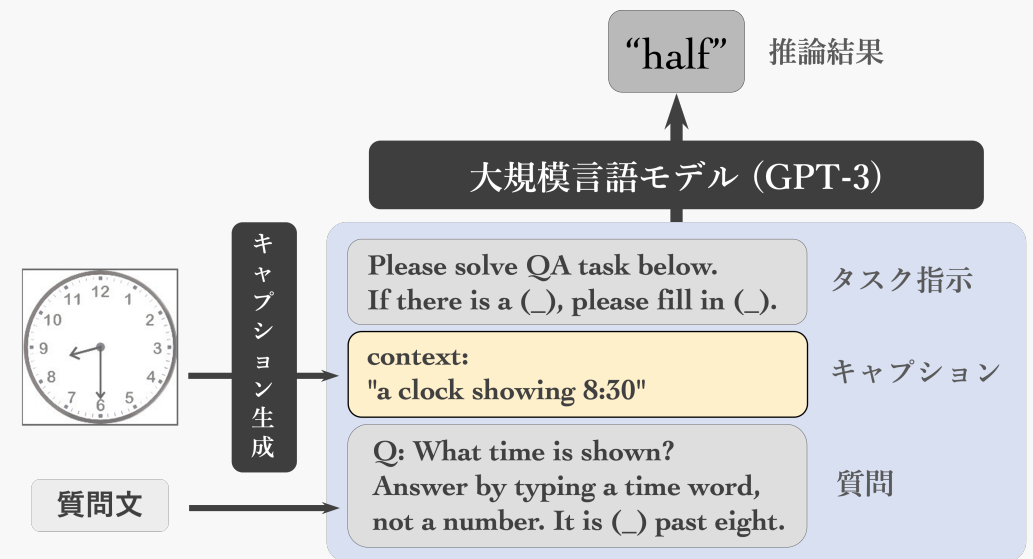


大規模言語モデルの推論能力を活用するため
人手 / BLIP を用いて視覚情報を言語化

GPT-3 を用いたゼロショット推論

- 言語モデル別の性能差を調査すべく、Patch-TRM でも同様に評価

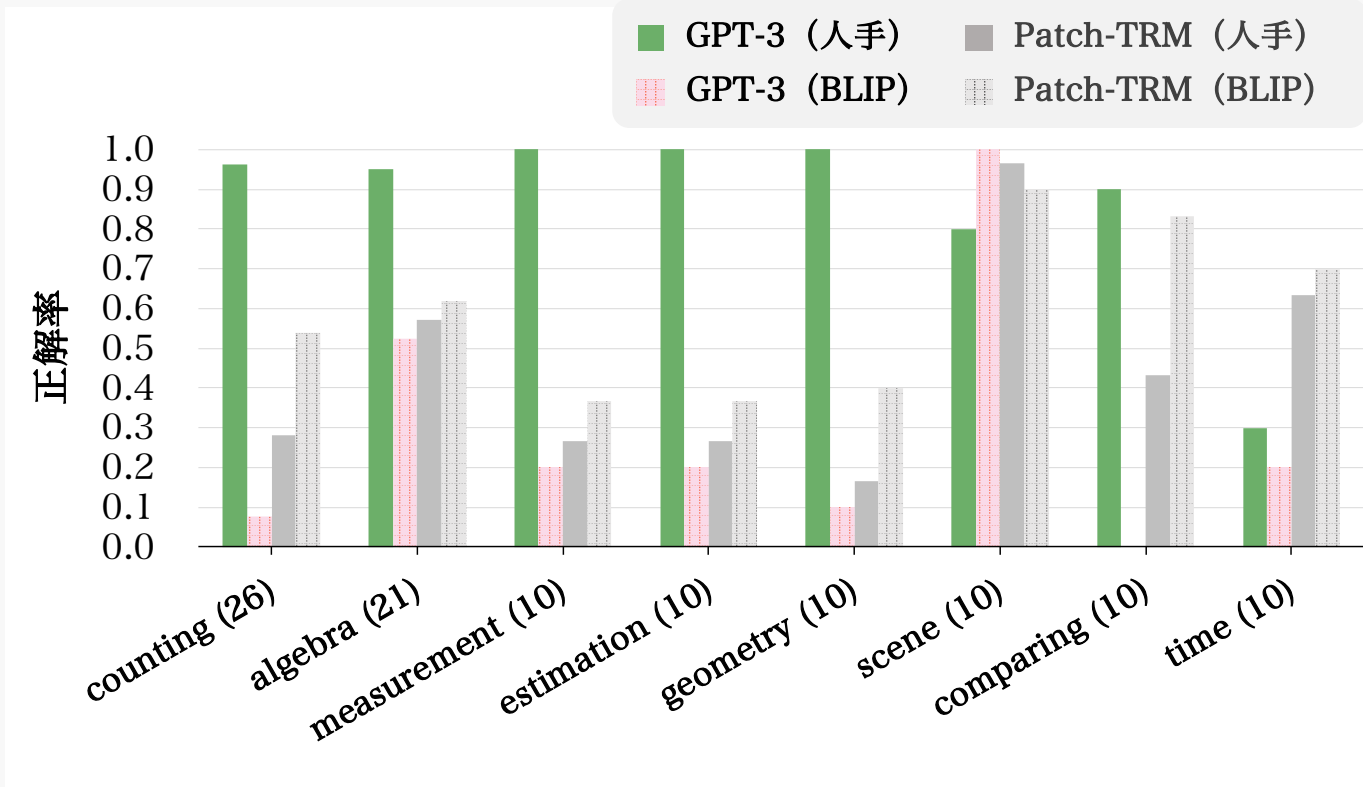
対象タスク	穴埋め問題
評価データ	57 件 (無作為抽出)
読解モデル	GPT-3 [Brown+'20]
キャプション生成モデル	人手 / BLIP [Li+'22]
評価指標	正解率
キャプション数	1



人手/BLIP によって視覚情報を言語化

大規模言語モデルの推論能力は IconQA にも有効か？

図3. IconQA 穴埋め問題の推論スキル別の正解率



適切なキャプションを用いることで読解性能が向上することを示唆

BLIP による記述内容、Patch-TRM のモデルアーキテクチャについて見直す必要があることを示唆

おわりに

画像キャプションを利用した IconQA タスクへのアプローチ

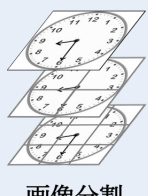
提案手法①：視覚と言語の意味表現を関連付ける

質問

Q: What time is shown?

Answer by typing a time word, not a number.
It is () past eight.

ダイアグラム画像



画像分割

提案領域

BLIP

a clock show...

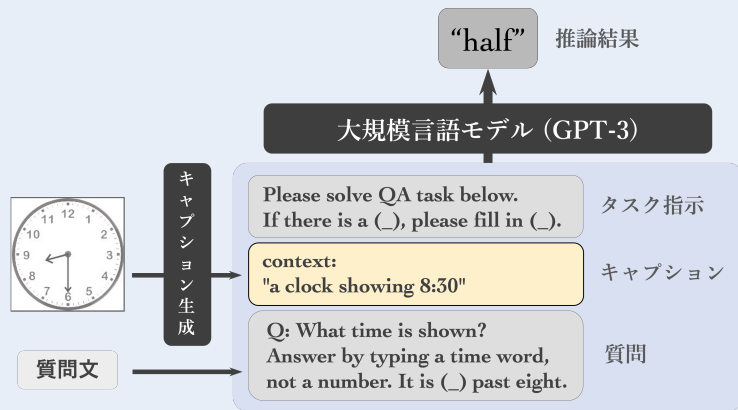
キャプション生成



解答

A: half

提案手法②：質問を正しく理解して推論する



結論

- キャプションによる視覚情報の拡張がダイアグラム画像の読解で有効である可能性を示唆
- 大規模言語モデルの推論能力がダイアグラム画像の読解で有効である可能性を示唆

将来の展望

- 視覚と言語の融合方法について調査
- 推論スキル別に適切な言語情報があるか調査
- モデルアーキテクチャの改善