P02

マニュアルやパンフレットなど絵や図表の多い資料を理解し答えてくれます

グラフィカルな文書を理解できる「tsuzumi」

IOWN Pick Up　NTT版大規模言語モデル

背景

大規模言語モデルの多くはテキスト情報しか理解できず、我々が普段扱う文書に含まれるグラフ、アイコン、イラスト、文字の大きさ、レイアウトといった視覚情報を理解できないという課題があります。

成果の概要

文書の画像から内容を理解し、ユーザの求める情報を提示する技術です。大量の文書画像から学習した大規模言語モデルにより、文書画像に関する自然文の指示に基づいて返答を出力します。また本モデルは、さまざまな文書・フォーマットに対応できます。

技術のポイント

- 文書画像を視覚的に読み解く技術およびLLMとの融合技術の開発で世界をリード（AAAI'24投稿中、AAAI'23・AAAI'21採択、InfographicsVQAコンペ2位）
- 数億ペアの日本語のテキストと画像から独自に学習した高精度な画像エンコーダを利用
- 世界最大規模の文書画像データセットを構築し、図表、webページ、手書き書類を含む多様な文書・指示が理解できる大規模言語モデルを学習

この研究がもたらす未来

文書情報を始めとした「人の目に映る世界」を言語と結び付けて理解することにより、オフィスで、さらには、あらゆる環境で、人間と協調して価値を創出する人工知能の実現をめざしています。

出展企業

日本電信電話株式会社

問い合わせ先

rdforum-exhibition@ml.ntt.com

Copyright © 2023 Nippon Telegraph and Telephone Corporation All Rights Reserved.