AI 活用で挑む学問の革新と創成　　2022 年度
2022 年度採択研究代表者　　年次報告書

八木 拓真

東京大学 生産技術研究所
特任研究員

大規模言語モデルからの知識抽出に基づく視覚スクリプトの創成

研究成果の概要

画像からのスクリプト予測の実現に向けた初期検討を行った。具体的には、(i) 画像からのスクリプト予測に必要な要素は何か (i) 既存の画像キャプショニングモデルおよび大規模言語モデル（LLM）を素朴に組み合わせた場合に人が見て妥当なスクリプトが生成できるか (iii) あるスクリプトを生成するための主役およびゴール（目標）も同様に生成できるか の 3 項目に関する検証を行った。

まず、(i)について、Shank & Abelson のオリジナルのスクリプトの定義を参考に場所、主役、ゴール、脇役、物体といったスクリプトに付随する構成要素を列挙し、約 40 枚の画像に対して人手で 122 件のアノテーションを行った。その結果、画像から妥当なスクリプトを生成するためには、画像に含まれない場所、主役、ゴールに関する情報が必要であることを見出した。続いて(ii)において、画像および場所・主役・ゴールに関する情報を入力として、最新の画像キャプショニングモデルおよび LLM（GPT-3）を in-context learning に基づいて組み合わせたスクリプト予測のための素朴なベースラインモデルを提案し、先述の 122 件のサンプルよりスクリプトを生成、その妥当性の人で評価を行った。また、(iii)についても類似の枠組みを用い、画像のみから主役よびゴールを生成するモデルを提案し同様に評価した。人手評価の結果、画像説明文などを GPT-3 に入力した場合に 9 割以上のサンプルにおいて人が見て画像に対して妥当なスクリプト・ゴールを生成できることを確認し、素朴なモデルでも LLM が実際にスクリプト知識を有すること、より厳密な評価指標が必要であることが示唆された。