Large Vision Language Model への文書画像内テキスト埋め込みの検証

會田勇斗 [1]　陳実 [1]　森長誠 [1]　近江崇宏 [1]　有馬幸介 [1]
[1] ストックマーク株式会社
{hayato.aida,shi.chen,makoto.morinaga,takahiro.omi,kosuke.arima}@stockmark.co.jp

概要

Large Vision Language Model (LVLM) は近年急速に進化しており、文書画像理解のタスクにおいても End-to-End で高い精度を達成している。一方で、GPT-4o クラスのモデルであっても画像内の文字認識誤りが発生することがあり、ビジネスでの実運用において課題となっている。本研究では、LVLM に文書画像内テキストを埋め込むことで、高精度な文書画像理解を実現することを目指す。既存の LVLM に文書画像内テキストをプラグイン的に埋め込む方式を提案し、その有効性を検証した。

1　はじめに

近年、LVLM の研究が盛んに行われている。写真のような自然画像から、グラフや図表など、高い精度で画像内の情報を理解することができる。特に 2024 年以降の LVLM はオープンソースのモデルでも画像入力の高解像化が進み、画像内のテキストを高精度で理解することが可能になった。一方で、ビジネスで利用される文書には高密度なテキストが含まれており、高解像度の LVLM であっても、小さい文字や装飾文字などで文字認識の誤りが発生することがある。我々はビジネスドメインでの LVLM の実用化を目指しており、文字認識誤り由来のハルシネーションを防ぐことは重要である。

ビジネスシーンで利用される文書は PDF 形式である場合が多く、文書画像内のテキストはその位置情報と共に取得することができる。また、スキャンした文書画像であっても、活字であれば OCR によりテキストとその位置情報を高精度で取得することができる。従って、LVLM の文字認識由来の誤りを防ぐためには、これらの文書画像内テキストを活用することが有効であると考えられる。

実際にエンコーダ系の分類モデルまで遡ると、文書画像内テキスト埋め込みの研究は一定の成果があり、最近ではそれらの手法の影響を受けて LLM,LVLM でも文書画像内テキスト埋め込みの研究が行われている。LayoutLM シリーズ [1, 2, 3] は文書画像内のテキストを活用するエンコーダ系モデルであり、文書分類、質問応答、情報抽出などのタスクで高い精度を達成している。また、LLM, LVLM では InstructDr[4], LayoutLLM[5], LayTextLLM[6], DocLLM[7] などのモデルが提案されており、文書内の OCR テキストとその位置情報を埋め込む。CLIP の Image Encoder を用いた LLaVA 系 [8] の LVLM では画像、文字、グラフ読み取りの性能が飛躍的に向上する一方で、文書画像理解に特化し、テキストとその bbox 座標を埋め込むマルチモーダル LLM は独自のアーキテクチャを採用している場合が多く、昨今の LLaVA 系モデルの性能向上の恩恵を十分に受けられていない。

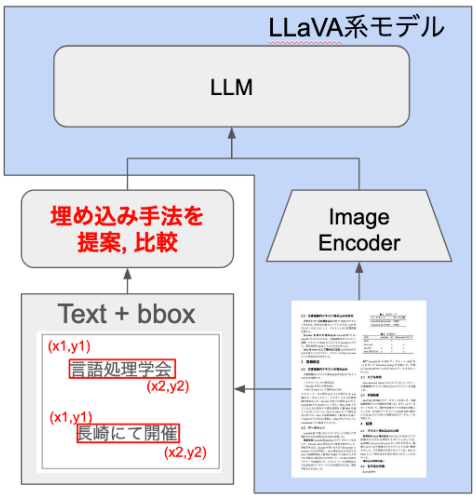

図 1　提案手法の概要

本研究では、LVLM に文書画像内のテキストをプラグイン的に埋め込む方式を提案し、LLaVA 系モデルの画像認識能力を活かしつつ、文書画像内テキストの読み取り精度を向上させることを目指す。その

ためにいくつかの埋め込み方式を提案、比較し、有効性を検証した。

2 関連研究

2.1 OCR free の LVLM

InternVL2[9, 10], Qwen2VL[11], LLaVA-onevision[12] などのモデルは、モデルアーキテクチャと学習データセットの工夫により、従来の LLaVA[8] や BLIP2[13] などのモデルよりも高い精度で画像内のテキストを理解し、文書画像理解タスクを解くことができる。アーキテクチャでは画像入力を高解像度で入力できるような改善が各モデルで共通しており、学習データセットでは大規模な OCR データセットを追加することで、テキスト理解の精度を向上させている。

2.2 文書画像内テキストと座標情報の埋め込み

文書画像内のテキストは、PDF や OCR のテキストとして取得することができる。PDF, OCR のテキストにおける共通した情報は、テキストとテキストを囲む bounding box(bbox) である。テキストと bbox を LLM に入力する方法はいくつか提案されているため、以下にその手法を示す。以後、文書画像内のテキストと bbox を text+bbox と呼ぶ。

テキストベースの埋め込み text+bbox を json 形式で入力する最もシンプルな手法 [14]。

Encoder を用いた埋め込み LayoutLLM[5] や InstructDr[4] などのモデルは、文書画像内のテキスト、画像、テキストの bbox を入力とする Attention 機構を持つ Encoder に入力し、潜在表現を LLM に入力する。

bbox 情報を埋め込む手法 LayTextLLM[6] や DocLLM[7] などのモデルは、単純な MLP を用いて bbox の情報を埋め込み、文書画像内のテキストは LLM の tokenizer を用いて入力する。

3 提案手法

3.1 文書画像内テキストの埋め込み

文書画像内テキストの埋め込み方式を以下の 3 つの手法で比較する。

DocPTM : LayoutLLM[5] では、大規模な文書画像データセットで学習された Document Pre-trained Model(DocPTM) を用いて文書画像内テキストを埋め込むことを提案している。今回は LayoutLLM の論文でも採用されている LayoutLMv3 を DocPTM として用い、text+bbox を LVLM に入力した。text+bbox は DocPTM と通して潜在表現に変換し、2 層 MLP を通して LLM の入力次元に変換、埋め込みを行った。

bbox-text : 先述したテキストベースの埋め込みを採用し、text+bbox を json 形式に変換し、LLM に入力した [14]。

bbox-proj : LayTextLLM[6] で提唱された、bbox を MLP で埋め込み、テキストと共に LLM に入力する方式。bbox の座標情報を 2 層 MLP(projector) を通して LLM の入力次元に変換し、テキストと concatenate して LLM に入力する。具体的には、一つの画像内テキストを位置情報と合わせて bbox1,text1 とすると次のような埋め込みを LLM に入力する。projector(bbox_1), embed(tokenize(text_1)), ...

表 1 学習データ

データセット	サンプル数
LayoutLLM-pretrain	115,955
LayoutLLM-sft 50%	149,528

3.2 データセット

表 1 に示すように、LayoutLLM にて提案されたデータセット [5] を用いて実験を行った。これらのデータセットは文字認識やレイアウト理解に重きを置いたデータセットであり、文書画像内テキストの埋め込みの有効性を検証するために選定した。

text+bbox 入力部分の事前学習 LVLM に新たなパラメータを追加した方式で実施した、事前学習について説明する。LayoutLLM-pretrain のデータセットの一部を用いて、docPTM, bbox-proj の事前学習を行った。DocPTM を用いる方式では DocPTM と projector のみ重みを更新し、bbox-proj では bbox の座標情報を 2 層 MLP を通して LLM の入力次元に変換する projector のみ学習した。なお、LVLM の本体のパラメータは固定した。bbox-text では LVLM のパラメータのみを利用するため、事前学習は行わなかった。

Supervised Fine-Tuning (SFT) LayoutLLM の SFT データセット (Chain of Thought なし) を用いて Instruction tuning を実施した。学習は LayoutLLM SFT の前半 50%のデータを用いて行った。

表 2 学習条件

手法	pretrain	sft	text+bbox
llava-ov-7b-sft	-	o	-
docPTM	o	o	o
bbox-text	-	o	o
bbox-proj	o	o	o
bbox-proj w/o PT	-	o	o

表 3 評価データ

データセット	サンプル数	split
DocVQA	5349	val
FUNSD	467	val (layoutLLM)
ChartQA	1920	val

3.3 モデル学習

表 2 に実施した学習条件を示す。比較的少ない学習データで高解像の画像理解を実現していることから、全ての実験は llava-onevision-qwen2-7b-ov(llava-ov-7b)[12] のモデルをベースにして行った。先述した 3 種類の text+bbox の入力方式を実装し、llava-ov-7b から追加学習することで、文書画像内テキストの埋め込みの有効性や妥当な入力方法を検証した。

3.4 評価指標

評価に利用したデータセットについて表 3 に示す。DocVQA[15], FUNSD[16] のデータセットを用いて、文字認識に関わる QA 性能を評価した。また、視覚的な図表理解性能を監視するために、ChartQA[17] による定量評価も実施した。なお、FUNSD データセットは LayoutLLM の論文にて QA モデル向けに変換されたデータセットを利用した。

4 結果

4.1 text+bbox 入力の効果

表 4 にて、画像入力のみと bbox-proj, bbox-text, docPTM の 3 つの埋め込み方式の性能を比較した。DocVQA, FUNSD では、bbox-proj, bbox-text, docPTM のいずれの方式も、画像のみの入力で訓練されたモデルよりも高い精度を示し、文字認識系のタスクにおいて text+bbox の入力が有効であることが確認できた。DocPTM では精度向上がわずかであるが、bbox-proj, bbox-text では精度が大幅に向上している。これは、bbox-proj, bbox-text の方式では、文書画像中

表 4 評価結果 (ANLS) ChartQA では画像のみを入力

手法	DocVQA	FUNSD	ChartQA
llava-ov-7b	0.401	0.227	**0.610**
llava-ov-7b-sft	0.449	0.267	0.348
docPTM	0.464	0.287	0.349
bbox-text	0.692	0.739	0.309
bbox-proj	0.705	0.729	0.252
bbox-proj w/o PT	**0.706**	**0.755**	0.292

のテキストを LLM へ直接入力するため、LLM が事前学習で獲得した豊富な表現力を活用できるためであると考えられる。docPTM では、transformer の encoder と projector を介して text+bbox を入力するため、LLM の表現力を十分に活用できていない可能性や、事前学習タスクの量と多様性が不足しており、DocPTM と LLM のアライメントが十分取れていない可能性が考えられる。

文字認識タスクにおいて最も高い精度を示したのは、bbox-proj w/o PT であった。より詳細な分析や、視覚的図表理解の精度とのトレードオフについて以下で述べる。

bbox-proj の事前学習 bbox-proj について、事前学習の有無で 2 種類の手法を比較した。結果として、DocVQA, FUNSD, ChartQA の全てのデータセットにおいて、事前学習を行わない場合の方が性能が高かった。これには 2 つの理由が考えられる。1 つ目は、事前学習タスクと SFT, 評価タスクの性質の違いによるものである。今回利用した LayoutLLM の事前学習タスクは、bbox の座標を回答させるものであり、SFT や評価で用いられる意味理解、情報抽出系の QA とは性質が異なる。2 つ目は、bbox-proj は 2 層 MLP でありパラメータが非常に少ないため、事前学習による効果が低くタスク転移へのロバスト性が低い可能性が考えられる。事前学習タスクの設定を見直し、より SFT に近いタスクを設定することで、事前学習の効果を高めることを今後の課題とする。

ChartQA の視覚的図表理解 LayoutLLM のデータセットで SFT を行った場合、公開モデルの llava-ov-7b よりも精度が低下した。まず llava-ov-7b-sft において、llava-ov-7b よりも精度が低下している。このことから、この精度低下の主な原因は text+bbox の入力ではなく、LayoutLLM-sft データセットの特性に起因するものであると考えられる。LayoutLLM-sft データセットは、テキスト認識に偏ったデータセッ

トであるため、このデータセットで学習したことにより、LLaVA の学習で獲得した視覚的図表理解能力が低下したと考えられる。さらに、bbox-proj, bbox-text の方式では、llava-ov-7b-sft よりも精度が低下している。先述したように、bbox-proj, bbox-text の方式では、LLM へ文書画像中のテキストを直接入力するため、LLM のパラメータと十分にアライメントが取れており、LLM が持つ豊富な表現力を活用することができる。その副反応的に画像情報への注意を相対的に下げるように学習が進み、画像のみの入力よりも精度が低下したと考えられる。

以上のことから、SFT データセットのテキスト認識タスクへの偏りの軽減や、text+bbox が存在しない条件を学習に含めることで、図表理解性能の劣化を防ぐことが今後の課題となる。

4.2 画像入力の効果

文書画像理解タスクにおける LLaVA 系モデルが持つ画像理解能力の効果を検証するために、text+bbox の入力方式において画像入力を省略した場合の精度を比較した。表 5 に示すように、DocVQA タスクにおいては、text+bbox の入力にどの方式を用いても、画像入力がある場合の精度が高い。これにより、文字認識を中心としたベンチマークであっても、一部のタスクにおいては画像情報に基づき回答を生成していることが考えられる。一方 FUNSD において, bbox-proj では画像入力が有りで精度が高いがその差はわずかであり、bbox-text では画像入力無しで精度が高くなっている。docPTM の場合は FUNSD, DocVQA 共に画像入力を省略すると大幅に精度が低下し、ほとんど回答ができない結果となった。docPTM の入力を LLM があまり活用できておらず、文書画像理解の多くを画像入力に依存するような学習をしていることがわかる。

FUNSD, DocVQA のスコア挙動の違い 二つのベンチマークの違いは、入力画像および質問の抽象度の違いにある。DocVQA は FUNSD に比べて図や表を含む文書画像が多く、画像情報も回答の参考になるため、画像入力を無くした場合の影響が大きい。また、FUNSD では"What is the content in the "DATE:" field?"のように、文書画像内の参照すべき箇所が"DATE:"とテキストで与えられており、注目箇所が明確で文書全体の意味を理解せずに回答ができる。一方で DocVQA では、"Where is the university located ?"のように、テキストの参照箇所がテキストで与えられておらず、画像内のどの箇所を参照すべきかを推論する必要がある。従って、DocVQA では画像全体の構造を理解することが重要であり、画像入力の有無で精度の差が出ると考えられる。

表 5 画像入力の効果 (ANLS)

手法	DocVQA	FUNSD
docPTM	**0.464**	**0.287**
docPTM w/o image	0.067	0.001
bbox-text	**0.692**	0.739
bbox-text w/o image	0.668	**0.741**
bbox-proj	**0.705**	**0.729**
bbox-proj w/o image	0.686	0.728
bbox-proj w/o PT	**0.706**	**0.755**
bbox-proj w/o PT, image	0.694	0.753

画像入力でのトークン消費することを考慮すると、想定質問や用途の抽象度が低くほとんどテキストで構成される画像の場合は、画像入力の代わりに text+bbox のみを入力することで、計算コストを削減しつつ、精度を維持することが期待できる。一方で、視覚的要素を含む画像や抽象度が高い質問に対しては画像を含めた入力が有効であるため、より汎用的な用途では画像入力の効果は大きいと言える。

5 おわりに

LVLM へ text+bbox を入力する手法を提案し、文字読み取り能力を向上させることを確認した。LVLM の文字認識系タスクの性能向上には、提案手法の中では bbox-proj の方式が最も有効であることが分かった。text+bbox を埋め込む方法は文書画像の分類や固有表現抽出などのタスクでも有効性が示されており、LayoutLMv3 がその代表格であったが、LVLM との接続においては LayoutLMv3 が best な選択では無かった点は興味深い。LLM が言語による大規模な事前学習や SFT によって獲得する知識活用の重要性を示唆しており、文書画像理解のみらずマルチモーダルモデル全般において意識すべき課題であると考えられる。

課題として、有効にテキスト読み取りタスクを解ける方式ほど、text+bbox の入力がない場合に性能が劣化してしまうことがわかった。本研究においてはテキスト読み取りに着目し、LayoutLLM のデータセットを活用したが、今後はより広いデータセットを用いて検証を進めることで、文書画像理解タスクにおける汎化性能の向上を目指す。

参考文献

[1] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. LayoutLM: Pre-training of text and layout for document image understanding. **arXiv [cs.CL]**, December 2019.

[2] Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. LayoutLMv2: Multi-modal pre-training for visually-rich document understanding. **arXiv [cs.CL]**, December 2020.

[3] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. LayoutLMv3: Pre-training for document AI with unified text and image masking. **arXiv [cs.CL]**, No. 1, April 2022.

[4] Ryota Tanaka, Taichi Iki, Kyosuke Nishida, Kuniko Saito, and Jun Suzuki. InstructDoc: A dataset for zero-shot generalization of visual document understanding with instructions. **arXiv [cs.CV]**, January 2024.

[5] Chuwei Luo, Yufan Shen, Zhaoqing Zhu, Qi Zheng, Zhi Yu, and Cong Yao. LayoutLLM: Layout instruction tuning with large language models for document understanding. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**, pp. 15630–15640, 2024.

[6] Jinghui Lu, Haiyang Yu, Yanjie Wang, Yongjie Ye, Jingqun Tang, Ziwei Yang, Binghong Wu, Qi Liu, Hao Feng, Han Wang, Hao Liu, and Can Huang. A bounding box is worth one token: Interleaving layout and text in a large language model for document understanding. **arXiv [cs.CL]**, July 2024.

[7] Dongsheng Wang, Natraj Raman, Mathieu Sibue, Zhiqiang Ma, Petr Babkin, Simerjot Kaur, Yulong Pei, Armineh Nourbakhsh, and Xiaomo Liu. DocLLM: A layout-aware generative language model for multimodal document understanding. **arXiv [cs.CL]**, December 2023.

[8] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. **arXiv [cs.CV]**, April 2023.

[9] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. **arXiv [cs.CV]**, December 2023.

[10] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, Ji Ma, Jiaqi Wang, Xiaoyi Dong, Hang Yan, Hewei Guo, Conghui He, Botian Shi, Zhenjiang Jin, Chao Xu, Bin Wang, Xingjian Wei, Wei Li, Wenjian Zhang, Bo Zhang, Pinlong Cai, Licheng Wen, Xiangchao Yan, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. How far are we to GPT-4V? closing the gap to commercial multimodal models with open-source suites. **arXiv [cs.CV]**, April 2024.

[11] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. **arXiv [cs.CV]**, September 2024.

[12] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy visual task transfer. **arXiv [cs.CV]**, August 2024.

[13] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. **arXiv [cs.CV]**, January 2023.

[14] Jiabang He, Lei Wang, Yi Hu, Ning Liu, Hui Liu, Xing Xu, and Heng Tao Shen. ICL-D3IE: In-context learning with diverse demonstrations updating for document information extraction. **arXiv [cs.CL]**, March 2023.

[15] Minesh Mathew, Dimosthenis Karatzas, and C V Jawahar. DocVQA: A dataset for VQA on document images. **arXiv [cs.CV]**, July 2020.

[16] Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. Funsd: A dataset for form understanding in noisy scanned documents, 2019.

[17] Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, **Findings of the Association for Computational Linguistics: ACL 2022**, pp. 2263–2279, Dublin, Ireland, May 2022. Association for Computational Linguistics.

A　評価タスクの画像サンプル

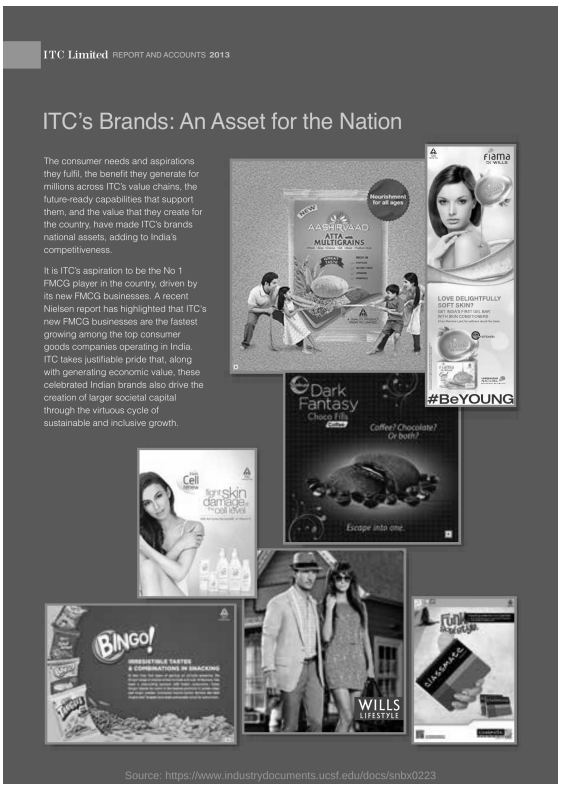

図 2　DocVQA : What is the name of the company? -> itc limited

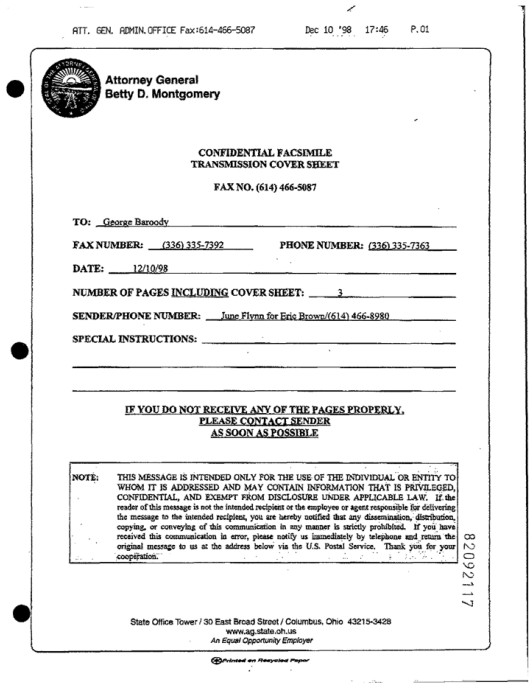

図 3　FUNSD : What is the content in the "DATE:" field? -> 12 /10 /98

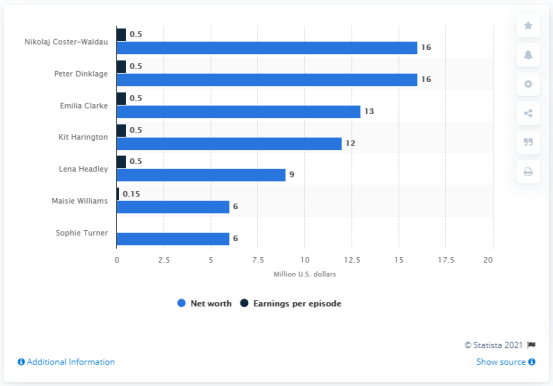

図 4　ChartQA : Who portrayed Jon Snow? -> Kit Harington