Génération augmentée de récupération multi-niveau pour répondre à des questions visuelles

Omar Adjali, Olivier Ferret, Sahar Ghannay, Hervé Le Borgne

Génération augmentée de récupération multi-niveau pour répondre à des questions visuelles

Omar Adjali[1] Olivier Ferret[1] Sahar Ghannay[2] Hervé Le Borgne[1]
(1) Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France
(2) Université Paris-Saclay, CNRS, LISN, 91400, Orsay, France
prenom.nom@lisn.upsaclay.fr, prenom.nom@cea.fr

RÉSUMÉ

La tâche de réponse à des questions visuelles à propos d'entités nommées, qui s'appuie sur la désambiguïsation des entités à l'aide d'informations textuelles et visuelles ainsi que de connaissances, se décompose principalement en deux étapes : recherche d'information puis recherche des réponses, souvent abordées indépendamment l'une de l'autre. La génération augmentée de récupération (RAG) offre une solution à ce manque d'interaction en utilisant les réponses générées comme signal pour l'entraînement de la recherche d'information. Le RAG s'appuie généralement sur des passages pseudo-pertinents extraits de bases de connaissances externes, ce qui peut conduire à des erreurs au niveau de la génération de réponses. Dans ce travail, nous proposons une approche de RAG à plusieurs niveaux améliorant la génération de réponses en associant recherche d'entités et expansion de requête. Plus précisément, nous définissons une fonction de perte RAG permettant de conditionner la génération de réponses à la fois par la recherche d'entités et celle de passages. Cette approche permet de dépasser les travaux existants sur le jeu d'évaluation ViQuAE, démontrant ainsi que les connaissances qu'elle va chercher sont plus pertinentes pour la génération de réponses.

ABSTRACT

Multi-Level Information Retrieval Augmented Generation for Knowledge-based Visual Question Answering.

The Knowledge-Aware Visual Question Answering about Entity task aims to disambiguate entities using textual and visual information, as well as knowledge. It usually relies on two independent steps, information retrieval then reading comprehension, that do not benefit each other. Retrieval Augmented Generation (RAG) offers a solution by using generated answers as feedback for retrieval training. RAG usually relies solely on pseudo-relevant passages retrieved from external knowledge bases which can lead to ineffective answer generation. In this work, we propose a multi-level information RAG approach that enhances answer generation through entity retrieval and query expansion. We formulate a joint-training RAG loss such that answer generation is conditioned on both entity and passage retrievals. We show through experiments new state-of-the-art performance on the ViQuAE KB-VQA benchmark and demonstrate that our approach can help retrieve more actual relevant knowledge to generate accurate answers.

MOTS-CLÉS : questions visuelles, multimodalité, recherche cross-modale, entités nommées.

KEYWORDS: visual question answering, multimodality, cross-modal retrieval, named entities.

ARTICLE : **Publié à** EMNLP 2024 (Adjali *et al.*, 2024). Code disponible à l'adresse suivante.

La réponse à des questions visuelles à propos d'entités nommées (KVQAE) a récemment fait l'objet d'une attention particulière en tant que tâche de référence pour évaluer les capacités des systèmes à comprendre en profondeur les informations visuelles et textuelles afin de répondre à des requêtes multimodales. Alors que dans la tâche standard de question-réponse visuelle (VQA), la réponse aux questions peut se faire uniquement à partir des images, la KQVAE impose la recherche des réponses dans une base de connaissances externes constituée à la fois de textes et d'images.

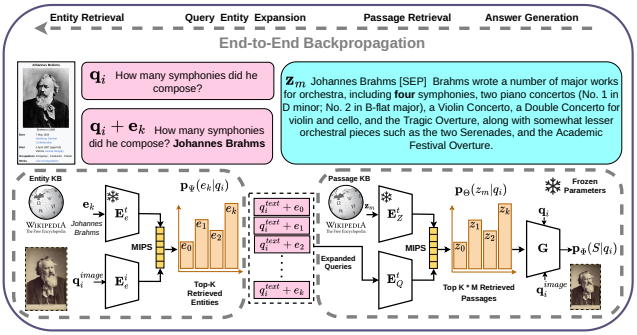

FIGURE 1 – Vue d'ensemble du modèle MiRAG.

Dans cet article, nous proposons le modèle MiRAG, qui s'appuie sur le cadre de la génération augmentée de récupération (*Retrieval Augmented Generation* : RAG (Lewis *et al.*, 2020)) pour résoudre la tâche KVQAE. La figure 1 illustre les principales étapes de notre approche. L'idée de base est d'effectuer la recherche à différents niveaux de granularité (entité et passage) à partir d'une requête multimodale avec des entrées textuelles et visuelles. Cette stratégie d'extraction à plusieurs niveaux affine progressivement la requête avant de générer des réponses à l'aide d'un modèle de génération. La recherche est tout d'abord effectuée à un premier niveau, celui des entités, afin d'identifier un ensemble d'entités candidates pertinentes pour la requête. Les entités ainsi récupérées sont ajoutées à la requête comme une forme d'expansion de requête avant d'effectuer une nouvelle recherche à un niveau plus fin, celui du passage. L'ajout d'entités aux requêtes permet de mieux comprendre le contexte de la requête, ce qui facilite la sélection de passages plus pertinents et fournit un contexte supplémentaire au générateur de réponses. Formellement, notre objectif est d'apprendre la probabilité $p(a|q, z, e)$ de générer une réponse a conditionnée par la requête q, un passage extrait z et l'entité extraite correspondante e. Cet apprentissage est réalisé de bout-en-bout en associant à la fois le modèle de recherche d'information et le modèle de génération des réponses.

	EM	**F1**
PaLM (few-shot) (Chen *et al.*, 2023)	31,5	-
ECA (Lerner *et al.*, 2023)	20,6	24,4
ILF (Lerner *et al.*, 2023)	21,3	25,4
DPR_{V+T} (Lerner *et al.*, 2024)	30,9	34,3
MiRAG (T5-large)	29,8	34,1
MiRAG (BLIP-2)	**36,6**	**41,2**

TABLE 1 – Évaluation de MiRAG et d'un ensemble de baselines sur l'ensemble de test du jeu de données ViQuAE. EM : Exact Match.

Le tableau 1 montre les résultats de l'évaluation du modèle MiRAG sur le jeu de données ViQuAE (Lerner *et al.*, 2022). PaLM est un très gros modèle génératif textuel tandis qu'ECA, ILF et DPR_{V+T} sont des modèles associant texte et image, le dernier étant plus spécifiquement cross-modal. On constate en premier lieu que la version textuelle de MiRAG (T5-large) est assez proche d'un modèle tel que PaLM alors qu'il contient 700 fois moins de paramètres, ce qui valide à la fois l'approche RAG et notre stratégie de recherche à plusieurs niveaux de recherche. Par ailleurs, il surpasse nettement deux des trois modèles de référence multimodaux. Les résultats de la version de MiRAG exploitant conjointement le texte et l'image en s'appuyant sur le modèle BLIP-2 montrent quant à eux l'intérêt de la prise en compte de l'image dans le cadre de MiRAG, avec un fort gain par rapport à la version texte, gain conduisant par ailleurs à dépasser assez notablement les performances de tous les modèles de référence considérés. L'évaluation de MiRAG sur d'autres jeux de données de KVQAE, comme EVQA (Mensink *et al.*, 2023) ou Infoseek (Chen *et al.*, 2023) fait partie des perspectives de ce travail, de même que le fait de pousser plus loin la multimodalité dans la recherche des entités et des passages.

Références

Adjali O., Ferret O., Ghannay S. & Le Borgne H. (2024). Multi-Level Information Retrieval Augmented Generation for Knowledge-based Visual Question Answering. In Y. Al-Onaizan, M. Bansal & Y.-N. Chen, Éds., *2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024)*, p. 16499–16513, Miami, Florida, USA : Association for Computational Linguistics. doi : 10.18653/v1/2024.emnlp-main.922.

Chen Y., Hu H., Luan Y., Sun H., Changpinyo S., Ritter A. & Chang M.-W. (2023). Can Pre-trained Vision and Language Models Answer Visual Information-Seeking Questions ? In H. Bouamor, J. Pino & K. Bali, Éds., *2023 Conference on Empirical Methods in Natural Language Processing*, p. 14948–14968, Singapore : ACL. doi : 10.18653/v1/2023.emnlp-main.925.

Lerner P., Ferret O. & Guinaudeau C. (2023). Multimodal Inverse Cloze Task for Knowledge-based Visual Question Answering. In *45th European Conference on Information Retrieval (ECIR 2024) : Advances in Information Retrieval*, p. 569–587, Dublin, Ireland : Springer Nature Switzerland.

Lerner P., Ferret O. & Guinaudeau C. (2024). Cross-modal Retrieval for Knowledge-based Visual Question Answering. In *46th European Conference on Information Retrieval (ECIR 2024) : Advances in Information Retrieval*, p. 421–438, Glasgow, Scotland : Springer Nature Switzerland.

Lerner P., Ferret O., Guinaudeau C., Le Borgne H., Besançon R., Moreno J. G. & Lovón Melgarejo J. (2022). ViQuAE, a dataset for knowledge-based visual question answering about named entities. In *45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 3108–3120, Madrid, Spain : ACM.

Lewis P., Perez E., Piktus A., Petroni F., Karpukhin V., Goyal N., Küttler H., Lewis M., Yih W.-t., Rocktäschel T., Riedel S. & Kiela D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, volume 33, p. 9459–9474 : Curran Associates, Inc.

Mensink T., Uijlings J., Castrejon L., Goel A., Cadar F., Zhou H., Sha F., Araujo A. & Ferrari V. (2023). Encyclopedic VQA : Visual Questions About Detailed Properties of Fine-Grained Categories. In *IEEE/CVF International Conference on Computer Vision (ICCV 2024)*, p. 3113–3124.