



HAL
open science

Multimodal inverse cloze task for knowledge-based visual question answering

Paul Lerner, Olivier Ferret, Camille Guinaudeau

► **To cite this version:**

Paul Lerner, Olivier Ferret, Camille Guinaudeau. Multimodal inverse cloze task for knowledge-based visual question answering. Journée scientifique du GDR TAL sur l'accès à l'information, Oct 2022, Rennes, France. hal-04379132

HAL Id: hal-04379132

<https://hal.science/hal-04379132v1>

Submitted on 8 Jan 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Multimodal Inverse Cloze Task for Knowledge-based Visual Question Answering

Paul Lerner¹, Olivier Ferret², and Camille Guinaudeau¹

¹ Université Paris-Saclay, CNRS, LISN, 91400, Orsay, France

² Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France
paul.lerner@lisn.upsaclay.fr

We present a new pre-training method, Multimodal Inverse Cloze Task, for Knowledge-based Visual Question Answering about named Entities (KVQAE, [2]). KVQAE is a recently introduced task that consists, as illustrated by Figure 1 in answering questions about named entities grounded in a visual context using a textual Knowledge Base. Therefore, the interaction between the modalities in this task is paramount and must be captured with complex fusion models. As these models require a lot of training data, we design this pre-training task from existing work in textual Question Answering [1]. It consists in considering a sentence as a pseudo-question and its context as a pseudo-relevant passage and is extended by considering images near texts in multimodal documents. Our method is applicable to different neural network architectures and leads to a 9% relative-MRR and 15% relative-F1 gain for retrieval and reading comprehension, respectively, over a no-pre-training baseline.



Visual Question (input)	Relevant visual passage in the Knowledge Base
 <p>“Which constituency did this man represent when he was Prime Minister?”</p>	 <p>“Macmillan indeed lost Stockton in the landslide Labour victory of 1945, but returned to Parliament in the November 1945 by-election in Bromley.”</p>

Fig. 1: Example of a question about a named entity with a visual context along with relevant passages from a Knowledge Base.

Keywords: Visual Question Answering · Named Entities · Pre-training · Multimodal Fusion.

References

1. Lee, K., Chang, M.W., Toutanova, K.: Latent Retrieval for Weakly Supervised Open Domain Question Answering. In: ACL 2019. pp. 6086–6096. Florence, Italy (Jul 2019)
2. Lerner, P., Ferret, O., Guinaudeau, C., Le Borgne, H., Besançon, R., Moreno, J.G., Lovón Melgarejo, J.: ViQuAE, a dataset for knowledge-based visual question answering about named entities. In: SIGIR 2022. New York, NY, USA (2022)