

Multimodal Inverse Cloze Task for Knowledge-based Visual Question Answering (KVQAE)


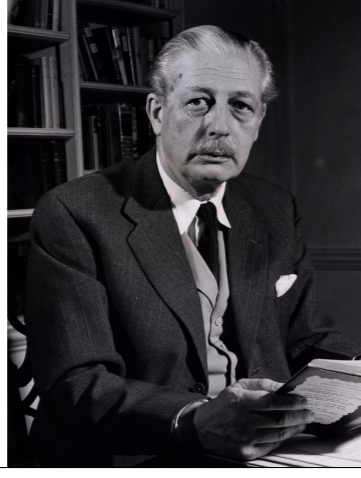


Paul Lerner¹, Olivier Ferret², Camille Guinaudeau¹

¹Universit  Paris-Saclay, CNRS, LISN, 91400, Orsay, France, ²Universit  Paris-Saclay, CEA, List, F-91120, Palaiseau, France

Contact: lerner@lisn.fr

KVQAE and ViQuAE Dataset and Knowledge Base

KVQAE = Answer questions about named entities grounded in a visual context using a Knowledge Base (KB).

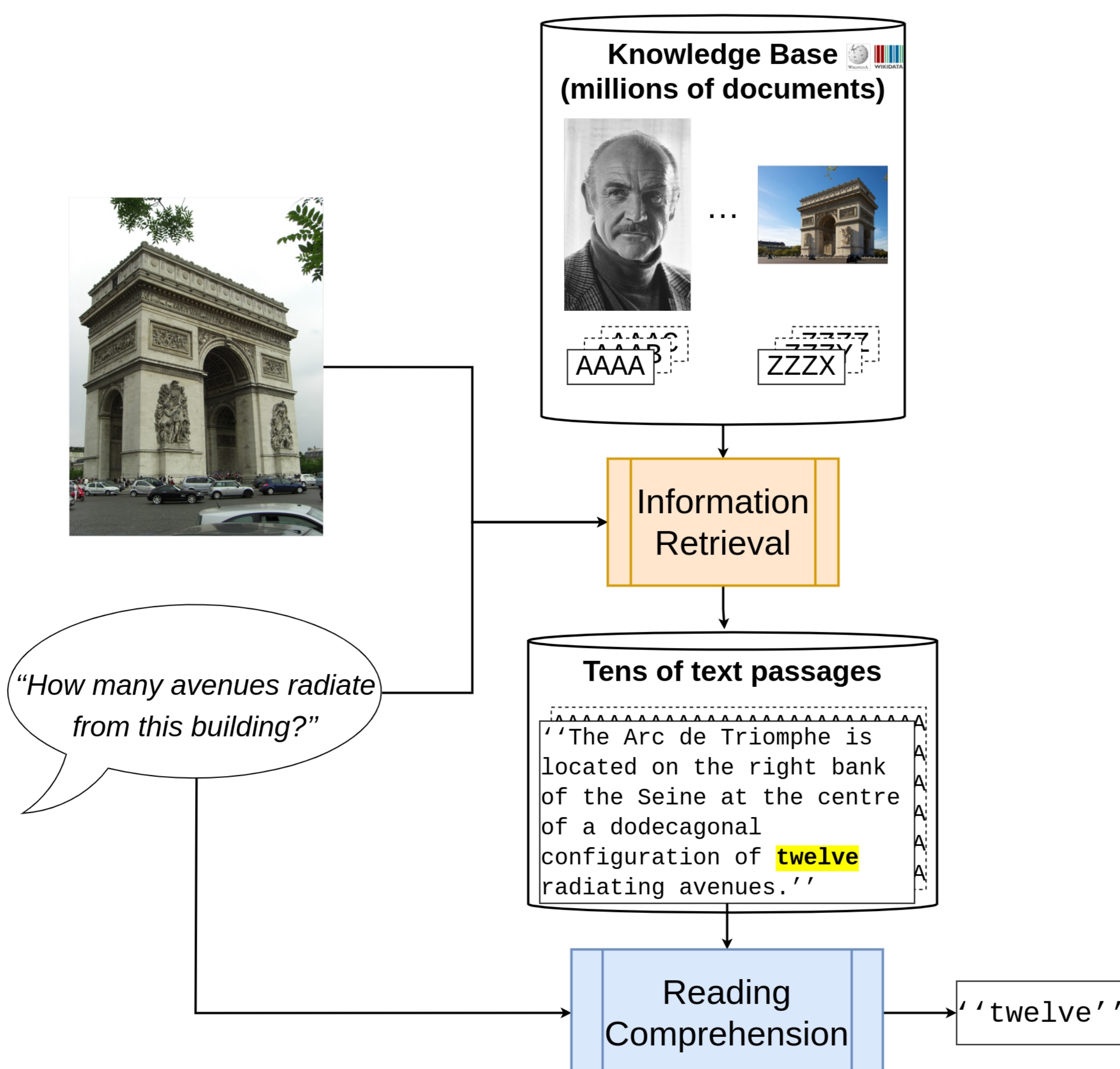
Visual Question (input)	Relevant visual passage in the KB
 <p>“Which constituency did this man represent when he was Prime Minister?”</p>	 <p>“Macmillan indeed lost Stockton in the landslide Labour victory of 1945, but returned to Parliament in the November 1945 by-election in Bromley.”</p>
 <p>“In which year did this ocean liner make her maiden voyage?”</p>	 <p>“Queen Elizabeth 2, often referred to simply as QE2, is a floating hotel and retired ocean liner built for the Cunard Line which was operated by Cunard as both a transatlantic liner and a cruise ship from 1969 to 2008.”</p>

ViQuAE: dataset of 3.7K visual questions (Lerner et al., 2022).

Knowledge Base:

- 1.5M Wikipedia articles with images
- Split in 12M text passages of 100 words

Information Retrieval + Reading Comprehension



Our focus: **Information Retrieval** (efficiently retrieves relevant information from the KB). Requires to *combine text and image*.

Late fusion baseline: results of text and image search fused at the score-level through linear interpolation. Neglects *interaction between the modalities*.

- Text search: **DPR** (Karpukhin et al., 2020): built upon two BERT models (Devlin et al., 2019)
- Image search: ResNet-50 (He et al., 2016) trained in the **CLIP** framework (Radford et al., 2021)

Acknowledgements

This work was supported by the ANR-19-CE23-0028 MEERQAT project (<https://www.meerqat.fr/>). This work was granted access to the HPC resources of IDRIS under the allocation 2021-AD011012846 made by GENCI.

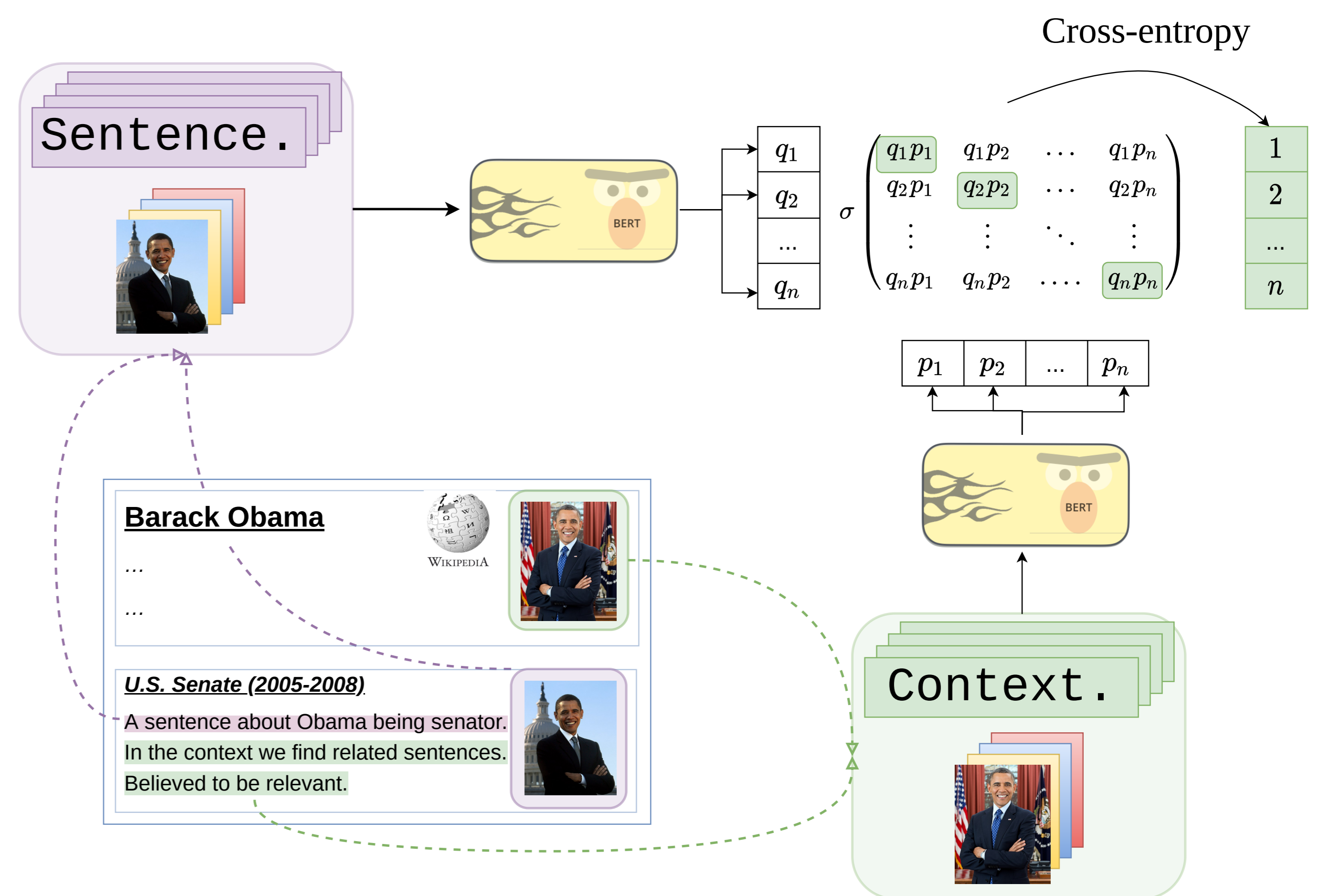
References

Paul Lerner, Olivier Ferret, Camille Guinaudeau, Herv  Le Borgne, Romaric Besan on, Jose G Moreno, and Jes s Lov n Melgarejo. 2022. ViQuAE, a dataset for knowledge-based visual question answering about named entities. In *Proceedings of The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'22*, New York, NY, USA. Association for Computing Machinery.

Multimodal Inverse Cloze Task (ICT)

ViQuAE: small dataset, **early fusion:** complex models

⇒ *pretraining* with Multimodal ICT. Extended from Lee et al. (2019).



Quantitative Results






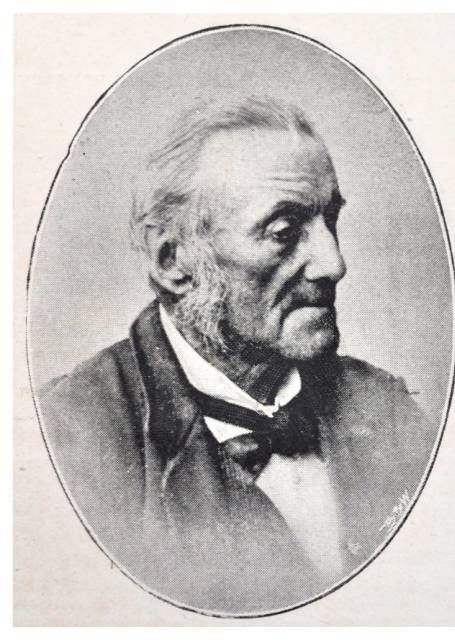
Two metrics for two evaluations:

- Mean Reciprocal Rank (MRR) for Information Retrieval (IR): directly evaluates the relevance of the retrieved visual passages
- F1-score (bag-of-words level) for Reading Comprehension (RC): evaluates the answer extracted by the reader module of Lerner et al.

Model	MRR (IR)	F1 (RC)
DPR (text-only)	32.8	20.1
DPR + CLIP (late fusion)	34.5	22.3
Early Cross-Attention (ECA)	37.8	24.4
Intermediate Linear Fusion (ILF)	37.3	25.4

- Both models outperform late fusion, thanks to cross-modal interactions
- Unexpected: ILF performs on-par with ECA

Qualitative Results

Visual Question	ECA top-1	DPR + CLIP top-1
 <p>“In which English palace was this man born?”</p>	 <p>Blenheim Palace was the birthplace of the 1st Duke's famous descendant, Winston Churchill [...]</p>	 <p>In 1762, George purchased Buckingham House (on the site now occupied by Buckingham Palace) for use as a family retreat. His other residences were Kew and Windsor Castle. St James's Palace was retained for official use.</p>
 <p>“Who designed this cathedral?”</p>	 <p>He was appointed [...] Surveyor of the Fabric of St Paul's Cathedral, where he was responsible for maintaining the building designed by Sir Christopher Wren.</p>	 <p>Sir George Gilbert Scott led the restoration of Salisbury Cathedral between 1863 – 1878. It was during this time that Skidmore created the cathedral's choir screen.</p>

- Unexpected: cross-modal interactions between question image and passage text
- Interactions between question text and image: harder to assess, *not* suggested by quantitative results (ILF ≈ ECA)