

Improving and Diagnosing Knowledge-Based Visual Question Answering via Entity Enhanced Knowledge Injection



The 1st International Workshop on Multimodal Understanding
for the Web and Social Media (MUWS) at WWW 22

Diego Garcia-Olano
Yasumasa Onoe
Joydeep Ghosh

April 26, 2022

Question: How many of them were born in the USA?

Image Caption: *Barack Obama* and his wife *Michelle* at the Civil Rights Summit at the LBJ Presidential Library, 2014.

Wikipedia Entities:

Barack_Obama Michelle_Obama



**Question
+ Image
Caption**

- VQA models are **expensive to pre-train** (many image, question pairs)
Can **we improve upon their performance during fine-tuning?**

- VQA models are **expensive to pre-train** (many image, question pairs)
Can **we improve upon their performance during fine-tuning?**
- Quite a bit of work studying if **LMs can be used as knowledge bases**
But less on **whether vision-language models can be?**

- VQA models are **expensive to pre-train** (many image, question pairs)
Can **we improve upon their performance during fine-tuning?**
- Quite a bit of work studying if **LMs can be used as knowledge bases**
But less on **whether vision-language models can be?**
- Poerner et al 2020 show improved performance on entity-centric text tasks by using a simple, entity based, **knowledge injection technique into LMs.**
Would this injection technique work as well for VQA models?

- VQA models are **expensive to pre-train** (many image, question pairs)
Can **we improve upon their performance during fine-tuning?**
- Quite a bit of work studying if **LMs can be used as knowledge bases**
But less on **whether vision-language models can be?**
- Poerner et al 2020 show improved performance on entity-centric text tasks by using a simple, entity based, **knowledge injection technique into LMs.**
Would this injection technique work as well for VQA models?
- Research on interpretability methods for single modalities is abundant,
How would knowledge injection affect bi-modal explainability?

E-BERT: Efficient-Yet-Effective Entity Embeddings for BERT (Poerner et al ACL 2020)

Wikipedia2Vec (Yamada 2016) $\mathcal{E}_{\text{Wikipedia}} : \mathbb{L}_{\text{Word}} \cup \mathbb{L}_{\text{Ent}} \rightarrow \mathbb{R}^{d_{\text{Wikipedia}}}$

E-BERT: Efficient-Yet-Effective Entity Embeddings for BERT (Poerner et al ACL 2020)

Wikipedia2Vec (Yamada 2016) $\mathcal{E}_{\text{Wikipedia}} : \mathbb{L}_{\text{Word}} \cup \mathbb{L}_{\text{Ent}} \rightarrow \mathbb{R}^{d_{\text{Wikipedia}}}$

E-BERT aligns **Wikipedia2Vec entity embeddings**
to **BERT's wordpiece vector space** for entities found in task text inputs

E-BERT: Efficient-Yet-Effective Entity Embeddings for BERT (Poerner et al ACL 2020)

Wikipedia2Vec (Yamada 2016) $\mathcal{E}_{\text{Wikipedia}} : \mathbb{L}_{\text{Word}} \cup \mathbb{L}_{\text{Ent}} \rightarrow \mathbb{R}^{d_{\text{Wikipedia}}}$

E-BERT aligns **Wikipedia2Vec entity embeddings**
to **BERT's wordpiece vector space** for entities found in task text inputs

Learn map **W** during training

$$\sum_{x \in \mathbb{L}_{\text{WP}} \cap \mathbb{L}_{\text{Word}}} \|\mathbf{W} \mathcal{E}_{\text{Wikipedia}}(x) - \mathcal{E}_{\text{BERT}}(x)\|_2^2$$

Learn map \mathbf{W} during training

$$\sum_{x \in \mathbb{L}_{WP} \cap \mathbb{L}_{Word}} \|\mathbf{W} \mathcal{E}_{\text{Wikipedia}}(x) - \mathcal{E}_{\text{BERT}}(x)\|_2^2$$

At Inference map Wiki ents to BERT via \mathbf{W}

$$\mathcal{E}_{\text{E-BERT}} : \mathbb{L}_{\text{Ent}} \rightarrow \mathbb{R}^{d_{\text{BERT}}}$$

$$\mathcal{E}_{\text{E-BERT}}(a) = \mathbf{W} \mathcal{E}_{\text{Wikipedia}}(a)$$

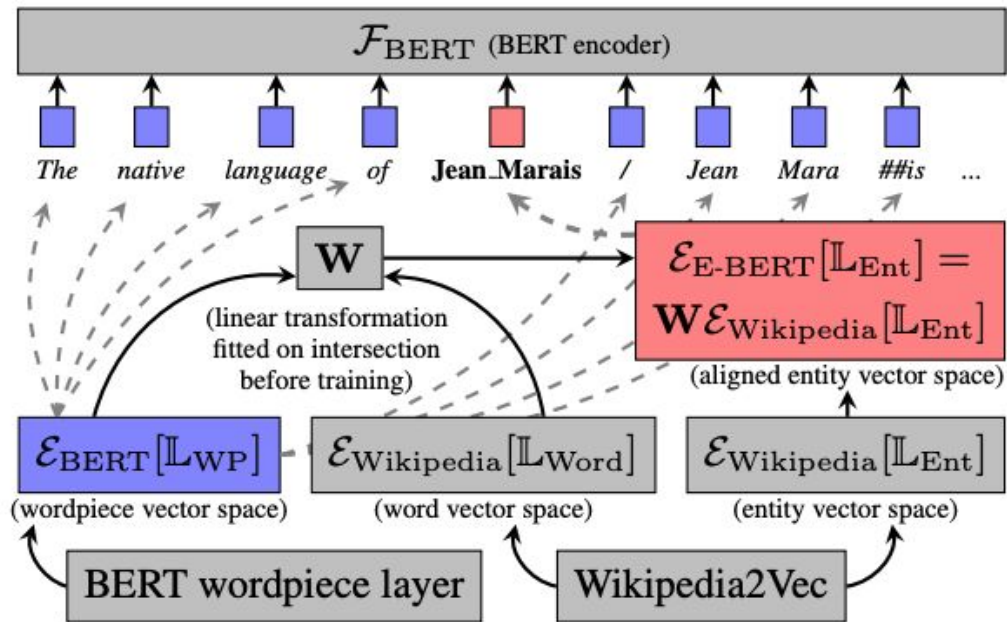


Figure 1: Schematic depiction of E-BERT-concat.

Question: How many of them were born in the USA?

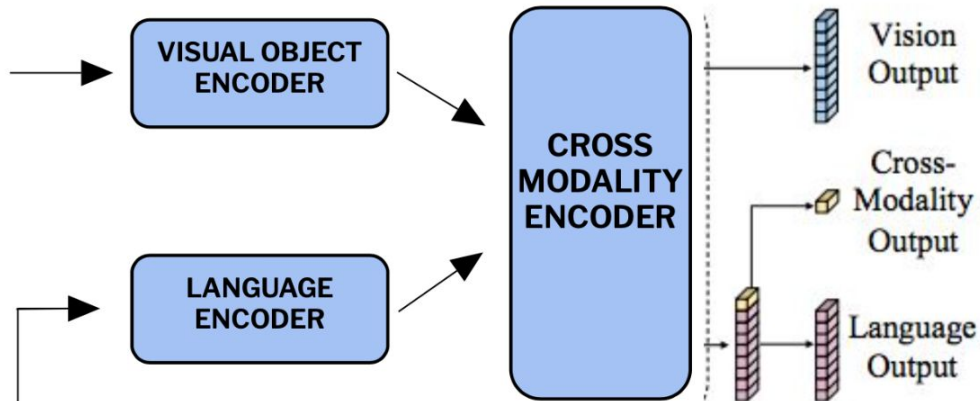
Image Caption: *Barack Obama* and his wife *Michelle* at the Civil Rights Summit at the LBJ Presidential Library, 2014.

Wikipedia Entities:
Barack_Obama Michelle_Obama



Question + Image Caption

LXMERT (Tan et al 2019)



KNOWLEDGE INJECTED INPUT

... in the USA ? *Barack_Obama* / *Barack Obama* and ...

E-BERT concat

(Poerner et al 2020)

$\mathcal{E}_{\text{BERT}}[\mathbb{L}_{\text{WP}}]$
wordpiece vector space

$\mathcal{E}_{\text{E-BERT}}[\mathbb{L}_{\text{Ent}}] =$
 $\mathbf{W}\mathcal{E}_{\text{Wikipedia}}[\mathbb{L}_{\text{Ent}}]$
(aligned entity vector space)

KVQA (Sanket Shah, et al. AAAI 19)

- 24K images with text captions of politicians, actors, athletes, etc
- 183K image/question QA pairs (~ 7 questions per image)
- Metadata for the 18.8K unique Wikipedia entities
- Rare entities (only 65% exist in top million most occurring Wiki entities)

OKVQA (Marino, et al. CVPR 19)

- 14k image/question pairs for commonsense reasoning tasks (fewer entities)
- 10 human generated answers per questions while KVQA only has 1

Entity span construction

KVQA

- 1) Question only (no spans)
- 2) Question + Image Caption (no spans)
- 3) **NERper** - only entities of people
- 4) **NERagro** - all entities, no filtering
- 5) **KVQAmeta** - use metadata provided
(less noise, more precise, only partial cover)

OKVQA

- 1) Question only (no spans)
- 2) **13K** - no filtering to obtain entity spans for 13K QA pairs (92.8% of questions)
- 3) **4K** - semi-automated rules based technique to identify poor candidate spans which filters the set to 4K (28.6% of questions).
- 4) **2.5K** - manual filtering over unique entity spans to filter it down to 2.5K (17.8% of questions).

Table 1: KVQA overall accuracy results over 5 splits and entity spans per question (ents per Q), E-BERT representations injected per question (eberts per Q) and the percent of questions with E-BERT injections (Qs w/ eberts) for split 1

prior work

	Model	Type	Acc	ents per Q	eberts per Q	Qs w/ eberts
	Shah 2019	-	49.50	-	-	-
	+ Caption	-	50.20	-	-	-
1.	Question	-	47.54	-	-	-
2.	+ Caption	-	50.25	-	-	-
3.	NERper	as is	50.37	2.5	1.5	.78
	NERper	links	50.42	1.8	1.5	.79
	NERper	noisy	50.69	2.5	2.3	.94
4.	NERagro	as is	50.26	4.0	2.6	.91
	NERagro	links	50.33	2.2	2.2	.97
	NERagro	noisy	50.77	3.3	3.2	.97
5.	KVQAMeta	as is	52.65	1.4	1.2	.87
	KVQAMeta	links	52.68	1.4	1.3	.95
	KVQAMeta	noisy	52.83	1.4	1.4	.99

- Using E-BERT with entity spans from **KVQAMeta** gives 2.5 points higher accuracy. These spans are the closest to “gold spans” (quality over quantity) however there is still plenty of room for improvement.
- Multi-hop and multi-relationship questions improve by 6 & 5 points respectively (Table 3)
- The improvement for the lower quality derived entity spans (NERper and NERagro) still give .5 accuracy improvement.
- In all cases, more context can be gathered via retrieval mechanisms and E-BERT could be used on top of those results.

Table 2: OKVQA model results over 5 runs. * denotes models based on GPT-3 that are not directly comparable

Model	Mean	Std	Max	Median
OKVQA best	27.84	-	-	-
Shevchenko [29]	39.04	-	-	-
Wu et al [39]	40.50	-	-	-
PICA-Base (best) [41] *	43.3	-	-	-
PICA-Full (best) [41] *	48.0	-	-	-
LXMERT Plain	43.51	0.23	43.87	43.34
+ EBERT 13K	40.59	0.09	40.69	40.59
+ EBERT 4K	43.67	0.13	43.88	43.66
+ EBERT 2.5K	43.61	0.36	44.10	43.34

- Overall using E-BERT on LXMERT for OKVQA has much less effect since the data has very few, as a percentage, questions with entities and image captions from COCO were not used (this is future work)
- Adding noisy entity spans (13K) hurts performance

Table 4: KVQA Bi-modal (BM) and Transformer attention (TRF) explanation results for Questions where an E-BERT injected entity is in top 5 most important tokens.

Model	Type	BM	BM	TRF	TRF
		ACC	Qs	Acc	Qs
NERper	as is	58.25	11.48	56.11	6.13
NERper	links	62.18	8.67	56.28	6.90
NERper	noisy	69.85	4.75	68.17	7.11
NERagro	as is	65.91	4.93	62.41	7.41
NERagro	links	52.74	14.75	49.31	18.52
NERagro	noisy	56.07	20.53	43.31	18.23
KVQAmeta	as is	61.00	2.77	70.03	6.30
KVQAmeta	links	68.97	4.26	79.67	12.57
KVQAmeta	noisy	42.72	5.15	39.65	10.02
Average		59.74	8.59	58.33	10.35

- For 7 of the 9 models, questions which include E-BERT entities amongst their top 5 using BM-GAE provide better accuracy.
- Suggests that when using either method, an entity appearing in the top 5 most important tokens correlates with higher accuracy (59.74 vs 51.04%) *

* Agrees with perturbation test results in Hila Chefer et al ICCV 2021. “Generic Attention-model Explainability for Interpreting Bi-Modal and Encoder-Decoder Transformers.”

BM-GAE EXPLANATION

26253_4,
Q:[CLS] in which continent was the person in
the image born ? civil war photograph of
nelson [SEP]

A:Europe
Qtype:['multi-hop', 'Multi-Relation']

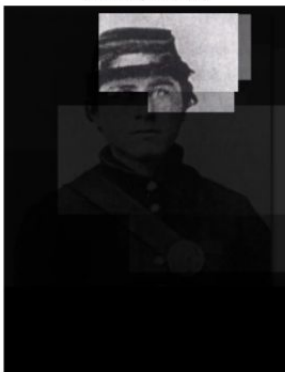
GroundEnts:Knute Nelson



M1

Plain: Europe = 1 1
probs:[0.08, -1.59, -1.92]
Preds:['Europe', 'Asia']
['North America', 'Africa']

Toks:('continent', 1.0)
('was', 0.1789)
('[SEP]', 0.1776)
('in', 0.1662)
('which', 0.1659)



M2

CAPT: North America = 0 1
probs:[1.99, -2.63, -2.89]
Preds:['North America', 'Asia']
['Oceania', 'South America']

Toks:('war', 1.0)
('nelson', 0.8931)
('civil', 0.6977)
('the', 0.0671)
('the', 0.0545)



M3

KVQAmeta Europe = 1 1 
probs:[4.14, -3.58, -4.1]
Preds:['Europe', 'North America']
['Asia', 'Africa']


Toks:('knute', 1.0) 
('<ebert>Knute Nelson</ebert>', 0.795)
('continent', 0.3186)
('/', 0.1777)
('war', 0.1519)
**Ent set:['Knute Nelson']



Figure 2: example of KVQA question where E-BERT is beneficial for KVQAmeta noisy entity set model. The rows show visual and token explanations for BM-GAE over the question/text (left column) and the 5 variants “Question”, “+Caption”, NERagro, NERper and KVQAmeta we explore. Next to each models name is their prediction and whether this top1 prediction is correct (1) or not, and then whether the correct answer exists in the top 5 predictions of the model which are additionally shown along with their logit values. Below that we see the top 5 most important tokens found by the explanation method followed by the set of Entities used for possible knowledge injection

- We analyzed how efficient, entity based knowledge injection via E-BERT during fine tuning affects the performance of an existing model LXMERT on the task of knowledge-based VQA in terms of accuracy & explainability.
- We show substantial improved accuracy on the entity rich KVQA dataset, 2.5% top 1 acc, without the need to redo any costly pre-training.
- Baseline model accuracy is never harmed by knowledge injection on KVQA, & only once for OKVQA, when the entity span set quality is very low.
- This work is complementary to SOTA methods which leverage retrieval based methods to gather additional context to improve VQA task performance since our method can be applied on top of those methods.

Thanks for listening!

Code/data: <https://github.com/diegoolano/kbvqa>

Pre-print: <https://arxiv.org/abs/2112.06888>

www.diegoolano.com

Twitter: @dgolano