Multimodal Differential Network for Visual Question Generation

Badri N. Patro, Sandeep Kumar, Vinod K. Kurmi, and Vinay P. Namboodiri

Indian Institute of Technology Kanpur, India

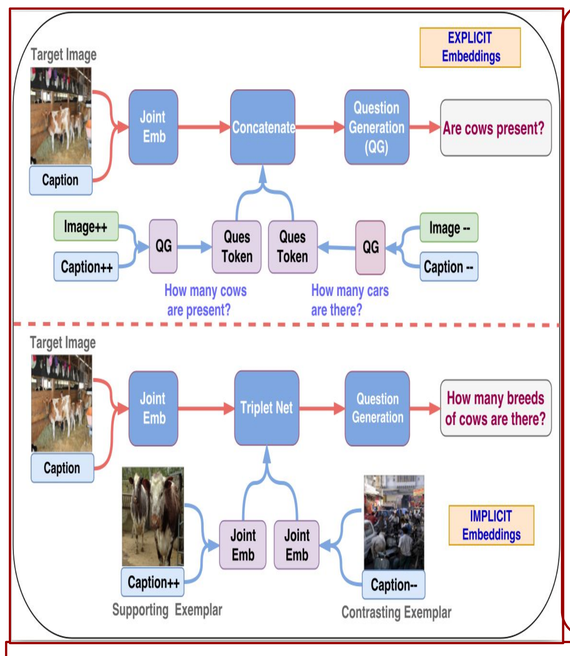
EMNLP 2018

Problem Statement

Goal: To generate natural language questions for given images

Contributions

- Provide a method to incorporate exemplars to learn differential embeddings that captures the subtle differences between supporting and contrasting examples.
- Advocate the use of Triplet Network to bring the target embedding closer to supporting exemplar's embedding and vice-versa.
- Propose Multimodal differential embeddings, as image or text alone doesn't capture the whole context and show that these outperform the ablations which incorporate single cue such as image, or tags or place information.

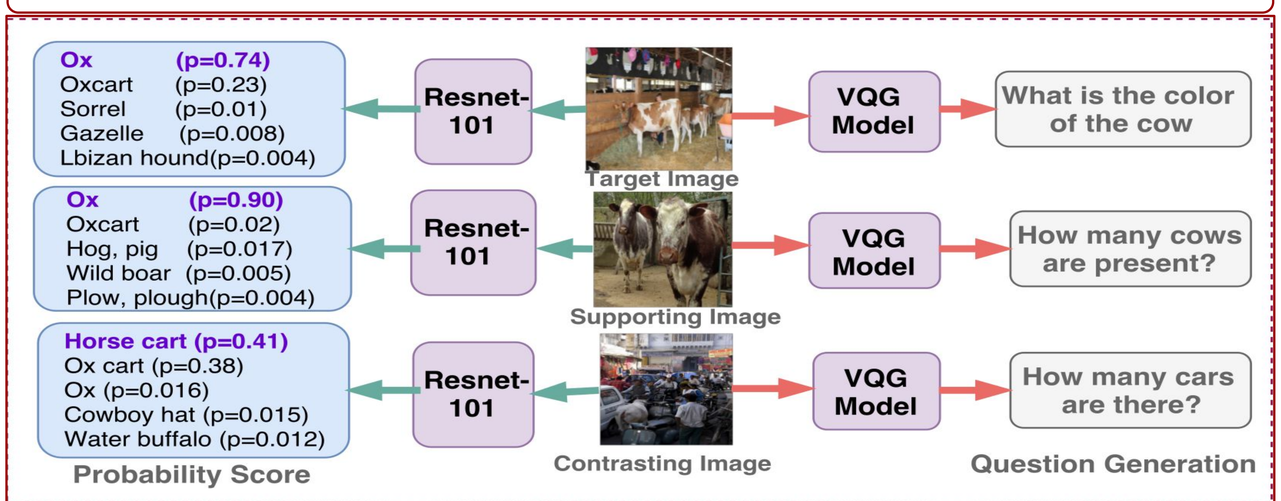


Approaches

- We experiment with different multimodal embeddings and find the image-caption joint embedding to perform the best on question generation.
- Empirical evidence also suggests the use of implicit embeddings over an explicit bag-of-words representation for generating the joint embeddings for exemplars.
- We use a triplet network which ensures that the target multimodal embeddings are closer to the supporting ones and vice-versa.

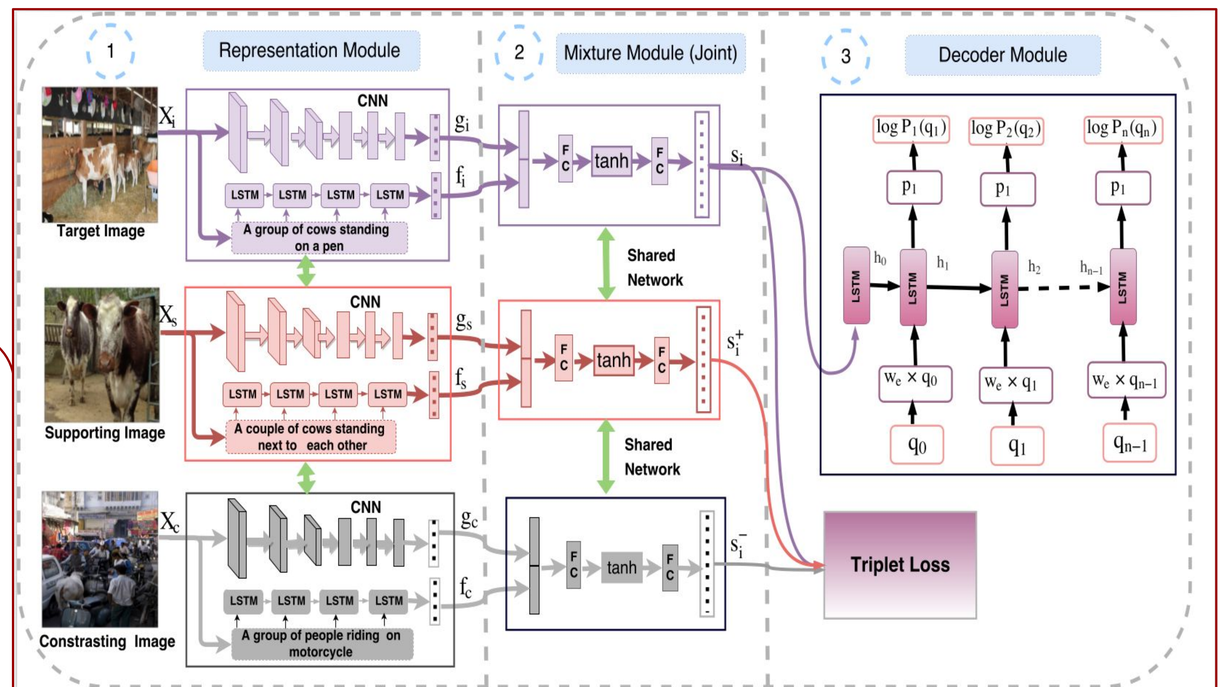
Motivation

- Images consists of multiple visual and language cues like places, captions and tags, but these are not sufficient for question generation in isolation.
- Supporting and Contrasting Exemplars also provide relevant context for Question Generation.



Method

- **Multimodal Differential Network:** We use Representation and Mixture modules to obtain a joint Image-Caption embedding and then a Decoder module to generate a natural language question.
- **Representation Module** consists of a VGG-19 CNN to obtain image representation and LSTM for caption embedding for Target, Supporting and Contrasting exemplars.
- **Mixture Module** takes in the image and caption embeddings and outputs 3 joint embeddings which are fed into a triplet network.
- **Decoder Module** takes the target image-caption embedding and produces a sequence of question words. Our method is trained end to end.

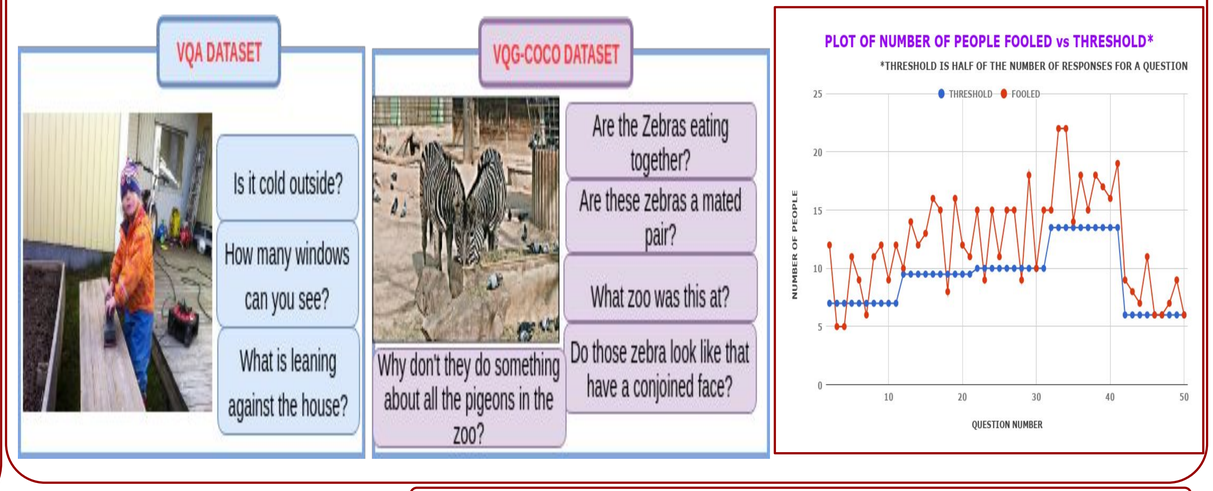


Quantitative Results

Context	BLEU1	METEOR	ROUGE	CIDEr
Natural2016	19.2	19.7	-	-
Creative2017	35.6	19.9	-	-
Image Only	20.8	8.6	22.6	18.8
Caption Only	21.1	8.5	25.9	22.3
Tag-Hadamard	24.4	10.8	24.3	55.0
PlaceCNN-Joint	25.7	10.8	24.5	**56.1**
Diff.Image-Joint	30.4	11.7	26.3	38.8
MDN-Joint (**Ours**)	**36.0**	**23.4**	**41.8**	50.7
Humans2016	**86.0**	**60.8**	-	-

Methods	BLEU1	METEOR	ROUGE	CIDEr
Sample(Yang,2015)	38.8	12.7	34.2	13.3
Max(Yang,2015)	59.4	17.8	49.3	33.1
Image Only	56.6	15.1	40.0	31.0
Caption Only	57.1	15.5	36.6	30.5
MDN-Attention	60.7	16.7	49.8	33.6
MDN-Hadamard	61.7	16.7	50.1	29.3
MDN-Addition	61.7	18.3	50.4	**42.6**
MDN-Joint (**Ours**)	**65.1**	**22.7**	**52.0**	33.1

Qualitative Results



$$L_{cross} = \frac{-1}{N}\sum_{t=1}^{N} y_t \log P(\hat{q}_t | I_i, C_i, \hat{q}_0, .. \hat{q_{t-1}})$$

$$T(s_i, s_i^+, s_i^-) = \max(0, D^+ + \alpha - D^-)$$

$$L = \frac{1}{M}\sum_{i=1}^{M}(L_{cross} + \gamma L_{triplet})$$

$$D(t(s_i), t(s_i^+)) + \alpha < D(t(s_i), t(s_i^-))$$
$$\forall (t(s_i), t(s_i^+), t(s_i^-)) \in M,$$

Email Id:{badri,sandepkr,vinodkk}@iitk.ac.in,
Project page: https://badripatro.github.io/MDN-VQG/
Acknowledgement: Special thanks to Microsoft India for travel support.