UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

MVA2023

Word to Sentence Visual Semantic Similarity for Caption Generation: Lessons Learned

Ahmed Sabir

TALP Research Center, Universitat Politècnica de Catalunya, Barcelona, Spain

asabir@cs.upc.edu

Background

- Although SoTA models generate human-like captions, they are known to lack lexical diversity because they do not possess a semantic understanding of the relation between objects in an image.
- We propose a post-process visual re-ranker that intends to **visually ground** the most relevant candidate caption to its related visual context in the image via semantic understanding.

Visual context: food
BL_{BeamS}: a plate of food on a table
$VR_{BERT+GloVe}$: a plate of food **and a drink** on a table
Human: a white plate with some food on it.

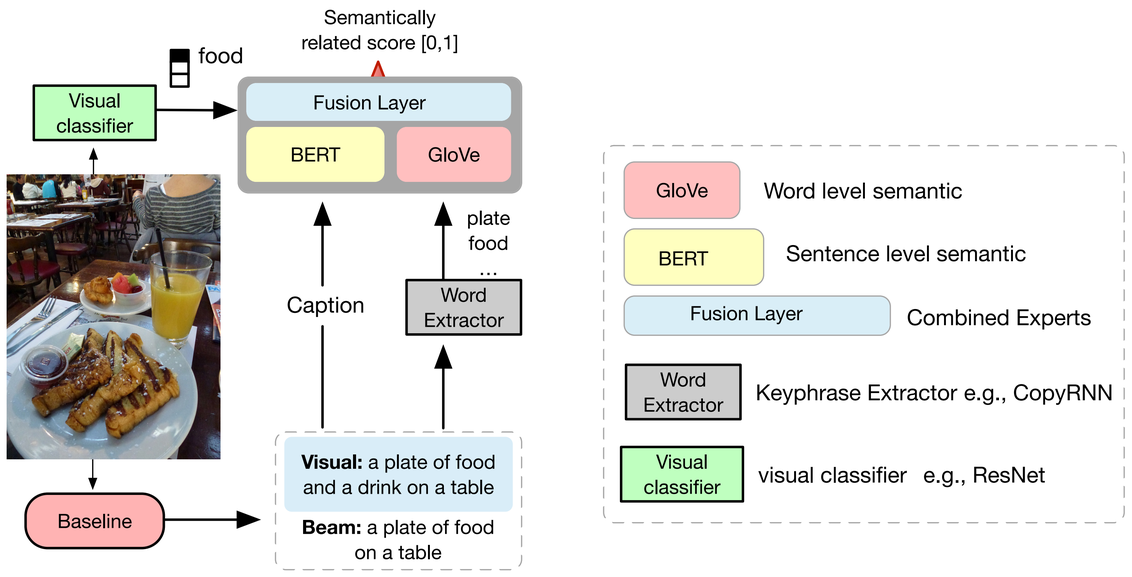
Visual context: chainlink fence
BL_{BeamS}: a black and white photo of train tracks
$VR_{BERT+GloVe}$: a black and white photo of a train **on the tracks**
Human: long train sitting on a railroad track

Visual context: bassinet
BL_{BeamS}: a baby sitting in front of a cake
$VR_{BERT+GloVe}$: a baby sitting in front of **a birthday cake**
Human: a woman standing over a sheet cake sitting on top of table

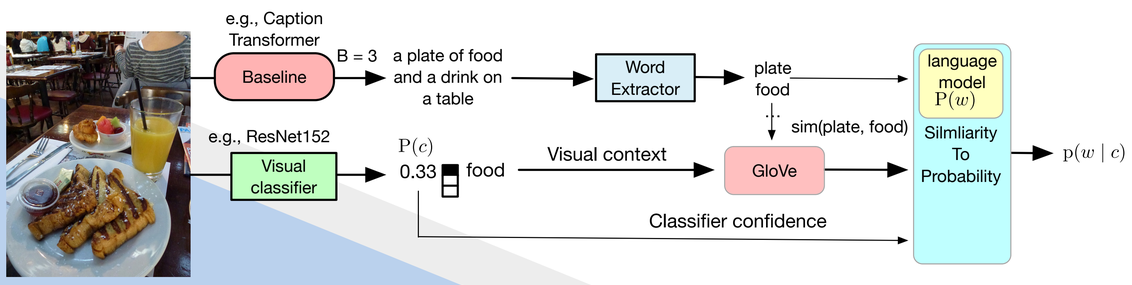
Visual context: trolleybus
BL_{Greedy}: a green bus parked in front of a building
$VR_{BERT+GloVe}$: a green double decker bus parked in front of a building ✗
Human: a passenger bus that is parked in front of a library

Architecture

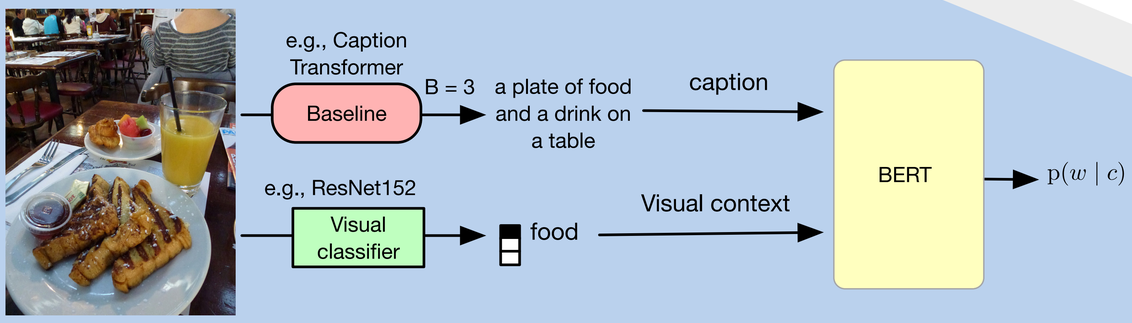
We introduce semantic relations between the visual context in the image and the caption at the word and sentence levels. We propose a joint BERT [9] with GloVe [28] to capture visual semantic similarity. The main components of the proposed Architecture:



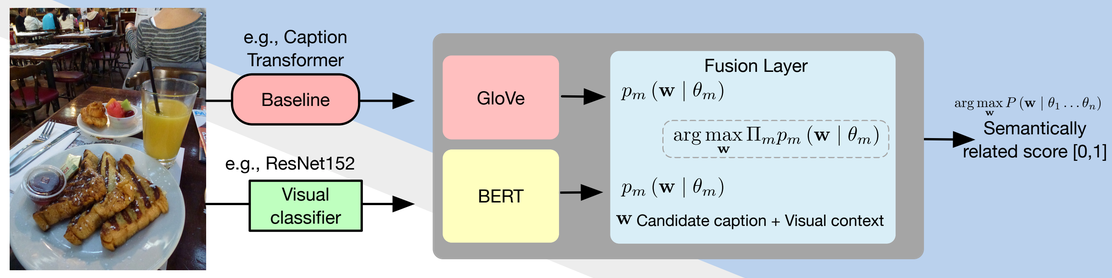
- **Word Level Model:** To enable word-level semantics with GloVe, we extract keyphrases [24] from the caption, and we then employ the confidence of the classifier in the image to convert the similarity into a probability [30].



- **Sentence Level Model:** We fine-tuned BERT on the Caption dataset, incorporating the top-k 3 visual context information extracted from each image [11], where the target is the **semantic relatedness** between the visual and the candidate caption. Also, we employ RoBERTa-sts [29] as out-of-the-box model.



- **Fusion Layers:** Inspired by Products of Experts[12], we merged the two experts through a Late Fusion layer. As this work aims to retrieve the closest candidate caption with the highest probability, the normalization step is unnecessary. The output is the combined probability of caption + visual (**w**).



Experiments

Dataset and Visual Context

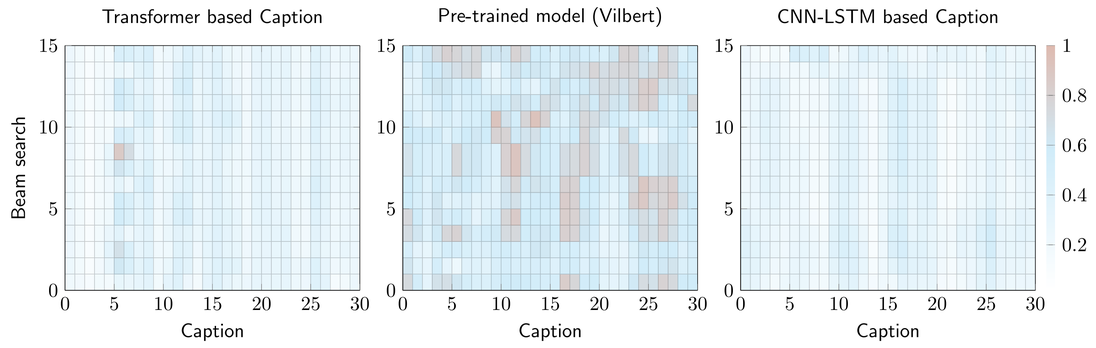
We evaluate the proposed approach on two different size datasets. The idea is to evaluate our approach with (1) a shallow model CNN-LSTM (*i.e.* less data scenario), and on a system that is trained on a huge amount of data (*e.g.* VilBERT and Transformer).

- **Caption Dataset:** For Training, we use the five human annotated captions from the COCO-Caption ♣ [18] and Flickr8k ♠ [13] datasets. For Testing, for the Transformer baselines: VilBERT and Caption Transformer: the 5k Karpathy test split, and for the CNN-LSTM baseline the Flickr8k test set 1730.
- **Visual Context:** We enrich the two datasets, as mentioned above, with textual visual context information using Object classifier ResNet-152 1000 classes.

Results

Model	B-1	B-2	B -3	B-4	M	R	C	BERTscore
Show and Tell (CNN-LSTM) [32] ♠								
$Tell_{BeamS}$	**0.331**	**0.159**	0.071	0.035	0.093	0.270	0.035	**0.8871**
$Tell+VR^{\cdot}V1_{BERT\text{-}GloVe}$	0.330	0.158	0.069	0.035	0.095	0.273	0.036	0.8855
$Tell+VR^{\cdot}V2_{BERT\text{-}GloVe}$	0.320	0.154	**0.073**	**0.037**	0.099	**0.277**	**0.041**	0.8850
$Tell+VR^{\cdot}V1_{RoBERTa\text{-}GloVe}$ (sts)	0.313	0.153	0.072	**0.037**	**0.101**	0.273	0.036	0.8839
VilBERT [21] ♣								
Vil_{BeamS}	0.739	0.577	0.440	0.336	0.271	0.543	1.027	0.9363
$Vil+VR^{\cdot}V1_{BERT\text{-}GloVe}$	0.739	0.576	0.438	0.334	**0.273**	0.544	1.034	0.9365
$Vil+VR^{\cdot}V2_{BERT\text{-}GloVe}$	**0.740**	0.578	0.439	0.334	**0.273**	**0.545**	1.034	0.9365
$Vil+VR^{\cdot}V2_{RoBERTa\text{-}GloVe}$ (sts)	**0.740**	**0.579**	**0.442**	**0.338**	0.272	**0.545**	**1.040**	**0.9366**
Transformer based Caption Generator [8] ♣								
$Trans_{BeamS}$	**0.780**	**0.631**	**0.491**	**0.374**	**0.278**	**0.569**	**1.153**	**0.9399**
$Trans+VR^{\cdot}V1_{BERT\text{-}GloVe}$	**0.780**	0.629	0.487	0.371	**0.278**	0.567	1.149	0.9398
$Trans+VR^{\cdot}V2_{BERT\text{-}GloVe}$	**0.780**	0.630	0.488	0.371	**0.278**	0.568	1.150	**0.9399**

Through these heatmap probabilities change after visual re-ranking, we can observe the advantages of incorporating visual re-ranking *e.g.* VilBERT.



Ablation study

We preformed an ablation study to investigate effectiveness of each expert, and why the negative result, by evaluating each model as stand-alone.

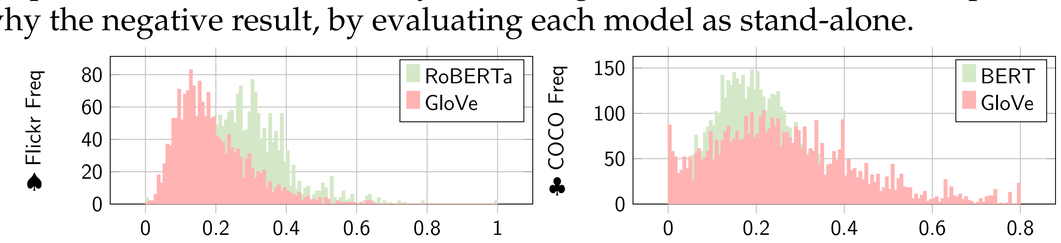


Figure ♠ (Left) Each Expert is contributing different probability confidence and therefore the model is learning the semantic relation. Figure (Right) ♣ shows that BERT is not contributing, as GloVe is dominating to become the expert, to the final score for two reasons: (1) short caption, and (2) less diverse beam.

Limitation

- **Word Model Similarity:** the fluctuating of independent stand-alone word similarity score *i.e.* extracted keywords from the caption with the visual.
- **Object detectors:** misclassified and hallucinated objects, which results in an inaccurate semantic score.