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

Ahmed Sabir

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

MVA 2023

Introduction

Although SoTA models generate human-like captions, they are known to lack lexical diversity due to the absence of the **semantic understanding** of the relation between objects in the image.

BL_{BeamS}: a plate of food on a table

Human: a white plate with some food on it.

BL_{BeamS}: a black and white photo of train tracks

Human: long train sitting on a railroad track.

BL_{BeamS}: a baby sitting in front of a cake

Human: a woman standing over a sheet cake sitting on top of table.

$\text{BL}_{\text{Greedy}}$: a green bus parked in front of a building

Human: a passenger bus that is parked in front of a library.

Contribution

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
$\mathbf{BL}_{\mathbf{BeamS}}$: a plate of food on a table
$\mathbf{VR}_{\mathbf{BERT+GloVe}}$: a plate of food **and a drink** on a table
Human: a white plate with some food on it.

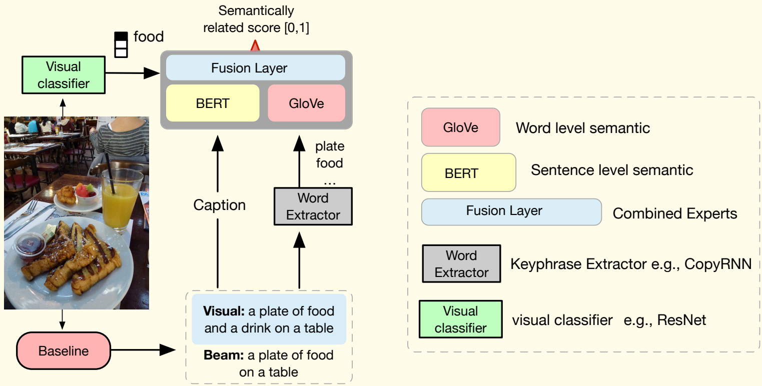
Visual context: chainlink fence
$\mathbf{BL}_{\mathbf{BeamS}}$: a black and white photo of train tracks
$\mathbf{VR}_{\mathbf{BERT+GloVe}}$: a black and white photo of a train **on the tracks**
Human: long train sitting on a railroad track

Visual context: bassinet
$\mathbf{BL}_{\mathbf{BeamS}}$: a baby sitting in front of a cake
$\mathbf{VR}_{\mathbf{BERT+GloVe}}$: 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
$\mathbf{BL}_{\mathbf{Greedy}}$: a green bus parked in front of a building
$\mathbf{VR}_{\mathbf{BERT+GloVe}}$: a green double decker bus parked in front of a building ✗
Human: a passenger bus that is parked in front of a library

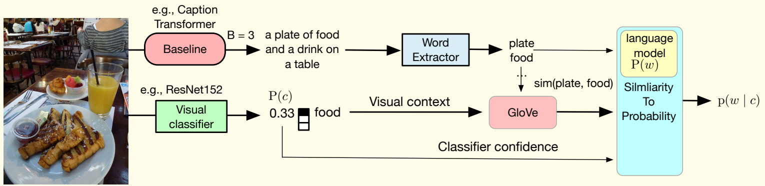
Architecture Overview

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.



Word-level Model

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

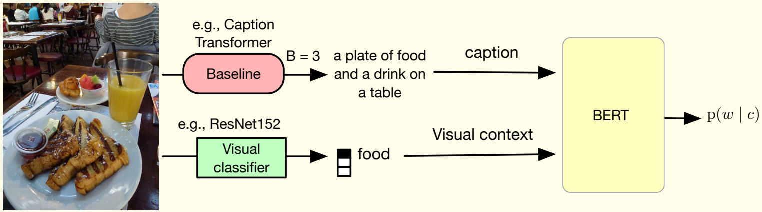


[24] Rui *et al.* Deep Keyphrase Generation. ACL2017

[30] Sabir *et al.* Visual Re-rankeing with Natural Language Understanding for Text Spotting. ACCV2018

Sentence-level Model

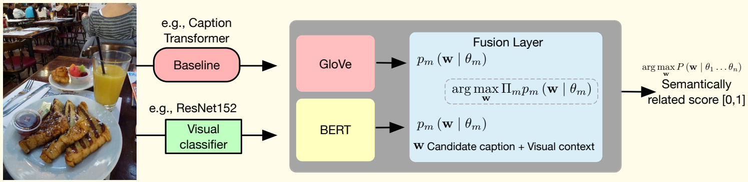
We fine-tuned BERT on the Caption dataset, incorporating the top-k 3 visual context information extracted from each image[11], where target is the **semantic relatedness** between the visual and the candidate caption.



[11] He *et al.* Deep residual learning for image recognition. CVPR2016

Fusion layer

Inspired by Products of Experts[12], we merged the two experts through a Fusion layer. As this work aims to retrieve the closest candidate caption with the highest probability, the normalization step is unnecessary.



[12] Hinton *et al.* Products of experts. ICANN1999

Results

We experiment with two datasets and three models (CNN-LSTM), Vision-and-Language BERT (ViLBERT) and Caption Transformer.

Model	B-1	B-2	B -3	B-4	M	R	C	BERTscore
Show and tell (CNN-LSTM) [32] ♠								
$\text{Tell}_{\text{BeamS}}$	**0.331**	**0.159**	0.071	0.035	0.093	0.270	0.035	**0.8871**
$\text{Tell+VR}_V1_{\text{BERT-Glove}}$	0.330	0.158	0.069	0.035	0.095	0.273	0.036	0.8855
$\text{Tell+VR}_V2_{\text{BERT-Glove}}$	0.320	0.154	**0.073**	**0.037**	0.099	**0.277**	**0.041**	0.8850
$\text{Tell+VR}_V1_{\text{RoBERTa-Glove}}$ (sts)	0.313	0.153	0.072	**0.037**	**0.101**	0.273	0.036	0.8839
ViLBERT [21] ♣								
$\text{Vil}_{\text{BeamS}}$	0.739	0.577	0.440	0.336	0.271	0.543	1.027	0.9363
$\text{Vil+VR}_V1_{\text{BERT-Glove}}$	0.739	0.576	0.438	0.334	**0.273**	0.544	1.034	0.9365
$\text{Vil+VR}_V2_{\text{BERT-Glove}}$	**0.740**	0.578	0.439	0.334	**0.273**	**0.545**	1.034	0.9365
$\text{Vil+VR}_V2_{\text{RoBERTa-Glove}}$ (sts)	**0.740**	**0.579**	**0.442**	**0.338**	0.272	**0.545**	**1.040**	**0.9366**
Transformer based caption generator [8] ♣								
$\text{Trans}_{\text{BeamS}}$	**0.780**	**0.631**	**0.491**	**0.374**	**0.278**	**0.569**	**1.153**	**0.9399**
$\text{Trans+VR}_V1_{\text{BERT-Glove}}$	**0.780**	0.629	0.487	0.371	**0.278**	0.567	1.149	0.9398
$\text{Trans+VR}_V2_{\text{BERT-Glove}}$	**0.780**	0.630	0.488	0.371	**0.278**	0.568	1.150	**0.9399**

♠Flickr8K dataset: Micah*et al.* Framing image description as a ranking task. JAIR 2013

♣COCO-Caption dataset: Lin*et al.* Microsoft coco: Common objects in context. ECCV2014

Results

Our re-ranker yielded mixed result (+) improving model accuracy, (−) struggles when dealing with less diverse captions *e.g*. Transformer baseline.

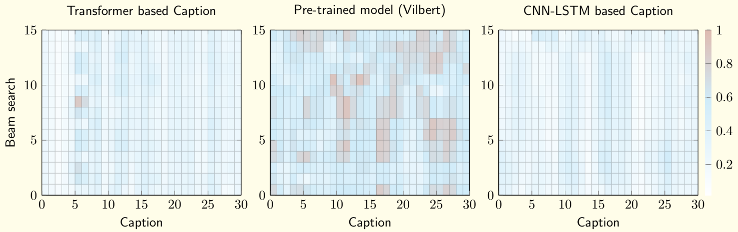
Model	B-1	B-2	B -3	B-4	M	R	C	BERTscore
Show and tell (CNN-LSTM) [32] ♠								
Tell$_{\text{BeamS}}$	**0.331**	**0.159**	0.071	0.035	0.093	0.270	0.035	**0.8871**
Tell+VR$_V1_{\text{BERT-Glove}}$	0.330	0.158	0.069	0.035	0.095	0.273	0.036	0.8855
Tell+VR$_V2_{\text{BERT-Glove}}$	0.320	0.154	**0.073**	**0.037**	0.099	**0.277**	**0.041**	0.8850
Tell+VR$_V1_{\text{RoBERTa-Glove}}$ (sts)	0.313	0.153	0.072	**0.037**	**0.101**	0.273	0.036	0.8839
VilBERT [21] ♣								
Vil$_{\text{BeamS}}$	0.739	0.577	0.440	0.336	0.271	0.543	1.027	0.9363
Vil+VR$_V1_{\text{BERT-Glove}}$	0.739	0.576	0.438	0.334	**0.273**	0.544	1.034	0.9365
Vil+VR$_V2_{\text{BERT-Glove}}$	**0.740**	0.578	0.439	0.334	**0.273**	**0.545**	1.034	0.9365
Vil+VR$_V2_{\text{RoBERTa-Glove}}$ (sts)	**0.740**	**0.579**	**0.442**	**0.338**	0.272	**0.545**	**1.040**	**0.9366**
Transformer based caption generator [8] ♣								
Trans$_{\text{BeamS}}$	**0.780**	**0.631**	**0.491**	**0.374**	**0.278**	**0.569**	**1.153**	**0.9399**
Trans+VR$_V1_{\text{BERT-Glove}}$	**0.780**	0.629	0.487	0.371	**0.278**	0.567	1.149	0.9398
Trans+VR$_V2_{\text{BERT-Glove}}$	**0.780**	0.630	0.488	0.371	**0.278**	0.568	1.150	**0.9399**

♠ Flickr8K dataset: Micah *et al*. Framing image description as a ranking task. JAIR 2013

♣ COCO-Caption dataset: Lin *et al*. Microsoft coco: Common objects in context. ECCV2014

Results

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



Results

Our re-ranker improve the lexical diversity, each selected caption has more Vocabulary, Unique words/total Words Per Caption.

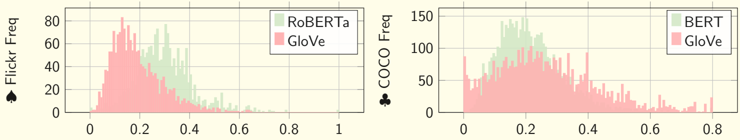
Model	Voc	TTR	Uniq	WPC
Show and Tell [32] ♠				
$\text{Tell}_{\text{BeamS}}$	304	0.79	**10.4**	12.7
$\text{Tell+VR}_{\text{RoBERTa-Glove}}$	**310**	**0.82**	9.42	**13.5**
VilBERT [21] ♣				
$\text{Vil}_{\text{BeamS}}$	894	**0.87**	8.05	10.5
$\text{Vil+VR}_{\text{RoBERTa-Glove}}$	**953**	0.85	**8.86**	**10.8**
Transformer [8] ♣				
$\text{Trans}_{\text{BeamS}}$	935	0.86	7.44	**9.62**
$\text{Trans+VR}_{\text{BERT-Glove}}$	**936**	0.86	**7.48**	8.68

♠Flickr8K dataset: Micah*et al.* Framing image description as a ranking task. JAIR 2013

♣COCO-Caption dataset: Lin*et al.* Microsoft coco: Common objects in context. ECCV2014

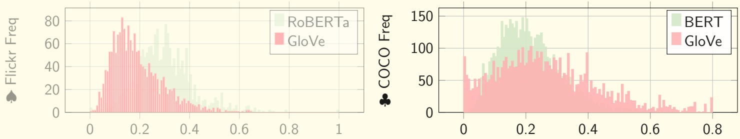
Ablation study

We performed an ablation study to investigate the effectiveness of each expert, by evaluating each model as stand-alone.



Ablation study

With our worst model (BERT-GloVe), with less diverse caption (*i.e.* less sentence context), word-level GloVe dominates as the main expert.



Thank You