Belief Revision based Caption Re-ranker with Visual Semantic

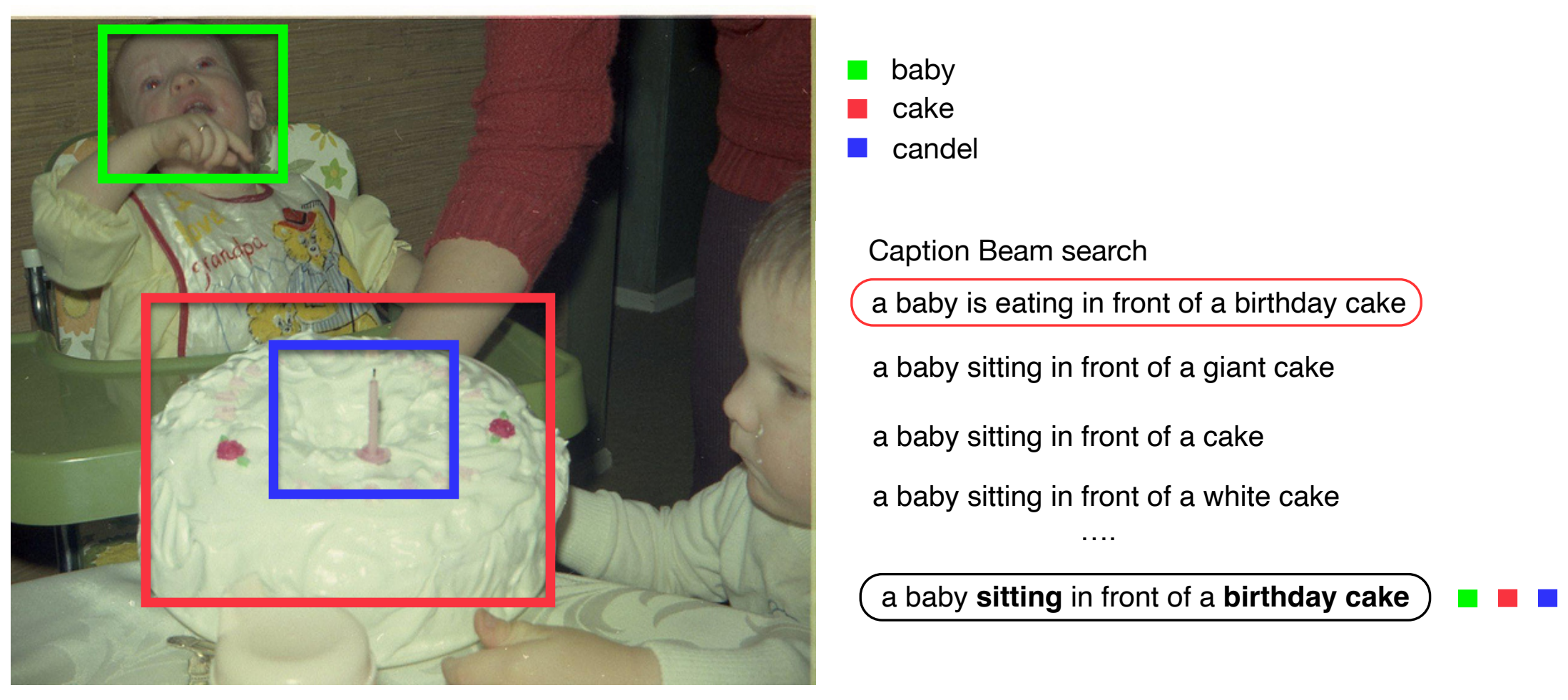
Ahmed Sabir[1], Francesc Moreno-Noguer[2], Pranava Madhyastha[3] and Lluís Padró[1]
[1]TALP Research Center, Universitat Politècnica de Catalunya, Barcelona, Spain
[2]Institut de Robòtica i Informàtica Industrial (CSIC-UPC), Barcelona, Spain
[3]City, University of London, UK
asabir@cs.upc.edu, fmoreno@iri.upc.edu, pranava@city.ac.uk, padro@cs.upc.edu

Background

- While SoTA models generate captions that are comparable to humans. They are known to lack lexical diversity. One of the limitations is that the **narrow beam search** may not result in the most description caption of the image.

  Also they are lack **semantic understanding** of the relation between objects in the image.
- Recent works use a beam search directly to produce diverse captions by forcing richer lexical **word choices** (Ippolito et al., 2019; Vijayakumar et al., 2018; Wang and Chan, 2019; Wang et al., 2020).

  However, these methods do not guarantee to include all objects in the image that are semantically related, which results in an incorrect diverse caption.
- We propose a post-process based **Visual Beam Re-ranker** (VR) that intends to visually ground the most closely related candidate beam to its related visual context. Our approach enhances the performance of any typical image captioning system without the necessity for additional training

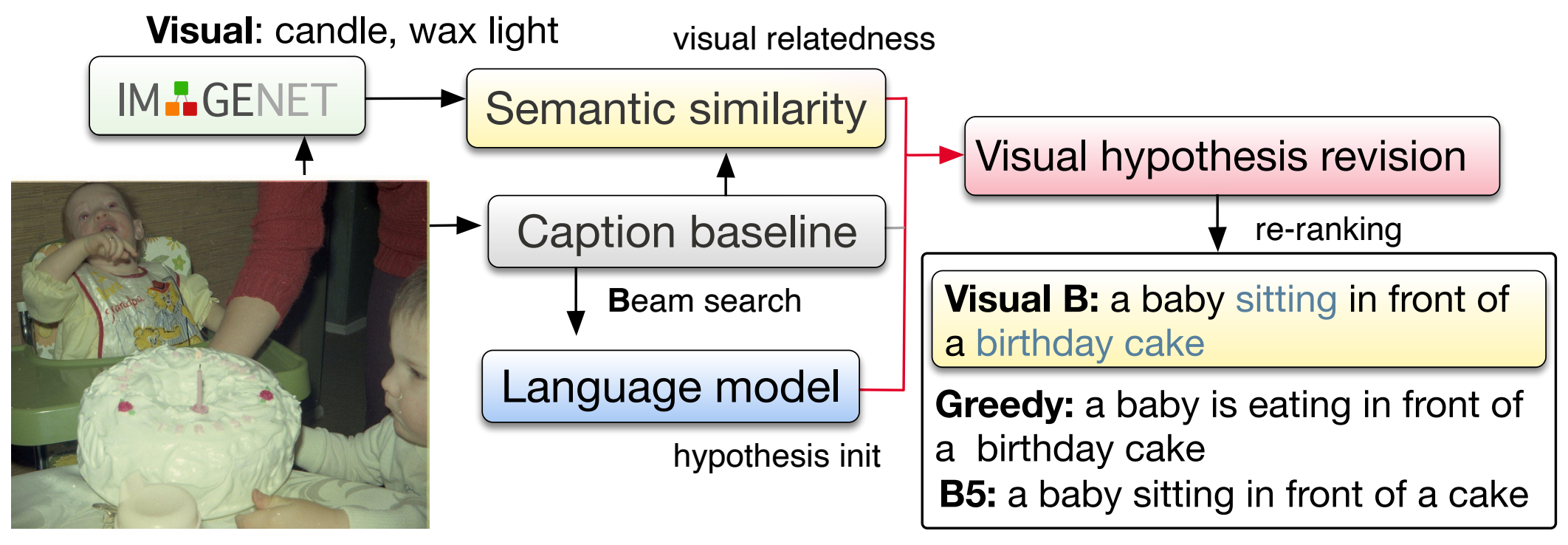

🟩 baby
🟥 cake
🟦 candel
Caption Beam search
a baby is eating in front of a birthday cake
a baby sitting in front of a giant cake
a baby sitting in front of a cake
a baby sitting in front of a white cake
....
a baby **sitting** in front of a **birthday cake** 🟩 🟥 🟦

Belief Revision Score

The Visual Re-ranker is based on **Probability from Similarity** (Blok et al., 2003). `SimProb` is a concept based on **belief revision** framework. Belief revision is a process of formatting a belief by **bring into account a new piece of information.**

Model Architecture: The main components of visual hypothesis revision VR:

- Language Model (Autoregressive LM *e.g.* GPT-2 (Radford et al., 2019))
- Visual Concept (Visual Classifier *e.g.* ResNet (He et al., 2016))
- Similarity (Mask Language Model *e.g.* BERT (Devlin et al., 2019))


Visual: candle, wax light
visual relatedness
IMAGENET
Semantic similarity
Visual hypothesis revision
Caption baseline
re-ranking
Beam search
Language model
hypothesis init
Visual B: a baby sitting in front of a birthday cake
Greedy: a baby is eating in front of a birthday cake
B5: a baby sitting in front of a cake

Philosophical Intuitions of `SimProb` Belief Revision: Let us consider the following statements:

obs 1 Tigers **can bit through wire**, therefore Jaguars **can bit through wire**.

obs 2 Kitten **can bit through wire**, therefore **Jaguars can bit through wire**.

obs 1 seem logical because it match the expectation. This obs 1 is consistent with our previous believe (Tigers are similar to Jaguars in terms of strength), and **no need to revise it.**

obs 2 is surprising because our prior belief is that kittens are not so strong, then we need to **revise and update our prior belief about kitten strength**.

The `SimProb` Model as VR can be written as:

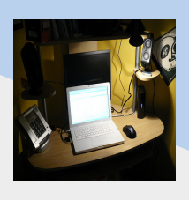
$$P(w|context = visualconcept) = P(w)^{\alpha}$$

where:

- $\alpha = \left(\frac{1-sim(w,c)}{1+sim(w,c)}\right)^{1-P(c)}$
- **Hypothesis** $P(w)$: The prior probabilities of original belief. As this approach is inspired by humans, the hypothesis P(w) needs to be initialized by a common observation such as a Language Model trained on a general text corpus.
- **Informativeness** $P(c)$: The information that causes hypothesis revision. We leverage ResNet (He et al., 2016) and an Inception-ResNet v2 based Faster R-CNN object detector (Huang et al., 2017) to extract textual visual context information from the image.
- **Similarities** $sim(w,c)$: Hypothesis revision is more likely if there is a close relation between the hypothesis and new information. We employ BERT to compute the similarity between the hypothesis (caption) and its related visual context.

Code: `https://github.com/ahmedssabir/Belief-Revision-Score`

Example

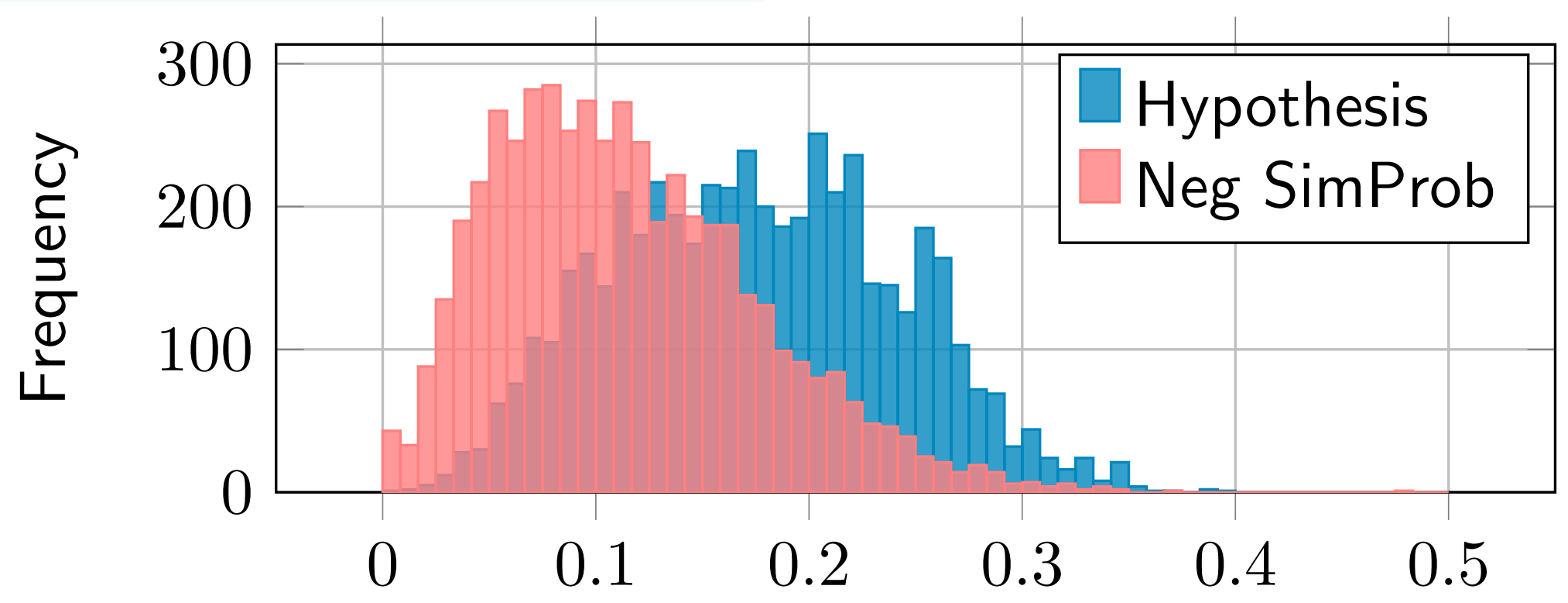
Model	Caption	BERTscore	SBERT-sts	Human%	Visual
BeamS	a close up of a plate of food	0.89	0.27	40	trifle
VR	piece of food sitting on top of a white plate	**0.91**	**0.53**	**60**	
Human refe	a white plate and a piece of white cake				
BeamS	a group of men on a field playing baseball	0.88	0.58	33.3	baseball
VR	a batter catcher and umpire during a baseball game	**0.91**	**0.84**	**66.7**	
Human refe	batter catcher and umpire anticipating the next pitch				
BeamS	a laptop computer sitting on top of a desk	0.91	0.69	25	desk
VR	a desk with a laptop and computer monitor	**0.95**	**0.77**	**75**	
Human refe	an office desk with a laptop and computer monitor				

Belief Revision Score with Negative Evidence

The Negative Evidence refers to the cases where the absence of visual evidence (¬c) leads to a decrease in the probability of the hypothesis

$$P(w \mid \neg c) = 1 - (1 - P(w))^{\alpha}$$

- **False Positive Visual Context (VR^{-low})** We employ the false-positive produced by the visual classifier as negative information to decrease the hypotheses.
- **Absent Visual Context (VR^{-high})** The negative information here is a set of visual information extracted from the original visual context that does not exist in the image but has some relation.
- **Positive Visual Context (VR^{-pos})** We approach this from a positive belief revision perspective but as negative evidence, as follows: (1) the similarity is computed without the context of the sentence, and (2) the static embedding is used and thus not knowing the sense of the word.



Dataset and Visual Context

We enrich COCO-Caption with textual visual context information.

- **Text hypothesis:** For Training, we use the five human annotated captions from the COCO-Caption dataset. For Testing we employ two baselines (1) VilBERT and Caption Transformer (Top-20 Beam search) from Karpathy split.
- **Visual Context:** Object classifier Resent152 (He et al., 2016) 1000 classes. Inception-ResNet Faster R-CNN (Huang et al., 2017) 80 classes.

Results

Model	B-1	B-4	M	R	C	S	BERTscore
VilBERT (Lu et al., 2020)							
$\text{Vil}_{\text{Greedy}}$	0.751	0.330	0.272	0.554	1.104	0.207	0.9352
$\text{Vil}_{\text{BeamS}}$	0.752	0.351	0.274	0.557	1.115	0.205	0.9363
$\text{Vil+VR}_{\text{W-Object}}$ (Fang et al., 2015)	**0.756**	0.348	0.274	**0.559**	1.123	0.206	0.9365
$\text{Vil+VR}_{\text{Object}}$ (Wang et al., 2018)	**0.756**	0.348	0.274	**0.559**	1.120	0.206	0.9364
$\text{Vil+VR}_{\text{Control}}$ (Cornia et al., 2019)	0.753	0.345	0.274	0.557	1.116	0.206	0.9361
$\text{Vil+VR}_{\text{RoBERTa}}$ (positive)	0.753	**0.353**	**0.276**	**0.559**	1.128	0.207	**0.9366**
$\text{Vil+VR}^{-low}_{\text{RoBERTa}}$	0.748	0.349	0.275	0.557	1.116	0.206	0.9362
$\text{Vil+VR}^{-high}_{\text{RoBERTa}}$	0.748	0.349	0.275	0.557	1.116	0.206	0.9364
$\text{Vil+VR}^{-pos}_{\text{GloVe}}$	0.751	0.351	0.276	0.558	1.123	0.207	0.9364
$\text{Vil+VR}^{-joint}_{\text{RoBERTa+GloVe}}$ (pos+neg)	0.750	0.351	**0.276**	**0.559**	**1.126**	**0.208**	**0.9365**
Transformer based caption generator (Cornia et al., 2020)							
$\text{Trans}_{\text{Greedy}}$	0.787	0.368	0.276	0.574	1.211	0.215	0.9376
$\text{Trans}_{\text{BeamS}}$	0.793	0.387	0.281	0.582	1.247	**0.220**	**0.9399**
$\text{Vil+VR}_{\text{W-Object}}$ (Fang et al., 2015)	0.786	0.348	0.274	0.559	1.123	0.206	0.9365
$\text{Trans+VR}_{\text{Object}}$ (Wang et al., 2018)	0.790	0.383	0.280	0.580	1.237	0.219	0.9391
$\text{Trans+VR}_{\text{Control}}$ (Cornia et al., 2019)	0.791	**0.388**	0.281	**0.583**	1.248	**0.220**	0.9398
$\text{Trans+VR}_{\text{BERT}}$ (positive)	0.793	**0.388**	**0.282**	**0.583**	**1.250**	**0.220**	**0.9399**
$\text{Trans+VR}^{-low}_{\text{BERT}}$	0.791	0.387	0.280	0.582	1.242	0.218	0.9396
$\text{Trans+VR}^{-high}_{\text{BERT}}$	0.793	0.385	0.282	0.582	1.243	0.219	0.9397
$\text{Trans+VR}^{-pos}_{\text{GloVe}}$ (negative)	**0.794**	**0.388**	**0.282**	**0.583**	**1.249**	**0.220**	**0.9399**
$\text{Trans+VR}^{-joint}_{\text{BERT+GloVe}}$	0.793	0.387	0.281	0.582	1.247	**0.220**	0.9398

Limitation

- **Semantic similarity score :** The unbalance similarity score (*e.g.* rare object) *sim*(visual/object, caption) negatively influences the revision.
- **Object detectors:** The failure cases of the visual classifier break the revision.

2022 – TALP Research Center, Universitat Politècnica de Catalunya