Belief Revision based Caption Re-ranker with Visual Semantic Information

Ahmed Sabir[1], Francesc Moreno-Noguer[2],
Pranava Madhyastha[3], 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

COLING 2022

Overview

Overview

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.

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**

Image credit: COCO-Captions (Lin et al., 2014)

- 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.

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
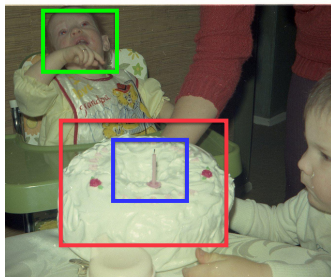
a baby sitting in front of a white cake

....

a baby **sitting** in front of a **birthday cake**

Image credit: COCO-Captions (Lin et al., 2014)

Background

- We propose a post-process based **Visual Beam** re-ranker that intends to visually ground the most closely related candidate beam to its related visual context in the image.

Image credit: COCO-Captions (Lin et al., 2014)

Background

Modern sophisticated image captioning systems focus heavily on visual grounding (object) to capture the detail of a static story in the image.

Visual Re-ranking
Fang et al. (2015)

visual detector to guide the image captioning

Object Freq Count
Wang et al. (2018)

investigates **informativeness** of object info

Controlled Caption
Cornia et al. (2019)

language grounding via object information

Semantic Coherency
Zhang el al. (2021)

explore reasoning in object grounding

Background

Modern sophisticated image captioning systems focus heavily on visual grounding (object) to capture the detail of a static story in the image.

Visual Re-ranking
Fang et al. (2015)

visual detector to guide the image captioning

Object Freq Count
Wang et al. (2018)

investigates **informativeness** of object info

Controlled Caption
Cornia et al. (2019)

language grounding via object information

Semantic Coherency
Zhang el al. (2021)

explore reasoning in object grounding

Background

Modern sophisticated image captioning systems focus heavily on visual grounding (object) to capture the detail of a static story in the image.

Visual Re-ranking
Fang et al. (2015)

visual detector to guide the image captioning

Object Freq Count
Wang et al. (2018)

investigates **informativeness** of object info

Controlled Caption
Cornia et al. (2019)

language grounding via object information

Semantic Coherency
Zheng el al. (2021)

explores reasoning in object grounding

Background

Modern sophisticated image captioning systems focus heavily on visual grounding (object) to capture the detail of a static story in the image.

Visual Re-ranking
Fang et al. (2015)

visual detector to guide the image captioning

Object Freq Count
Wang et al. (2018)

investigates **informativeness** of object info

Controlled Caption
Cornia et al. (2019)

language grounding via object information

Semantic Coherency
Zhang el al. (2021)

explores reasoning in object grounding

Contributions

- Enhance the performance of any typical image captioning system without the necessity for additional training.

- Propose a human-inspired general approach that aims to calibrate the original likelihood of top-n captions by beam search to re-rank the most closely related caption to the visual information in the image.

Visual: candle, wax light

IM GENET

Visual Revision

Baseline Beam search → Re-ranking

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

...

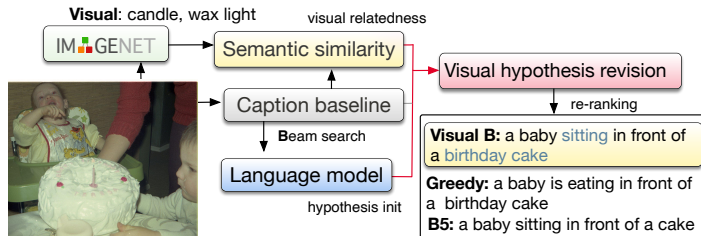a baby **sitting** in front of a **birthday cake**

Overview

Proposed Architecture: Visual Hypothesis Revision

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

Visual: candle, wax light

visual relatedness

IM**·**GENET → Semantic similarity

Caption baseline

Beam search

Language model

hypothesis init

Visual hypothesis revision

re-ranking

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

Belief Revision

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.**

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.**

Blok, Sergey, Douglas Medin, and Daniel Osherson. "Probability from similarity." AAAI. 2003.

Belief Revision

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.**

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 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**.

Blok, Sergey, Douglas Medin, and Daniel Osherson. "Probability from similarity." AAAI. 2003.

SimProb Model

Probability from Similarity `SimProb`

`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.**

$$P(Q_c|Q_a) = P(Q_c)^{\alpha}$$

Hypothesis: $P(Q_c)$

Informativeness: $1 - P(Q_a)$

Similarities: $\alpha = \left[\dfrac{1 - \text{sim}(a, c)}{1 + \text{sim}(a, c)} \right]^{1 - P(Q_a)}$

Probability from Similarity `SimProb`

`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.**

$$P(Q_c|Q_a) = P(Q_c)^{\alpha}$$

Hypothesis: $P(Q_c)$ Original belief

Informativeness: $1 - P(Q_a)$ New information

Similarities: $\alpha = \left[\dfrac{1 - \mathsf{sim}(a, c)}{1 + \mathsf{sim}(a, c)}\right]^{1 - P(Q_a)}$ Degree of similarity

SimProb Model

Probability from Similarity `SimProb`

`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.**

$$P(Q_c|Q_a) = P(Q_c)^{\alpha}$$

Hypothesis: $P(Q_c)$ Language Model

Informativeness: $1 - P(Q_a)$ Visual Context Information

Similarities: $\alpha = \left[\dfrac{1 - \mathrm{sim}(a,c)}{1 + \mathrm{sim}(a,c)}\right]^{1 - P(Q_a)}$ Semantic Similarity

SimProb Model

Sent Embedding is trained on hypothesis (w) to visual context (c)

$$C = \sum_{\substack{c_i \in Image \\ P(c_i) \geq \beta}} sim(w, c_i)$$

$$sim(w, c) = \frac{\vec{w} \cdot \vec{c}}{|\vec{w}| \cdot |\vec{c}|}$$

We convert the semantic score to probability according to assumption $p(w|c) \geqslant p(w)$. Thus the visual context asset the language model:

$$P(w|c) = P(LM)^{\alpha} \quad \text{where } \alpha = \left(\frac{1 - sim(w,c)}{1 + sim(w,c)}\right)^{1 - P(c)}$$

If there is no visual context information, α goes to $= 1$ the bare original belief probability is used.
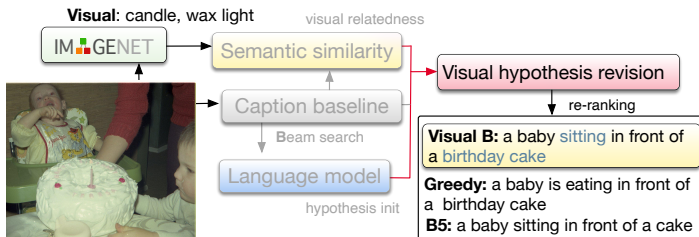
Approach

Hypothesis

- As this approach is inspired by humans, the hypothesis P(w) needs to be initialized by a common observation from general text.

- We employ GPT-2 (Radford et al., 2019) to initialize the hypothesis. We set P(w) as the mean token probability.
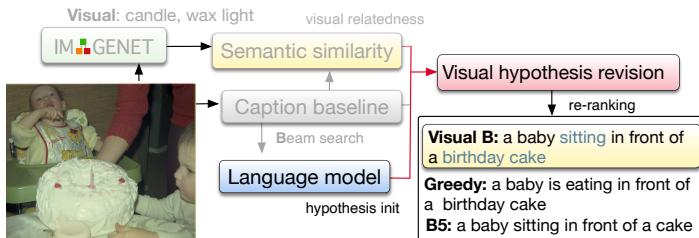
Informativeness

- The informativeness is the new information set $P(c)$ that causes the hypothesis revision.

- We leverage ResNet and Inception-ResNet v2 based faster R-CNN (Huang et al., 2017) to extract the visual information from the image.

Visual: candle, wax light

IM**GE**NET → Semantic similarity (visual relatedness) → Visual hypothesis revision

Caption baseline → Beam search → Language model (hypothesis init)

re-ranking

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

Approach

Similarities

- Hypothesis revision is more likely if there is a close relation between the hypothesis and new information.

- We employ BERT (fine-tuned) to compute the semantic similarity between the hypothesis (caption) and its related visual context.
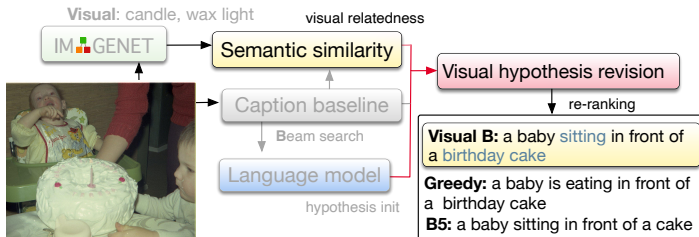
Text hypothesis

- Training: the five human annotated caption from COCO-Caption dataset (460k captions: 373k for training, and 87k for validation)

- Testing: Top-20 Beam search from the baseline 5k Karpathy Split

Visual context

- Object classifier Resent152 (He et al., 2016) 1000 classes.

- Inception-ResNet Faster R-CNN (Huang et al., 2017) 80 classes.

VC_1	VC_2	VC_3	Text hypothesis (Caption)
cheeseburger	plate	hotdog	a plate with a hamburger fries and tomatoes
bakery	dining table	web site	a table having tea and a cake on it
gown	groom	apron	its time to cut the cake at this couples wedding
racket	scoreboard	tennis ball	a crowd is watching a tennis game being played
laptop	screen	desktop computer	a grey kitten laying on a windows laptop
washbasin	toilet	seat tub	a bathroom toilet sitting on a stand next to a tub and sink

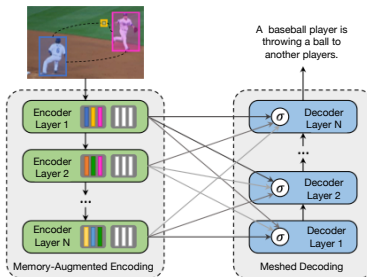# Overview

Baselines

- Transformer Caption Generator (12-layer Mech Transformer)
- ViLBERT Pre-trained Model (12-in-1 datasets)
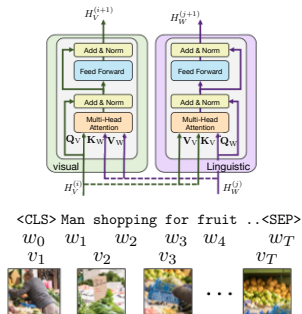

VilBERT Pre-trained Model (Lu et al., 2020)

Mesh-Caption Transformer (Cornia et al., 2020)

Result

Model	B-1	B-4	M	R	C	S	BERTscore
VilBERT (Lu et al., 2020)							
Vil_{Greedy}	0.751	0.330	0.272	0.554	1.104	0.207	0.9352
Vil_{BeamS}	0.752	0.351	0.274	0.557	1.115	0.205	0.9363
$Vil+VR_{W-Object}$ (Fang et al., 2015)	0.756	0.348	0.274	0.559	1.123	0.206	0.9365
$Vil+VR_{Object}$ (Wang et al., 2018)	0.756	0.348	0.274	0.559	1.120	0.206	0.9364
$Vil+VR_{Control}$ (Cornia et al., 2019)	0.753	0.345	0.274	0.557	1.116	0.206	0.9361
$Vil+VR_{BERT}$ (only *sim*)	0.753	0.343	0.273	0.556	1.112	0.206	0.9361
$Vil+VR_{BERT}$	0.752	0.351	0.274	0.557	1.115	0.205	0.9365
$Vil+VR_{BERT-Object}$	0.752	0.352	**0.277**	0.560	1.129	**0.208**	0.9365
$Vil+VR_{RoBERTa}$	0.753	**0.353**	0.276	0.559	1.128	0.207	**0.9366**
$Vil+VR_{RoBERTa-Object}$	**0.758**	0.344	0.262	0.555	1.234	0.206	0.9365
$Vil+VR_{BERT-Multi-class}$	0.753	**0.353**	0.276	0.559	1.131	**0.208**	0.9365
$Vil+VR_{BERT-Multi-object}$	0.752	0.351	0.276	0.558	1.123	**0.208**	0.9364
$Vil+VR_{RoBERTa-Multi-class}$	0.751	0.351	**0.277**	**0.561**	**1.137**	**0.208**	**0.9366**
$Vil+VR_{RoBERTa-Multi-object}$	0.752	**0.353**	**0.277**	0.559	1.131	**0.208**	**0.9366**
Transformer based caption generator (Cornia et al., 2020)							
$Trans_{Greedy}$	0.787	0.368	0.276	0.574	1.211	0.215	0.9376
$Trans_{BeamS}$	0.793	0.387	0.281	0.582	1.247	**0.220**	**0.9399**
$Trans+VR_{W-Object}$ (Fang et al., 2015)	0.786	0.378	0.277	0.579	1.228	0.216	0.9388
$Trans+VR_{Object}$ (Wang et al., 2018)	0.790	0.383	0.280	0.580	1.237	0.219	0.9391
$Trans+VR_{Control}$ (Cornia et al., 2019)	0.791	**0.388**	0.281	**0.583**	1.248	**0.220**	0.9398
$Trans+VR_{BERT}$ (only *sim*)	0.789	0.380	0.279	0.579	1.234	0.219	0.9389
$Trans+VR_{BERT}$	0.793	**0.388**	**0.282**	**0.583**	**1.250**	**0.220**	**0.9399**
$Trans+VR_{BERT-Object}$	0.793	0.385	0.281	0.581	1.242	0.219	0.9396
$Trans+VR_{RoBERTa}$	0.792	0.386	0.280	0.582	1.244	0.219	0.9395
$Trans+VR_{RoBERTa-Object}$	0.792	0.386	0.281	0.582	1.242	0.219	0.9396
$Trans+VR_{BERT-Multi-class}$	**0.794**	0.385	0.281	0.582	1.248	**0.220**	0.9395
$Trans+VR_{BERT-Multi-object}$	0.792	0.385	0.281	0.582	1.244	**0.220**	0.9395
$Trans+VR_{RoBERTa-Multi-class}$	0.791	0.385	0.280	0.581	1.244	0.219	0.9395
$Trans+VR_{RoBERTa-Multi-object}$	0.791	0.385	0.281	0.582	1.243	0.219	0.9395

Multiple Visual Contexts

- We observe that longer captions benefit from more context via multiple visuals.

Visual Context   Multi-Visual Contexts
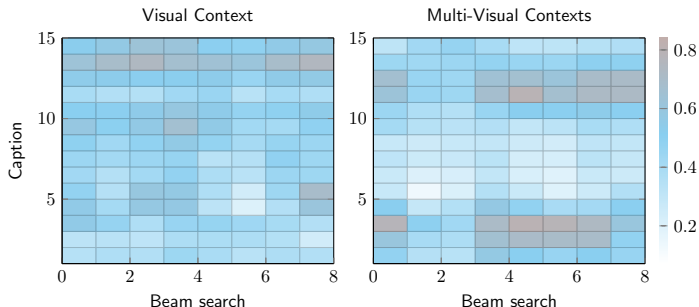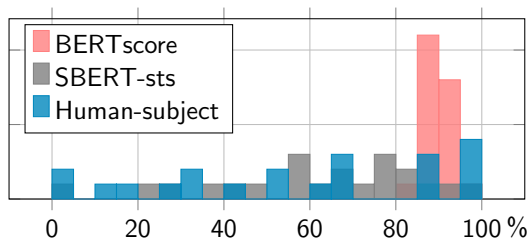
Diversity Evaluation

- **Lexical Diversity:** Textual Lexical Diversity (MTLD), Type Token Ratio (TTR) and Word per Caption (WPC)
- **N-gram Diversity:** Div-1 (n-gram), Div-2 (bi-gram) and mBLEU
- **Semantic Diversity:** Self-CIDEr and sentence SBERT-sts

| | Lexical Diversity | | | | Voc | mBLEU↓ | | n-gram Diversity | | Semantic Diversity | |
	MTLD	TTR	Uniq	WPC	Dist	best-5	best Beam*	Div-1	Div-2	Self-CIDEr	SBERT-sts
Human	19.56	0.90	9.14	14.5	3425						
ViLBERT											
Vil_BeamS	17.28	0.87	8.05	10.5	894	0.899	0.454	0.38	0.44	0.661	0.7550
Vil+VR_w-Object	15.90	0.87	8.02	9.20	921	0.899	0.455	0.38	0.44	**0.662**	**0.7605**
Vil+VR_Object	15.77	0.87	8.03	9.19	911	0.899	0.455	0.38	0.44	0.661	0.7570
Vil+VR_Control	15.69	0.87	8.07	9.21	**935**	0.899	0.452	0.38	0.44	0.661	0.7567
Vil+VR_RoBERTa (ours)	**17.70**	0.87	**8.14**	**10.8**	892	**0.896**	**0.451**	0.38	0.44	0.661	0.7562
Transformer based caption generator											
Trans_BeamS	14.77	0.86	7.44	9.62	935	**0.954**	0.499	0.26	0.29	0.660	0.7707
Trans+VR_w-Object	13.14	0.85	7.37	8.62	965	0.958	0.498	0.25	0.29	0.660	0.7709
Trans+VR_Object	13.38	0.86	**7.45**	8.69	**982**	0.958	0.495	0.25	0.29	0.660	0.7700
Trans+VR_Control	13.25	0.86	7.44	8.64	961	0.958	0.498	0.25	0.29	0.660	**0.7716**
Trans+VR_BERT (ours)	**14.78**	0.86	**7.45**	**9.76**	980	0.963	**0.338**	0.26	**0.30**	0.660	0.7711

please refer to the paper for all the metrics references.

Human Evaluation

- We conducted a human study to investigate human preferences over the visual re-ranked caption.

- We can observe that 46% of **native speakers** agreed with our visual re-ranker. Meanwhile, the result for **non-native** speakers is 61%

Result - Examples

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				

Ablation study

- Belief Revision relies on a different block (*i.e.* LM, similarity and visual context) to make the final revision

- We perform an ablation study over a random 100 samples from the test set to investigate the effectiveness of the proposed setup

Model	B-4	M	R	C	S
ViLBERT-VR-GPT-2 + ResNet					
+ RoBERTa (proposed)	0.346	0.266	**0.541**	**1.171**	0.205
+ DistilSBERT (Reimers et al., 2019)	0.335	0.266	0.537	1.128	0.205
+ SimCSE (Gao et al., 2021)	0.324	0.263	0.529	1.122	**0.207**
+ SimCSE (unsupervised)	**0.349**	**0.267**	0.539	1.164	0.205
+ CLIP (Radford et al., 2021)	0.335	0.261	0.527	1.142	0.202
Trans - VR-GPT-2 + ResNet					
+ BERT (proposed)	**0.363**	**0.268**	**0.565**	**1.281**	0.207
+ DistilSBERT (Reimers et al., 2019)	0.355	0.260	0.557	1.249	0.205
+ SimCSE (Gao et al., 2021)	0.356	0.265	0.564	1.272	0.207
+ SimCSE (unsupervised)	0.356	0.263	0.560	1.253	**0.208**
+ CLIP (Radford et al., 2021)	0.349	0.260	0.555	1.243	0.203

Belief Revision with Negative Evidence

`SimProb` with Negative Evidence (Blok et al., 2007)

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

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

Sergey V Blok, Douglas L, and Daniel Osherson."Induction as conditional probability judgment." Memory & Cognition 2007.

Belief Revision with Negative Evidence

`SimProb` with Negative Evidence (Blok et al., 2007)

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

Sergey V Blok, Douglas L, and Daniel Osherson."Induction as conditional probability judgment." Memory & Cognition 2007.

Belief Revision with Negative evidence
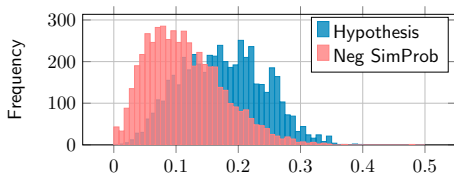
A **False Positive Visual Context (VR^{-low}):** We employ the false-positive produced by the visual classifier as negative information to decrease the hypotheses.

A **False Positive Visual Context (VR^{-low}):** We employ the false-positive produced by the visual classifier as negative information to decrease the hypotheses.

B **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.

Belief Revision with Negative evidence

A **False Positive Visual Context (VR^{-low}):** We employ the false-positive produced by the visual classifier as negative information to decrease the hypotheses.

B **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.

C **Positive Visual Context (VR^{-pos}):** We approach this from a positive belief revision perspective but as negative evidence. (1) the similarity is measured without the context of the sentence and (2) the static embedding is computed without knowing the sense of the word.

Result

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

Visual information: Object classifier failure cases

Model	Caption	BERTscore	SBERT-sts	Human%	Visual
BeamS	a pile of trash sitting inside of a building	0.88	**0.38**	**100**	✗ vacuum
VR	a pile of trash sitting in front of a building ✗	0.88	0.27	0	
Human refe	an older floor light sits deserted in an abandoned hospital				
BeamS	a kitchen with black counter tops and wooden cabinets	0.88	**0.44**	**100**	✗ barbershop
VR	a kitchen counter with a black counter top	0.88	0.40	0	
Human refe	a kitchen with a sink bottles jars and a dishwasher				

Overview

Conclusion

Contributions

- We demonstrate that the Belief Revision (BR) approach that works well with human judgment can be applied to Image Captioning by employing human-inspired reasoning via a pre-trained model.

- We propose two models to improve image caption systems (1) BR with positive visual evidence (increase the hypothesis) and (2) negative evidence (decrease the hypothesis), with wrong visual.

Future Work
We plan to apply the Belief Revision Score to many re-ranking tasks in NLP such as text generation, multimodel MT, and lexical selection.

Thank You