Belief Revision based Caption Re-ranker with Visual Semantic Information

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

[1] Computer Science Department, Universitat Politècnica de Catalunya, Barcelona, Spain
[2] Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Barcelona, Spain
[3] City, University of London, London, UK

Abstract

In this work, we focus on improving the captions generated by image-caption generation systems. We propose a novel re-ranking approach that leverages visual-semantic measures to identify the ideal caption that maximally captures the visual information in the image. Our re-ranker utilizes the Belief Revision framework (Blok et al., 2003) to calibrate the original likelihood of the top-n captions by explicitly exploiting the semantic relatedness between the depicted caption and the visual context. Our experiments demonstrate the utility of our approach, where we observe that our re-ranker can enhance the performance of a typical image-captioning system without the necessity of any additional training or fine-tuning.[1]

1 Introduction

Image caption generation is a task that predominantly lies at the intersection of the areas of computer vision and natural language processing. The task is primarily aimed at generating a natural language description for a given image. Caption generation systems usually consist of an image encoder that encodes a given image (usually by using a CNN) whose encoding is fed to a decoder (usually by using a generative model such as RNN) to generate a natural language sentence which describes the image succinctly. The most widely used approaches include a CNN-RNN end-to-end system (Vinyals et al., 2015; Anderson et al., 2018), end-to-end systems with attention that attend to specific regions of the image for generation (Xu et al., 2015; You et al., 2016) and systems with reinforcement learning based methods (Rennie et al., 2017; Ren et al., 2017). Furthermore, recent advances have resulted in end-to-end systems that use Transformer based architecture for language generation and have become the current state-of-the-art
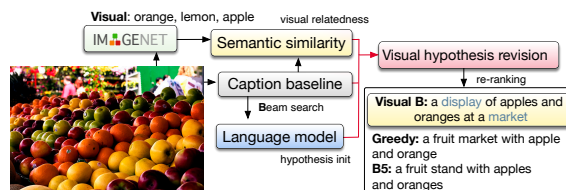

Figure 1: An overview of our hypothesis revision based visual re-ranker. We use the visual context from the image to *revise* and re-rank the most closely related caption to its visual context. These semantic relatedness measures are learned at the word-to-sentence level. In this example, we showcase our visual re-ranker (**V**isual **B**eam), a post-processing approach, which is able to re-rank the most 'descriptive caption' from the 5-Best **B**eam (Cornia et al., 2020).

(Huang et al., 2019; Cornia et al., 2020; Zhang et al., 2021b).

While the state-of-the-art models generate captions that are comparable to human level captions, they are known to lack lexical diversity, are often not very distinct, and sound synthetic. We here highlight a few recent approaches that have focused on this problem, these include Dai et al. (2017) that uses generative adversarial networks towards generating diverse and human like captions. Vedantam et al. (2017) use a beam search with a distractor image to force the model to produce diverse captions by encouraging the models to be discriminative. Other 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). In this work, we follow a similar line of research and focus on the problem of improving diversity and making captions natural and human like and propose a novel re-ranking approach. In this approach, we use n-best reranking with a given beam that explicitly uses the semantic correlation between the caption and the visual context through belief revision (an approach inspired by human logic). We refer the reader to Figure 1, where the

---

[1] https://github.com/ahmedssabir/Belief-Revision-Score

approach results in a caption that is a) visually relevant and b) the most natural and human like.

Our primary contributions in this paper are:

- We demonstrate the utility of the Belief Revision (Blok et al., 2003) framework, which has been shown to correlate highly with human judgment and has demonstrate its applicability to the task of Image Captioning. We do this by employing vision-language joint semantic measures using state-of-the-art pre-trained language models.
- Our approach is a *post-processing* method and is devised to be a drop-in replacement for any caption system.
- Through our experiments, we report that our proposal selects better captions as reported using automated metrics, as well as being validated by human evaluations.

2 Belief Revision with `SimProb` Model

In this section, we briefly introduce SimProb, which is based on the philosophical intuitions of Belief Revision, an idea that helps to convert similarity measures to probability estimates. Blok et al. (2003) introduce a conditional probability model that assumes that the preliminary probability result is updated or revised to the degree that the hypothesis proof warrants. The range of revision is based on the informativeness of the argument and its degree of similarity. That is, the similarity to probability conversion can be defined in terms of **Belief Revision**. Belief Revision is a process of forming a belief by taking into account a new piece of information.

Let us consider the following statements:

```
1 Tigers can bite through wire,
  therefore Jaguars can bite through
  wire.
2 Kittens can bite through wire,
  therefore Jaguars can bite through
  wire.
```

In the first case, the statement seems logical because it matches our prior belief *i.e.* jaguars are similar to tigers, so we expect them to be able to do similar things. We hence consider that the statement is consistent with our previous belief, and there is no need to revise it. In the second case, the statement is surprising because our prior belief is that kittens are not as similar to jaguars, and thus, not so strong. But if we assume the veracity of the statement, then we need to revise and update our prior belief about the strength of kittens.

This work formalizes belief as probabilities and revised belief as conditional probabilities and provides a framework to compute them based on the similarities of the involved objects. According to the authors, belief revision should be proportional to the similarity of the involved objects (*i.e.* in the example, the statement about kittens and jaguars would cause a stronger belief revision than *e.g.* the same statement involving pigeons and jaguars because they are less similar). In our case, we use the same rationale and the same formulas to convert similarity (or relatedness) scores into probabilities suitable for reranking.

`SimProb` Model To obtain the likelihood revisions based on similarity scores, we need three parameters: (1) **Hypothesis**: prior probabilities, (2) **Informativeness**: conclusion events and (3) **Similarities**: measuring the relatedness between involved categories. The goal is to predict a conditional probability of statements, given one or more other statements. In order to predict the conditional probability of the argument's conclusion, given its premise or hypothesis, we will need only the prior probabilities of the statements, as well as the similarities between the involved categories (*e.g.* kittens and tigers).

Formulation of `SimProb` The conditional probability $\mathrm{P}(Q_c|Q_a)$ is expressed in terms of the prior probability of the conclusion statement $\mathrm{P}(Q_c)$, the prior probability of the premise statement $\mathrm{P}(Q_a)$, and the similarity between the conclusion and the premise categories sim(a, c).

$$\mathrm{P}\left(Q_c \mid Q_a\right) = \mathrm{P}\left(Q_c\right)^\alpha \text{ where } \alpha = \left[\frac{1-\mathrm{sim}(a,c)}{1+\mathrm{sim}(a,c)}\right]^{1-\mathrm{P}(Q_a)}$$

Belief Revision Elements As we discussed above, there are two factors that determine the hypothesis probability revision: 1) the sufficient relatedness to the category: as $sim(a,c) \rightarrow 0$, $\alpha \rightarrow 1$, and thus $\mathrm{P}(Q_c|Q_a) = \mathrm{P}(Q_c)$, *i.e.* no revision takes place, as there are no changes in the original belief. While as $sim(a,c) \rightarrow 1$, $\alpha \rightarrow 0$, and the hypothesis probability $\mathrm{P}(Q_c)$ is revised and is raised closer to 1; 2) the informativeness of the new information $1 - \mathrm{P}(Q_a)$: as $\mathrm{P}(Q_a) \rightarrow 1$ and in consequence is less informative, $\alpha \rightarrow 1$, as there is no new information, and hence no revision is required.

3 Visual Re-ranking for Image Caption

3.1 Problem Formulation

The beam search is the dominant method for approximate decoding in structured prediction tasks such as machine translation, speech recognition

and image captioning. A larger beam size allows the model to perform a better exploration in the search space compared to greedy decoding. The main idea of the beam search is to explore all possible captions in the search space by keeping a set of *top candidates*.

Our goal is to leverage the visual context information of the image to re-rank the candidate sequences obtained through the beam search, thereby moving the most visually relevant candidate up in the list, as well as moving wrong candidates down. For this purpose, we experiment with different re-rankers, based on the relatedness between the candidate caption and the semantic context observed in the image through the idea of Belief Revision.

Caption Extraction We employ two recent Transformer based architectures for caption generation to extract the top candidate captions using different beam sizes ($B = 1 \dots 20$) (Vijayakumar et al., 2018). The first baseline is based on a multi-task model for discriminative Vision and Language BERT (Lu et al., 2020) that is fine-tuned on 12 downstream tasks. The second baseline is the vanilla Transformer (Vaswani et al., 2017) with the Meshed-Memory based caption generator (Cornia et al., 2020) with pre-computed top-down visual features (Anderson et al., 2018).

3.2 Proposal

One approach of using word-level semantic relations for scene text correction with the visual context of an image was introduced in Sabir et al. (2018), which allows for the establishment of learning semantic correlations between a visual context and a text fragment. In our work, this semantic relatedness is between a visual context and a given candidate caption (*i.e.* beam search), and uses Belief Revision (BR) via `SimProb` to re-visit and re-rank the original beam search based on the similarity to the *image objects/labels* c (a proxy for image context). The BR in this scenario is a conditional probability which assumes that the caption preliminary probability (*hypothesis*) $P(w)$ is revised to the degree approved by the semantic similarity with visual context $\text{sim}(w, c)$. The final output caption w for a given visual context c is written as:

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

where the main components of visual based hypothesis revision:

Hypothesis: $P(w)$

Informativeness: $1 - P(c)$

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

where $P(w)$ is the *hypothesis* probability (beam search candidate caption) and $P(c)$ is the probability of the evidence that causes hypothesis probability revision (visual context from the image). We next discuss the details of each component in `SimProb` as visual based re-ranker.

Hypothesis: 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 (LM) trained on a general text corpus. Therefore, we employ a Generative Pre-trained Transformer (GPT-2) (Radford et al., 2019) a LM to initialize the hypothesis probability. We set $P(w)$ as the mean of LM token probability.

Informativeness: Inversely related to the probability of set $P(c)$ 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)[2] to extract textual visual context information from the image. We use the classifier probability confidence with a threshold to filter out non-existent objects in the image. For each image, we extract visual information as follows: (1) top-1 concept (2) multi concept top-3 (label class or object category) visual information. For the single concept, we employ a unigram LM, based on the 3M-token opensubtitles corpus (Lison and Tiedemann, 2016), to initialize the informativeness of the visual information. For multiple concepts, we take the mean probability of the three concepts. Note that we are initializing the single visual context with LM to maximize the visual context score while computing the informativeness.

Similarities: Hypothesis revision is more likely if there is a close relation between the hypothesis and the new information (candidate caption and visual context in our case). We rely on two of the most recent state-of-the-art pre-trained Transformer-based language models to compute the semantic similarity between the caption and its visual context information with contextual embedding. In particular, we utilize the visual as context for the sentence (*i.e.* caption) to compute the cosine distance:

- **BERT (Devlin et al., 2019):** BERT achieves remarkable results on many sentence level tasks and especially in the textual semantic

---

[2]TensorFlow Object Detection API

similarity task (STS-B) (Cer et al., 2017). Therefore, we fine-tuned BERT$_{base}$ on the training dataset, (textual information, 460k captions: 373k for training and 87k for validation) *i.e.* visual context, caption, label [semantically related or not related]), with a binary classification cross-entropy loss function [0,1] where the target is the semantic similarity between the visual and the candidate caption, with batch size 16 for two epochs with a learning rate $2e-5$.

- **RoBERTa (Liu et al., 2019):** RoBERTa is an improved version of BERT, trained on a large amount of data, using dynamic masking strategies to prevent overfitting. It achieves a 2.4% improvement over BERT$_{Large}$ in the STS task. Since RoBERTa$_{Large}$ is more robust, we use an off-the-shelf model tuned on STS-B task. In particular, we follow the traditional approach to compute the semantic similarity with a BERT based model with a mean pool, over the last hidden layer, to extract a meaningful vector to compute the cosine distance.

4 Experiments

4.1 Dataset

COCO-Caption (Lin et al., 2014): This dataset contains around 120k images and each image is annotated with five different human-written captions. We use the split provided by (Karpathy and Fei-Fei, 2015), where 5k images are used for testing, 5k for validation, and the rest for model training.

Visual Context Enrichment: We enrich COCO-Caption with textual visual context information. To automate visual context generation and without the need for a human label, for the training dataset, we use only ResNet152, which has 1000 label classes, to extract the top-k three label class visual context information for each image in the caption dataset. For testing, we rely only on the top-k visual information as a concept, and we also employ the Inception-ResNet v2 based Faster R-CNN object detector with 80 object classes. In particular, each single annotated caption has three visual context information.

Evaluation Metric: We use the official COCO offline evaluation suite, producing several widely used caption quality metrics: **B**LEU (Papineni et al., 2002) **M**ETEOR (Banerjee and Lavie, 2005), **R**OUGE (Lin, 2004), CIDEr (Vedantam
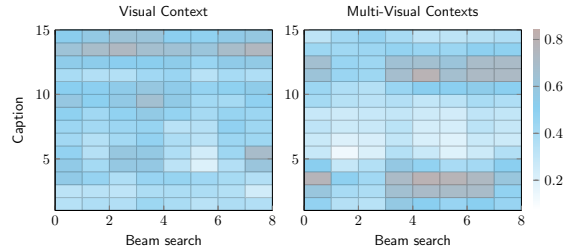
Figure 2: Visualization of top-k nine re-ranked Beam search via `SimProb` with, Vil+VR$_{RoBERTa}$ (Right) multiple visual and (Left) one concept visual context. The longer caption benefits from using multiple concepts.

et al., 2015), **S**PICE (Anderson et al., 2016) and BERTscore (Zhang et al., 2020).

4.2 Results and Discussion

We use visual semantic information to re-rank candidate captions produced by out-of-the-box state-of-the-art caption generators. We extract the top-20 beam search candidate captions from two state-of-the-art models: VilBERT (Lu et al., 2020), fine-tuned on a total of 12 different vision and language datasets such as caption image retrieval and visual question answering, and a specialized caption-based Transformer (Cornia et al., 2020).

Experiments applying different rerankers to the each base system are shown in Table 1. The tested rerankers are: (1) VR$_{BERT}$ using BERT similarity between the candidate caption and the visual context of the image, transforming it to a probability using Equation 1, and combines the result with the original candidate probability to obtain the reranked score. (2) VR$_{RoBERTa}$ carrying out the same procedure using similarity produced by RoBERTa. A simpler model is also tested –VR$_{BERT}$ (only sim) in Table 1–, which replaces Equation 1 with $P(w \mid c) = sim(w, c)^{P(c)}$, that is, it does not rely on the original caption probability.

First, we compare our work with the original visual caption re-ranker with multiple word objects as concepts from the image, that are extracted via Inception-ResNet v2 based Faster RCNN (*i.e.* person, van, *etc.*), VR$_{w-Object}$ (Fang et al., 2015). However, to make a fair comparison, we use the Sentence-RoBERTa$_{Large}$ for the sentence semantic similarity model *i.e.* cosine(word objects, caption). Secondly, we compare our model against two approaches that uses object information to improve image captioning: First, Wang et al. (2018) investigates the benefit of object frequency counts for generating a good captions. We train an LSTM

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

Table 1: Performance of compared baselines on the Karpathy test split with/without **V**isual semantic **R**e-ranking. For each base system, we report performance using a greedy search and the best beam search. Re-ranking is applied to the top-20 results of each system using BERT or RoBERTa for caption-context similarity. The visual contexts are extracted using ResNet152 and Inception Resnet v2 based Faster R-CNN **object** detector. We also report results for Bert-based similarity without a hypothesis probability (rows marked *only sim*).

decoder (*i.e.* language generation stage) with an object frequency counts dictionary on the training dataset. The dictionary is a Fully Connected layer, concatenated with the LSTM and a dense layer, that adds more weight to the most frequent counts object that are seen by the caption and the visual classifier. Second, Cornia et al. (2019) that introduce a controllable grounded captions via a visual context. We train the last stage (decoder), attention and language model LSTM, on the training dataset to visually ground the generated caption based on the visual context.

One observation, shown in Table 1, is that the benefit of using multiple visual contexts for longer captions, which can increase the chance of re-ranking the most visually related candidate caption, as shown in Figure 2 SimProb score with VilBERT.

Also, we investigate the statistical significance, using approximate randomization and bootstrapping resampling (Koehn, 2004), to detect minute differences in BLEU and METEOR, and NIST-BLEU[3] (Doddington, 2002) scores, as shown in

Table 2, in which we observe the improvement with our re-ranker over BLEU, METEOR[4] and NIST-BLEU. We would like to remark here that, with regards to subtle variations, the statistical significance of metrics such as BLEU, NIST-BLEU, and SacreBLEU (Post, 2018) tend to disagree with human judgement (Mathur et al., 2020; Kocmi et al., 2021). We therefore also conduct a human evaluation study (Section 6).

Figure 4 shows SimProb distribution over 40k samples with a pre-trained RoBERTa_Large similarity score. (Left Figure) Before applying the revision, overall re-ranking scores are relatively low, and (Right Figure) after the visual revision, overall scores increased with more confident about each selected caption. The SimProb score positively shifts the distribution over all the samples.

4.3 Limitation

We note that, the quality of the Beam search influences our re-ranker, since non-diverse, repeated captions or fewer novel ones, will make the re-

---

[3]An improved version of BLEU rewarded infrequently

used words by giving greater weighting to rarer words.

[4]It has been previously observed that METEOR correlates better with human judgments than BLEU.

Vil_BeamS: a group of people standing around a white truck
Vil+VR_BERT: a group of people standing around a white *van*
Human: there is a white van that is stopped on the road

Visual: Minibus

Vil_BeamS: a close up of a plate of food
Vil+VR_BERT: *piece* of food sitting on top of *a white plate*
Human: a white plate and a piece of white cake

Visual: trifle

Vil_BeamS: a group of men on a field playing baseball
Vil+VR_BERT: *a batter catcher and umpire during a baseball game*
Human: batter catcher and umpire anticipating the next pitch
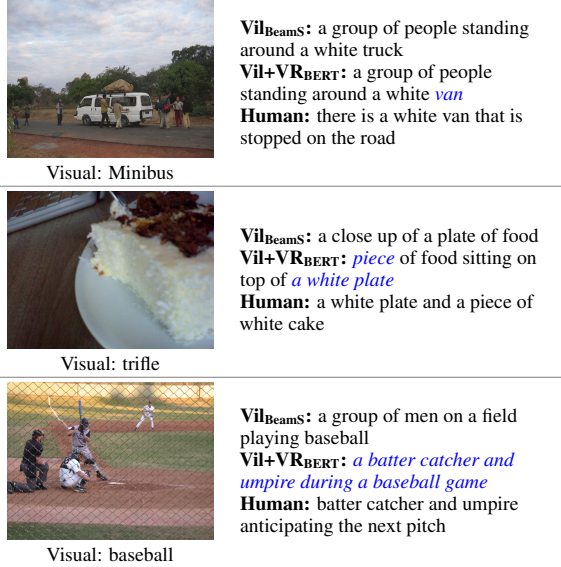
Visual: baseball

Figure 3: Example of captions re-ranked by our Visual Re-ranker and the original caption (Best-beam) from the base system. Re-ranked captions are more precise, have a higher lexical diversity, or provide more details.

ranking less effective, as shown in the Unique words per caption in Table 2 with the Transformer baseline.

5 Evaluation of Diversity

We follow the standard diversity evaluation metrics (Shetty et al., 2017; Deshpande et al., 2019): (1) *Div*-1 the ratio of unique unigram to the number of words in caption (2) *Div*-2 the ratio of unique bi-gram to the number of words in the caption, (3) *mBLEU* is the BLEU score between the candidate caption against all human captions (lower value indicate diversity). However, since even though we obtained the top-20 candidates from the base systems, many of them are the same or have very small differences (beam search drawback), which will reflect in small performance differences before and after re-ranking. Therefore, some of the standard metrics are not able to capture these small changes, as shown in Table 2. Consequently, to try to capture the changes and the effect of the re-ranking, we also measured the lexical and semantic diversity with the following metrics:

Type-Token Ratio (TTR): TTR (Brown, 2005) is the number of unique words or types divided by the total number of tokens in a text fragment.

Measure of Textual Lexical Diversity (MTLD): MTLD (McCarthy and Jarvis, 2010) is based on TTR, and measures the average length of subsequences in the text for which a certain TTR is main-
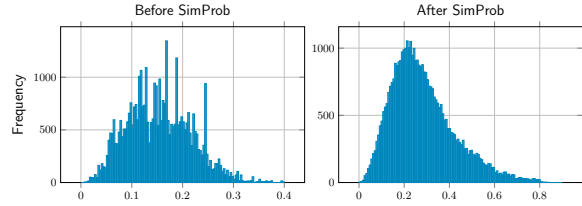

Figure 4: Visualization of the distribution change in re-rank scores on the 40k random sample from the test set. (Left) the score distribution before applying Belief Revision via `SimProb` with LM-GPT-2 initialization. (Right) the score distribution after applying the revision via similarity RoBERTa_Large with the visual context.

tained, thus, unlike TTR, being length-invariant.

For semantic diversity, we use the standard metric (Wang and Chan, 2019) Self-CIDEr. Also, inspired by BERTscore and following (Song et al., 2021) that introduce Sentence Semantic Semantic (SSS) for machine translation, we use SBERT-sts (fine-tuned on sts task) with a cosine score to measure the sentence level semantic correlation against all human references. We observe that the SBERT-sts capture the semantic content better than Self-CIDEr.

Table 2 shows that visual re-ranking selects longer captions and with higher lexical diversity than the base system beam search. Figure 3 shows some examples where visual context re-ranking selected captions with more precise lexica (*van* vs. *truck*), higher diversity –*i.e.* adding details about objects (*white plate* vs. *plate*)– and even selecting a more specific abstraction level (*batter, catcher and umpire* vs. *a group of men*).

6 Human Evaluation

We conducted a human study to investigate human preferences over the visual re-ranked caption. We randomly select 26 test images and give 12 reliable human subjects the option to choose between two captions: (1) Best-beam (BeamS)[5] and (2) **V**isual **R**-ranker. We obtain mixed results, as some re-ranked captions are grammatically incorrect, such as singulars instead of plurals and *sitting on* for objects instead of subjects. Overall, we can observe that 46% of native speakers agreed with our visual re-ranker. Meanwhile, the result for non-native speakers is 61%. In some details, we observe that our model and the non-native human subjects chose those re-ranked captions because they correlated more closely with the visual information regarding

---

[5]The best result by the baseline in standard metrics.

	Lexical Diversity				Vocabulary		Accuracy 4-gram (p-value)						mBLEU↓		n-gram Diversity		Semantic Diversity	
	MTLD	TTR	Uniq	WPC	Dist	Dist*	BLEU	p	M	p	NIST	p	best-5	best Beam*	Div-1	Div-2	Self-CIDEr	SBERT-sts
Human	19.56	0.90	9.14	14.5	3425	3326												
VilBERT																		
Vil$_{BeamS}$	17.28	0.87	8.05	10.5	894	842	0.337	-	0.265	-	0.755	-	0.899	0.454	0.38	0.44	0.661	0.7550
Vil+VR$_{w\text{-}Object}$ (Fang et al., 2015)	15.90	0.87	8.02	9.20	921	866	0.335	0.109	0.266	0.46	0.764	0.00	0.899	0.455	0.38	0.44	**0.662**	**0.7605**
Vil+VR$_{Object}$ (Wang et al., 2018)	15.77	0.87	8.03	9.19	911	854	0.335	0.131	0.266	0.57	0.761	0.043	0.899	0.455	0.38	0.44	0.661	0.7570
Vil+VR$_{Control}$ (Cornia et al., 2019)	15.69	0.87	8.07	9.21	**935**	**878**	0.331	0.016	0.266	0.46	0.758	0.118	0.899	0.452	0.38	0.44	0.661	0.7567
Vil+VR$_{RoBERTa}$ (ours) (Table 1 Best)	**17.70**	0.87	**8.14**	**10.8**	892	838	**0.339**	0.147	**0.267**	0.04	**0.764**	0.002	**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	897	0.341	-	0.272	-	0.781	-	**0.954**	0.499	0.26	0.29	0.660	0.7707
Trans+VR$_{w\text{-}Object}$ (Fang et al., 2015)	13.14	0.85	7.37	8.62	965	923	0.364	0.00	0.272	0.10	0.789	0.001	0.958	0.498	0.25	0.29	0.660	0.7709
Trans+VR$_{Object}$ (Wang et al., 2018)	13.38	0.86	**7.45**	8.69	**982**	**940**	0.369	0.00	0.271	0.04	0.798	0.00	0.958	0.495	0.25	0.28	0.660	0.7700
Trans+VR$_{Control}$ (Cornia et al., 2019)	13.25	0.86	7.44	8.64	961	921	0.373	0.00	0.272	0.00	0.796	0.00	0.958	0.498	0.25	0.29	0.660	**0.7716**
Trans+VR$_{BERT}$ (ours) (Table 1 Best)	14.78	0.86	**7.45**	9.76	980	939	**0.374**	0.00	**0.273**	0.19	**0.806**	0.00	0.963	**0.338**	0.26	**0.30**	0.660	0.7711

Table 2: Diversity statistics and statistical tests. Measuring the diversity of caption before/after re-ranking. Uniq and WPC columns indicate the average of unique/total words per caption, respectively. The BLEU, METEOR and NIST are an average result, with an approximate randomization test with 1k trials, to estimate the statistically significant improvement with/without our re-ranker. mBLEU and mBLEU* are computed with respect to the top 5-captions and best-beam, respectively. We also report the Distinct vocabulary (Dist* filtering out stop words).

the grammatical error in the sentence and unlike the native speaker.

7 Ablation study

Belief Revision relies on a different block (*i.e.* LM, similarity and visual context) to make the final revision. In this study, we perform an ablation study over a random 100 samples from the test set to investigate the effectiveness of the proposed setup. Table 3 shows result with different settings.

Language Model Block: One of the principal intentions in initializing the original hypothesis with a LM is the ability to combine different models. We experimented with product probability, although the mean LM probability achieved better results.

Similarity Block: The degree of similarity between the caption and its visual context is major factor in hypothesis revision. Thus, we experimented with a light model (Distil SBERT) and unsupervised/supervised **Sim**ple **C**ontrastive **S**entence **E**mbedding (SimCSE) for learning sentence similarity. The results show that unsupervised, via dropout with the sentence itself, contrastive learning based similarity performs well in the case of the longer captions, as shown in VilBERT Table 3.

Visual Context Block: We experimented with the most recent model of Contrastive Language-Image Pre-Training (CLIP) (Radford et al., 2021) with Zero-Shot Prediction to extract the visual context. Although, CLIP can predict rare objects better, there is no improvement over ResNet152 with a huge computational cost.

8 Negative Evidence: an extension

Until now, following Blok et al. (2003), we considered only the cases when the visual context increase

Model	B-4	M	R	C	S
VilBERT-VR-GPT-2$_{mean}$ + ResNet					
+ RoBERTa (Table 1 Best)	0.346	0.266	**0.541**	**1.171**	0.205
+ LM-GPT-2$_{product}$	0.335	0.266	0.535	1.142	0.205
+ DistilSBERT	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$_{mean}$ + ResNet					
+ BERT (Table 1 Best)	**0.363**	**0.268**	**0.565**	**1.281**	0.207
+ LM-GPT-2$_{product}$	0.360	0.261	0.561	1.254	0.205
+ DistilSBERT	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

Table 3: Ablation study using different information from various baselines in each block (*i.e.* LM, similarity and visual context).

the belief of the hypothesis (Equation 1). Blok et al. (2007) also propose Equation 2 for the case where the absence of evidence leads to a decrease in the probability of the hypothesis.

$$\mathrm{P}(w \mid \neg c) = 1 - (1 - \mathrm{P}(w))^{\alpha} \qquad (2)$$

In our case, we introduce negative evidence in three ways:

False Positive Visual Context (VR^{-low}): We employ the false-positives produced by the visual classifier as negative information to decrease the hypotheses. In this case, we have lower similarity measures as the relation between the visual context and caption are farther apart.

Absent Visual Context (VR^{-high}): The negative information here is a set of visual information extracted from the original visual context (*i.e.* from the visual classifier) which does not exist in the image. Thus, the visual context produced by the classifier is used as a query on a pre-trained 840B GloVe (Pennington et al., 2014), with cosine similarity, to retrieve the closest visual context in the same semantic space (*e.g.* visual: river, closest visual: valley).

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

Table 4: Comparison between positive (single concept VR) and Negative Belief Revision (NBR) on the Karpathy split. The NBR uses a *high similarity* VR^{-high} object related to the positive visual but not in the image, *low similarity* VR^{-low} false positive from the visual classifier, and positive visual via static word level similarity VR^{-pos}. **Boldface** fonts reflect improvement over the baseline.

Positive Visual Context (VR^{-pos}): As the previous two-approaches produced unexpected results with low and high similarities as shown in Table 4, we approach this from a positive belief revision perspective but as negative evidence. Until now, all approaches use sentence-level semantic similarity, but in this experiment, we convert the similarity from sentence to word level. For this first, we employ an LSTM based CopyRNN keyphrase extractor (Meng et al., 2017), which is trained on a combined pre-processed wikidump (*i.e.* keyword, short sentence) and SemEval 2017 Task 10 (Keyphrases from scientific publications) (Augenstein et al., 2017). Secondly, GloVe is used to compute the cosine similarity with the visual context in a word-level manner. We consider this as negative evidence for the following reasons: (1) the similarity is computed without the context of the sentence and (2) the static embedding is computed without knowing the sense of the word. The advantage of VR^{-pos} is that a high similarity and confident visual information are present and thus satisfies the revision.

Joint Evidence (VR^{-joint}): Finally, we combined the best model VR^{-pos} with the best positive evidence (baseline+VR$_{\text{BERT/RoBERTa}}$) with a simple multiplication.

Table 4 shows that there is some refinement results with the negative evidence over both baselines with VR^{-pos} and VR^{-joint}. However, there is no improvement over the original positive evidence.

9 Discussion: Limitations

Object Classifier Failure Cases: As the belief revision approach relies heavily on the object in the image for the likelihood revision, the quality of the object classifier is critical for the final decision. Here, we show some failure cases when the visual classifier struggle with complex background (*i.e.* wrong visual, object hallucination, *etc.*) as shown in Figure 5. Note that, if no related visual is present in the image the belief revision score will back off to 1 (no revision needed).

Object-to-Caption Similarity Score: Another limitation is the low/high cosine similarity score, which unbalances the likelihood revision. For example, a visual context *paddle* and the caption: *a man riding a surfboard on a wave* have low cosine scores when using a pre-trained model that is fine-tuned on sentence-to-sentence semantic similarity tasks (*i.e.* STS-B). Note that, we tackle this problem by adding multiple visual contexts as shown in Figure 2.

Related work

Modern sophisticated image captioning systems focus heavily on visual grounding to capture real-world scenarios. Early work Fang et al. (2015) builds a visual detector to guide and re-ranked image captioning with global similarity. The work of (Wang et al., 2018) investigates the informa-

Vil_BeamS: a pile of trash sitting inside of a building
Vil+VR_BERT: a pile of trash sitting in front of a building ✗
Human: an older floor light sits deserted in an abandoned hospital

Visual: vacuum ✗

Trans_BeamS: a kitchen with black counter tops and wooden cabinets
Tans+VR_BERT: a kitchen counter with a black counter top ✗
Human: a kitchen with a sink bottles jars and a dishwasher

Visual: barbershop ✗

Figure 5: Failure cases of the object detectors. The object classifier struggle with images with complex background and out-of-context object.

tiveness of visual or object information (*e.g.* object frequency count, size and position) in an end-to-end caption generation. Another work Cornia et al. (2019) proposes controlled caption language grounding through visual regions from the image. More recently, Gupta et al. (2020) introduce weakly supervised contrastive learning via object context and language modeling (*i.e.* BERT) for caption phrase grounding. Inspired by these works, (Fang et al., 2015) carried out re-ranking via visual information, (Wang et al., 2018; Cornia et al., 2019; Chen et al., 2020) explored the benefits of object information in image captioning, (Gupta et al., 2020) exploited the benefits of language modeling to extract contextualized word representations and the exploitation of the semantic coherency in caption language grounding (Zhang et al., 2021a), we purpose an object based re-ranker via human inspired logic reasoning with Belief Revision to re-rank the most closely related captions with contextualized semantic similarity.

Unlike the earlier approaches, our methods employ state-of-the-art tools and pre-trained models. Therefore, the system will keep improving in the future as better systems become available. In addition, our model can be directly used as a drop-in complement for any caption system that outputs a list of candidate hypothesis.

Conclusion

In this work, we aim to demonstrate that the Belief Revision approach that works well with human judgment can be applied to Image Captioning by employing human-inspired reasoning via a pre-trained model (*i.e.* GPT, BERT). Belief Revision (BR) is an approach for obtaining the likelihood re-

visions based on similarity scores via human judgment. We demonstrate the benefits of the approach by showing that two state-of-the-art Transformer-based image captioning results are improved via simple language grounding with visual context information. In particular, we show the accuracy gain in a benchmark dataset using two methods: (1) BR with positive visual evidence (increase the hypothesis) and (2) negative evidence (decrease the hypothesis), with wrong visual *i.e.* false positive by the classifier. However, this adaptation could be applied to many re-ranking tasks in NLP (text generation, multimodel MT, lexical selection, *etc.*), as well as in Computer Vision applications such as visual storytelling.

Ethical Considerations

The core contribution of our paper is algorithmic for the task of image captioning. As such we do not foresee any downstream harms propagated immediately by our proposal. We however acknowledge that due to the very nature of the data driven processing, there could be an amplification or propagation of potential biases existing in the datasets.

References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, pages 265–283.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*, pages 382–398. Springer.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Isabelle Augenstein, Mrinal Das, Sebastian Riedel, Lakshmi Vikraman, and Andrew McCallum. 2017. Semeval 2017 task 10: Scienceie-extracting keyphrases and relations from scientific publications. *arXiv preprint arXiv:1704.02853*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings*

of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, pages 65–72.

Sergey Blok, Douglas Medin, and Daniel Osherson. 2003. Probability from similarity. In *AAAI Spring Symposium on Logical Formalization of Commonsense Reasoning*.

Sergey V Blok, Douglas L Medin, and Daniel N Osherson. 2007. Induction as conditional probability judgment. *Memory & Cognition*, 35(6):1353–1364.

Keith Brown. 2005. *Encyclopedia of language and linguistics*, volume 1. Elsevier.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*.

Shizhe Chen, Qin Jin, Peng Wang, and Qi Wu. 2020. Say as you wish: Fine-grained control of image caption generation with abstract scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9962–9971.

Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2019. Show, control and tell: A framework for generating controllable and grounded captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8307–8316.

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10578–10587.

Bo Dai, Sanja Fidler, Raquel Urtasun, and Dahua Lin. 2017. Towards diverse and natural image descriptions via a conditional gan. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2970–2979.

Aditya Deshpande, Jyoti Aneja, Liwei Wang, Alexander G Schwing, and David Forsyth. 2019. Fast, diverse and accurate image captioning guided by part-of-speech. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10695–10704.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145.

Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. 2015. From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1473–1482.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Tanmay Gupta, Arash Vahdat, Gal Chechik, Xiaodong Yang, Jan Kautz, and Derek Hoiem. 2020. Contrastive learning for weakly supervised phrase grounding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 752–768. Springer.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, et al. 2017. Speed/accuracy trade-offs for modern convolutional object detectors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. 2019. Attention on attention for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4634–4643.

Daphne Ippolito, Reno Kriz, Maria Kustikova, João Sedoc, and Chris Callison-Burch. 2019. Comparison of diverse decoding methods from conditional language models. *arXiv preprint arXiv:1906.06362*.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. *arXiv preprint arXiv:2107.10821*.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 388–395.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 2020. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10437–10446.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in bleu: Reevaluating the evaluation of automatic machine translation evaluation metrics. *arXiv preprint arXiv:2006.06264*.

Philip M McCarthy and Scott Jarvis. 2010. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392.

Rui Meng, Sanqiang Zhao, Shuguang Han, Daqing He, Peter Brusilovsky, and Yu Chi. 2017. Deep keyphrase generation. *arXiv preprint arXiv:1704.06879*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matt Post. 2018. A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Zhou Ren, Xiaoyu Wang, Ning Zhang, Xutao Lv, and Li-Jia Li. 2017. Deep reinforcement learning-based image captioning with embedding reward. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 290–298.

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7008–7024.

Ahmed Sabir, Francesc Moreno-Noguer, and Lluís Padró. 2018. Visual re-ranking with natural language understanding for text spotting. In *Asian Conference on Computer Vision*, pages 68–82. Springer.

Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, and Bernt Schiele. 2017. Speaking the same language: Matching machine to human captions by adversarial training. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4135–4144.

Yurun Song, Junchen Zhao, and Lucia Specia. 2021. Sentsim: Crosslingual semantic evaluation of machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3143–3156.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. 2017. Context-aware captions from context-agnostic supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 251–260.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Ashwin Vijayakumar, Michael Cogswell, Ramprasaath Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2018. Diverse beam search for improved description of complex scenes. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

Jiuniu Wang, Wenjia Xu, Qingzhong Wang, and Antoni B Chan. 2020. Compare and reweight: Distinctive image captioning using similar images sets. In *European Conference on Computer Vision*, pages 370–386. Springer.

Josiah Wang, Pranava Madhyastha, and Lucia Specia. 2018. Object counts! bringing explicit detections back into image captioning. *arXiv preprint arXiv:1805.00314*.

Qingzhong Wang and Antoni B Chan. 2019. Describing like humans: on diversity in image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4195–4203.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057.

Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659.

Tianyi Zhang, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Wenqiao Zhang, Haochen Shi, Siliang Tang, Jun Xiao, Qiang Yu, and Yueting Zhuang. 2021a. Consensus graph representation learning for better grounded image captioning. In *Proc 35 AAAI Conf on Artificial Intelligence*.

Xuying Zhang, Xiaoshuai Sun, Yunpeng Luo, Jiayi Ji, Yiyi Zhou, Yongjian Wu, Feiyue Huang, and Rongrong Ji. 2021b. Rstnet: Captioning with adaptive attention on visual and non-visual words. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15465–15474.

A Appendix: Additional Result

SimProb with Two Visual Contexts. Until now, we experimented with one or multiple visual contexts at the same time; however, when we have more than one evidence supporting the revision, we can reason and choose which one to use to revise the hypothesis. In this scenario we begin with the conditional probability that comes from the dominant premise alone $P(Q_c|Q_a)$ (let's assume Q_a is dominant). Then we add a fraction of the remaining lack of trust or confidence $1 - P(Q_c|Q_a)$ that the dominant conditional *leaves behind*. The similarity between premise categories defines the size of the portion or fraction $sim(a, b)$ and separates the influence of the nondominant premise on the conclusion prior $P(Q_c|Q_b) - P(Q_c)$. The component of similarity is designed to reduce the impact of the non-dominant premise when the premises are redundant. Note that the proposed method in equation below guarantees an increase in strength with additional premises. In addition, this property, $\Pr(Qc|Qa, Qb) \geq \Pr(Qc|Qa), \Pr(Qc|Qb)$, is noncompetitive in the sense that one category does not reduce the probability of concept for another. Following the notation of Equation 1, we write the two visual contexts SimProb as:

$$\mathrm{P}(w \mid c_1, c_2) = \beta M + (1 - \beta)S, \text{ where}$$

$$\beta = \max \left\{ \begin{array}{l} sim(w, c_1) \\ sim(w, c_2) \\ sim(c_1, c_2) \\ 1.0 - sim(w, c_1) \\ 1.0 - sim(w, c_2) \\ \mathrm{P}(c_1) \\ \mathrm{P}(c_2) \end{array} \right\}$$

$$M = \max\{\mathrm{P}(w \mid c_1), \mathrm{P}(w \mid c_2)\}, S =$$
$$\mathrm{P}(w \mid c_1) + \mathrm{P}(w \mid c_2) - \mathrm{P}(w \mid c_1) \times \mathrm{P}(w \mid c_2)$$

where $\mathrm{P}(w \mid c_1)$ and $\mathrm{P}(w \mid c_2)$ are defined by Equation 1, and the two visual contexts are: (1) c_1 ResNet is the label **Class** and (2) the c_2 COCO **Object** categories are from Inception-ResNet v2 based Faster RCNN. Note that β takes the max of all models, and thus it is not breaking the formation if one of the similarities or probabilities is not confident enough (*i.e.* if it is below the threshold).

The last rows in Table 8 show the result of the SimProb selecting the best visual context of the two visuals. However, although there is some improvement over the other approaches (*i.e.* single and multi-visual contexts), it is not significant enough to justify the computational cost as a post-processing approach.

Figure 7 shows the benefit of employing common observation via Unigram LM in comparison to the classifier confident, (Left Figure) a denser SimProb score caption re-ranking.

Additional Statistical Significance Analysis. Table 5 shows the full results of the statistically significant test via pair bootstrap resampling (Koehn,

	BLEU (*p*-value)				NIST (*p*-value)			
Model	B1	B2	B3	B4	N1	N2	N3	N4
VilBERT								
Vil$_{BeamS}$	0.740	0.578	0.441	0.337	0.492	0.672	0.731	0.755
Vil+VR$_{w\text{-}Object}$ (Fang et al., 2015)	0.744 (0.019)	**0.581** (0.063)	0.442 (0.371)	0.335 (0.109)	**0.497** (0.00)	**0.680** (0.00)	**0.740** (0.001)	0.764 (0.00)
Vil+VR$_{Object}$ (Wang et al., 2018)	**0.745** (0.01)	**0.581** (0.091)	0.441 (0.429)	0.335 (0.131)	0.496 (0.003)	0.677 (0.015)	0.737 (0.012)	0.761 (0.043)
Vil+VR$_{Control}$ (Cornia et al., 2019)	0.741 (0.233)	0.577 (0.29)	0.439 (0.104)	0.331 (0.016)	0.494 (0.015)	0.676 (0.046)	0.735 (0.064)	0.758 (0.118)
Vil+VR$_{RoBERTa}$ (ours) (Table 1 Best)	0.741 (0.382)	0.580 (0.129)	**0.443** (0.00)	**0.339** (0.147)	**0.497** (0.00)	**0.680** (0.00)	0.740 (0.00)	**0.764** (0.002)
Transformer based caption generator								
Trans$_{BeamS}$	0.726	0.584	0.451	0.341	0.492	0.688	0.756	0.781
Trans+VR$_{w\text{-}Object}$ (Fang et al., 2015)	0.775 (0.00)	0.625 (0.00)	0.482 (0.00)	0.364 (0.00)	0.498 (0.00)	0.696 (0.00)	0.764 (0.001)	0.789 (0.001)
Trans+VR$_{Object}$ (Wang et al., 2018)	0.777 (0.00)	0.628 (0.00)	0.486 (0.00)	0.369 (0.00)	0.504 (0.00)	0.703 (0.00)	0.773 (0.00)	0.798 (0.00)
Trans+VR$_{Control}$ (Cornia et al., 2019)	0.780 (0.00)	0.623 (0.00)	**0.490** (0.00)	0.373 (0.00)	0.502 (0.00)	0.700 (0.00)	0.771 (0.00)	0.796 (0.00)
Trans+VR$_{BERT}$ (ours) (Table 1 Best)	**0.781** (0.00)	**0.631** (0.00)	**0.490** (0.00)	**0.374** (0.00)	**0.509** (0.00)	**0.710** (0.00)	**0.781** (0.00)	**0.806** (0.00)

Table 5: Result with pair bootstrapping resampling test via 1k trial (Koehn, 2004) on the significant improvement before and ranking with BLEU and NIST.

	SacreBLEU				
Model	Baseline Avg	New Avg	delta	Baseline better confidence %	New better confidence %
VilBERT (Lu et al., 2020)					
Vil$_{BeamS}$	9.10				
Vil+VR$_{W\text{-}Object}$ (Fang et al., 2015)		9.18	0.08	27.60	72.40
Vil+VR$_{Object}$ (Wang et al., 2018)		8.89	-0.22	93.50	6.50
Vil+VR$_{Control}$ (Cornia et al., 2019)		9.01	-0.09	75.60	24.40
Vil+VR$_{RoBERTa}$ (ours) (Table 1 Best)		**9.29**	0.18	8.50	**91.50**
Transformer Caption Generator (Cornia et al., 2020)					
Trans$_{BeamS}$	10.16				
Trans+VR$_{W\text{-}Object}$ (Fang et al., 2015)		10.01	-0.16	93.80	6.20
Trans+VR$_{Object}$ (Wang et al., 2018)		10.18	0.02	43.40	56.60
Trans+VR$_{Control}$ (Cornia et al., 2019)		10.09	-0.07	83.80	16.20
Trans+VR$_{BERT}$ (ours) (Table 1 Best)		**10.36**	0.19	1.10	**98.90**

Table 6: Result with pair bootstrapping resampling test via 1k trial (Koehn, 2004) on the significant improvement before and ranking with Sacrebleu (Post, 2018).

2004)[6] with BLEU and NIST. The NIST metric is an improved version of BELU that rewards infrequently used words by giving greater weighting to rarer words.

Also, we employ Sacrebleu[7] (Post, 2018) to investigate the statistically significant improvement, using delta, of our re-ranker with a BLEU score using the same approach as that above. Table 6 shows that our method performs better as **new better confidence** than the two baselines.

Additional Diversity Analysis. Table 7 shows part-of-speech tagging (POS) results before and after visual re-ranking. The proposed model VR yields a richer output in all POS tags in both baselines.

Full Experimental Results. Table 8 and Table 10 show the full results of our experiments, with the most common metrics used for image captioning, for positive and negative evidence, re-

spectively. Also, as we mentioned before, inspired by BERTscore and following (Song et al., 2021) we employ sentence-to-sentence semantic similarity score to compare candidate captions with human references with pre-trained Sentence-RoBERTa$_{LARGE}$ (Reimers and Gurevych, 2019) tuned for general STS task. Unlike BERTscore which aligns word-to-word similarites, SBERT-sts builds a semantic vector for the whole sentence, which can be used to compare candidate captions with human references.

Additional Ablation Study. Table 11 shows additional ablation study experiments with different information.

Training Dataset. As shown in Figure 6, we use two approaches to match and filter out not related visual context: 1) Threshold: to filter out the probabilities prediction when the visual classifier is not confident. 2) Semantic alignment: to match the most related caption to its environmental context. In more detail, we use cosine similarity with GloVe to match the visual with its context. Table 9 illustrates samples of the enriched human annotation, and caption dataset, with visual context informa-
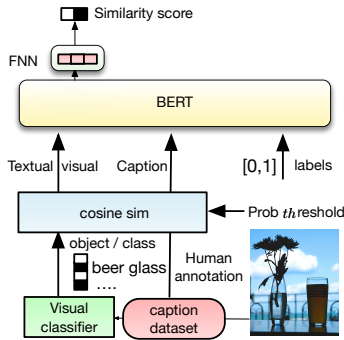
Figure 6: Dataset preprocessing until training. We use two methods to filter our non-related visual context (1) probability threshold: to filter out the visual context, and (2) semantic alignment with the caption via cosine distance (semantic relation).
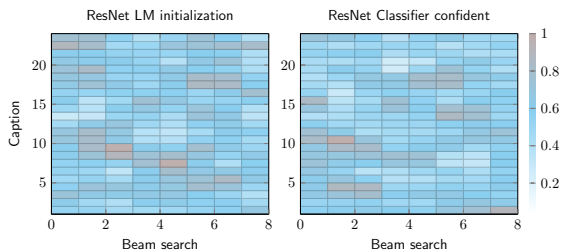

Figure 7: `SimProb` score of top-8 Beam search caption re-ranking (Right) with the visual classifier confidence probability without any initialization, (Left) with visual context that initialized by general common observation *i.e.* LM.

tion.

Model	Noun	Verb	Adj	Conj
VilBERT				
Vil$_{\text{BeamS}}$	14094	3586	3220	7914
Vil+VR$_{\text{RoBERTa}}$	**14403**	**3739**	**3325**	**8233**
Transformer				
Trans$_{\text{BeamS}}$	13961	3111	2004	7458
Trans+VR$_{\text{BERT}}$	**14203**	**3146**	**2056**	**7563**

Table 7: Most frequent POS tag before and after visual re-ranking. The result shows that after Visual Re-ranking both captions have more noun, verb, *etc*.

Why Positive Visual as Negative Evidence ? As we mentioned in the main script, we consider this as negative evidence for the following reasons: (1) the similarity is computed without the context of the sentence and (2) the static embedding (without knowing the sense of the word, *e.g.* bar for alcoholic drinking or rectangular solid piece of block). Although this approach relies on positive information, which is not the main intention of the negative evidence, the results demonstrate that there is a new direction of research that can be conducted using
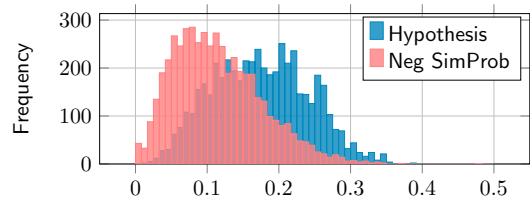
Figure 8: Visualization of Negative Evidence Neg-`SimProb` distribution on random samples from the test set (karpathy split). The negative visual information VR^{-pos} decreases the hypothesis that is initialized by LM-GPT-2.

positive information as negative evidence. Figure 8 shows that the original hypothesis is decreased with the positive information.
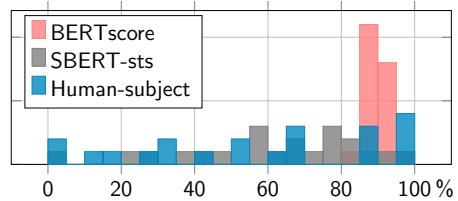
Figure 9: Comparison results between native human subject, BERTscore, and sentence level metric SBERT-sts on the test set.
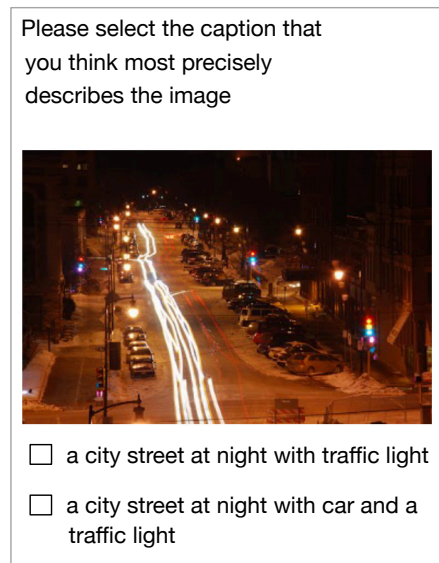
Figure 10: **Human Evaluation**. The user interface presented to our human subjects through the survey website asking them to re-rank the most descriptive caption candidates based on the visual information.

B   Human Evaluation

We conducted a human study to investigate human preferences over the visual re-ranked caption. We randomly selected 26 test images and gave 12

| Model	B-1	B-2	B-3	B-4	M	R	C	S	BERTscore	SBERT-sts
VilBERT (Lu et al., 2020)										
Vil$_{Greedy}$	0.751	0.587	0.441	0.330	0.272	0.554	1.104	0.207	0.9352	0.7550
Vil$_{BeamS}$	0.752	0.592	0.456	0.351	0.274	0.557	1.115	0.205	0.9363	0.7550
Vil+VR$_{W-Object}$ (Fang et al., 2015)	0.756	0.595	0.456	0.348	0.274	0.559	1.123	0.206	0.9365	**0.7605**
Vil+VR$_{Object}$ (Wang et al., 2018)	0.756	0.594	0.455	0.348	0.274	0.559	1.120	0.206	0.9364	0.7570
Vil+VR$_{Control}$ (Cornia et al., 2019)	0.753	0.591	0.453	0.345	0.274	0.557	1.116	0.206	0.9361	0.7565
Vil+GPT-2$_{mean}$ (only LM)	0.749	0.590	0.455	0.351	0.276	0.558	1.124	0.208	0.9364	0.7546
Vil+VR$_{BERT}$ (only sim)	0.753	0.591	0.452	0.343	0.273	0.556	1.112	0.206	0.9361	0.7562
Vil+VR$_{BERT}$	0.752	0.592	0.456	0.351	0.274	0.557	1.115	0.205	0.9365	0.7567
Vil+VR$_{BERT-Object}$	0.752	0.592	0.457	0.352	0.277	0.560	1.129	0.208	0.9365	0.7562
Vil+VR$_{RoBERTa}$	0.753	0.594	0.458	**0.353**	0.276	0.559	1.128	0.207	**0.9366**	0.7562
Vil+VR$_{RoBERTa-Object}$	**0.758**	**0.611**	**0.465**	0.344	0.262	0.555	1.234	0.206	0.9365	0.7554
Vil+VR$_{BERT-Multi-class}$	0.753	0.593	0.458	**0.353**	0.276	0.559	1.131	0.208	0.9365	0.7586
Vil+VR$_{BERT-Multi-object}$	0.752	0.592	0.456	0.351	0.276	0.558	1.123	0.208	0.9364	0.7566
Vil+VR$_{RoBERTa-Multi-class}$	0.751	0.591	0.456	0.351	**0.277**	**0.561**	**1.137**	0.208	**0.9366**	0.7589
Vil+VR$_{RoBERTa-Multi-object}$	0.753	0.593	0.458	**0.353**	0.276	0.559	1.131	0.208	0.9365	0.7586
Vil+VR$_{BERT-Class+object}$	0.752	0.592	0.455	0.350	0.276	0.559	1.126	**0.209**	0.9365	0.7563
Vil+VR$_{RoBERTa-class+object}$	0.752	0.592	0.457	0.352	**0.277**	0.559	1.127	0.208	0.9365	0.7558
Transformer Caption Generator (Cornia et al., 2020)										
Trans$_{Greedy}$	0.787	0.634	0.488	0.368	0.276	0.574	1.211	0.215	0.9376	0.7649
Trans$_{BeamS}$	0.793	0.645	0.504	0.387	0.281	0.582	1.247	**0.220**	**0.9399**	0.7707
Trans+VR$_{W-Object}$ (Fang et al., 2015)	0.786	0.638	0.497	0.378	0.277	0.579	1.228	0.216	0.9388	0.7709
Trans+VR$_{Object}$ (Wang et al., 2018)	0.790	0.642	0.501	0.383	0.280	0.580	1.237	0.219	0.9391	0.7700
Trans+VR$_{Control}$ (Cornia et al., 2019)	0.791	0.644	**0.505**	**0.388**	0.281	**0.583**	1.248	**0.220**	0.9398	0.7716
Trans+GPT-2$_{mean}$ (only LM)	0.791	0.643	0.503	0.386	0.281	0.582	1.242	0.219	0.9396	0.7714
Trans+VR$_{BERT}$ (only sim)	0.789	0.640	0.498	0.380	0.279	0.579	1.234	0.219	0.9389	0.7693
Trans+VR$_{BERT}$	0.793	**0.646**	**0.505**	**0.388**	**0.282**	**0.583**	1.250	**0.220**	**0.9399**	0.7711
Trans+VR$_{BERT-Object}$	0.793	0.644	0.503	0.385	0.281	0.581	1.242	0.219	0.9396	0.7695
Trans+VR$_{RoBERTa}$	0.792	0.644	0.504	0.386	0.280	0.582	1.244	0.219	0.9395	0.7705
Trans+VR$_{RoBERTa-Object}$	0.792	0.644	0.503	0.386	0.281	0.582	1.242	0.219	0.9396	0.7701
Trans+VR$_{BERT-Multi-class}$	**0.794**	0.645	0.503	0.385	0.281	0.582	1.248	**0.220**	0.9395	**0.7717**
Trans+VR$_{BERT-Multi-object}$	0.792	0.644	0.502	0.385	0.281	0.582	1.244	**0.220**	0.9395	0.7693
Trans+VR$_{RoBERTa-Multi-class}$	0.791	0.643	0.503	0.385	0.280	0.581	1.244	0.219	0.9395	0.7710
Trans+VR$_{RoBERTa-Multi-object}$	0.791	0.643	0.502	0.385	0.281	0.582	1.243	0.219	0.9395	0.7712
Trans+VR$_{RBERTa-Class+Object}$	0.793	0.645	0.504	0.387	0.281	0.582	1.247	0.220	0.9397	0.7705
Trans+VR$_{BERT-Class+Object}$	0.793	0.645	**0.505**	**0.388**	**0.282**	**0.583**	**1.251**	0.220	**0.9399**	0.7695

Table 8: **Positive Evidence: full result with all evaluation metrics**. Performance of compared baselines on the Karpathy test split with/without semantic re-ranking. For each base system, we report performance using a greedy search and the best beam search. Re-ranking is applied to the top-20 results of each system using BERT or RoBERTa for caption-context similarity. The visual contexts are extracted using ResNet152 and the Inception Resnet v2 based Faster R-CNN **object** detector. We also report results for Bert-based similarity without a hypothesis probability (rows marked *only sim*).

VC$_1$	VC$_2$	VC$_3$	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

Table 9: **Training Dataset**. The visual context (VC) is from a pre-trained visual classifier (*i.e.* ResNet152) and the caption is from COCO-Caption dataset (human-annotated).

reliable human subjects the option to choose between two captions: (1) Best-beam (BeamS) and (2) **V**isual **R**-ranker as shown in Figure 10.

Also, inspired by BERTscore and following (Song et al., 2021) that introduce Sentence Semantic Semantic (SSS) for machine translation, we employ a sentence-to-sentence semantic similarity score to compare candidate captions with human references. We use pre-trained Sentence-RoBERTa$_{LARGE}$ tuned for general STS-B task.

Consequently, the embedding will be more robust semantically than lexically for the STS tasks. Figure 9 shows that our results with SBERT-sts agrees more with human judgment than the BERTscore. Figure 11 and Figure 12 show some examples of when humans agree/disagree with our re-ranker.

Model	B-1	B-2	B-3	B-4	M	R	C	S	BERTscore	SBERT-sts
VilBERT (Lu et al., 2020)										
$\text{Vil}_{\text{Greedy}}$	0.751	0.587	0.441	0.330	0.272	0.554	1.104	0.207	0.9352	0.7550
$\text{Vil}_{\text{BeamS}}$	0.752	0.592	0.456	0.351	0.274	0.557	1.115	0.205	0.9363	0.7550
$\text{Vil+VR}_{\text{W-Object}}$ (Fang et al., 2015)	**0.756**	**0.595**	0.456	0.348	0.274	**0.559**	1.123	0.206	0.9365	**0.7605**
$\text{Vil+VR}_{\text{Object}}$ (Wang et al., 2018)	**0.756**	0.594	0.455	0.348	0.274	**0.559**	1.120	0.206	0.9364	0.7570
$\text{Vil+VR}_{\text{Control}}$ (Cornia et al., 2019)	0.753	0.591	0.453	0.345	0.274	0.557	1.116	0.206	0.9361	0.7565
$\text{Vil+VR}_{\text{RoBERTa}}$ Table 1 (positive)	0.753	0.594	**0.458**	**0.353**	**0.276**	**0.559**	1.128	0.207	**0.9366**	0.7562
$\text{Vil+VR}_{\text{RoBERTa}}^{-low}$	0.748	0.588	0.453	0.349	0.275	0.557	1.116	0.206	0.9362	0.7531
$\text{Vil+VR}_{\text{RoBERTa}}^{-high}$	0.748	0.588	0.453	0.349	0.275	0.557	1.116	0.206	0.9364	0.7546
$\text{Vil+VR}_{\text{GloVe}}^{-pos}$	0.751	0.591	0.455	0.351	0.276	0.558	1.123	0.207	0.9364	0.7556
$\text{Vil+VR}_{\text{RoBERTa+GloVe}}^{-joint}$	0.750	0.591	0.455	0.351	**0.276**	**0.559**	**1.126**	**0.208**	0.9365	0.7548
Transformer Caption Generator (Cornia et al., 2020)										
$\text{Trans}_{\text{Greedy}}$	0.787	0.634	0.488	0.368	0.276	0.574	1.211	0.215	0.9376	0.7649
$\text{Trans}_{\text{BeamS}}$	0.793	0.645	0.504	0.387	0.281	0.582	1.247	**0.220**	**0.9399**	0.7707
$\text{Trans+VR}_{\text{W-Object}}$ (Fang et al., 2015)	0.786	0.638	0.497	0.378	0.277	0.579	1.228	0.216	0.9388	0.7709
$\text{Trans+VR}_{\text{Object}}$ (Wang et al., 2018)	0.790	0.642	0.501	0.383	0.280	0.580	1.237	0.219	0.9391	0.7700
$\text{Trans+VR}_{\text{Control}}$ (Cornia et al., 2019)	0.791	0.644	0.505	**0.388**	0.281	**0.583**	1.248	**0.220**	0.9398	**0.7716**
$\text{Trans+VR}_{\text{BERT}}$ Table 1 (positive)	0.793	**0.646**	**0.505**	**0.388**	0.282	**0.583**	**1.250**	**0.220**	**0.9399**	0.7711
$\text{Trans+VR}_{\text{BERT}}^{-low}$	0.791	0.643	0.504	0.387	0.280	0.582	1.242	0.218	0.9396	0.7682
$\text{Trans+VR}_{\text{BERT}}^{-high}$	0.793	0.644	0.503	0.385	0.282	0.582	1.243	0.219	0.9397	0.7686
$\text{Trans+VR}_{\text{GloVe}}^{-pos}$	**0.794**	**0.646**	**0.506**	**0.388**	0.282	**0.583**	1.249	**0.220**	**0.9399**	0.7702
$\text{Trans+VR}_{\text{BERT+GloVe}}^{-joint}$	0.793	0.645	0.504	0.387	0.281	0.582	1.247	**0.220**	0.9398	0.7704

Table 10: **Negative Evidence: full result with all evaluation metrics**. Comparison results between positive Belief Revision (single concept VR) (gray color) and Negative Belief Revision (NBR) on the Karpathy test split. The NBR uses a *high similarity* VR^{-high} object related to the positive visual but not in the image, *low similarity* VR^{-low} uses false positive from the visual classifier, and positive visual via static word level similarity VR^{-pos}. **Boldface** fonts reflect the improvement over the baseline.

Model	B-1	B-2	B-3	B-4	M	R	C	S	BERTscore	SBERT-sts
VilBERT-VR-GPT-2$_{\text{mean}}$ + ResNet										
+ RoBERTa (Table 1 Best)	**0.753**	**0.594**	**0.458**	**0.353**	0.276	0.559	1.128	0.207	0.9366	0.7562
+ LM-GPT-2$_{\text{product}}$	0.749	0.590	0.455	0.351	0.276	0.558	1.124	0.208	0.9364	0.7486
+ DistilSBERT	0.751	0.591	0.456	0.352	**0.277**	0.559	1.130	0.209	0.9365	0.7567
+ SimCSE-BERT (Gao et al., 2021)	0.752	0.593	0.457	0.352	0.276	0.559	1.130	**0.209**	0.9365	0.7558
+ SimCSE-RoBERTa	0.750	0.590	0.455	0.351	0.276	0.558	1.125	0.208	0.9365	0.7549
+ SimCSE-BERT-V$_1$ (unsupervised)	0.750	0.591	0.455	0.351	0.276	0.558	1.128	0.207	0.9365	0.7560
+ SimCSE-BERT-V$_2$ (unsupervised)	0.752	0.593	0.457	**0.353**	**0.277**	0.559	1.132	0.208	0.9365	0.7560
+ CLIP-V (Radford et al., 2021)	**0.753**	**0.594**	**0.458**	**0.353**	0.276	**0.561**	1.131	0.208	**0.9367**	**0.7579**
Trans - VR-GPT-2$_{\text{mean}}$ + ResNet										
+ BERT (Table 1 Best)	0.793	**0.646**	**0.505**	**0.388**	**0.282**	**0.583**	**1.250**	**0.220**	**0.9399**	**0.7711**
+ LM-GPT-2$_{\text{product}}$	0.787	0.642	0.503	0.386	0.279	0.581	1.236	0.219	0.9398	0.7683
+ DistilSBERT	**0.794**	**0.646**	**0.505**	0.387	**0.282**	**0.583**	1.247	**0.220**	0.9396	0.7704
+ SimCSE-BERT (Gao et al., 2021)	0.792	0.644	0.503	0.386	0.281	0.581	1.243	0.219	0.9394	0.7694
+ SimCSE-RoBERTa	**0.794**	0.645	0.504	0.387	0.281	0.582	1.244	0.219	0.9395	0.7698
+ SimCSE-V$_1$ (unsupervised)	0.792	0.645	0.504	0.386	0.281	0.582	1.244	0.219	0.9397	0.7705
+ SimCSE-V$_2$ (unsupervised)	0.792	0.645	**0.505**	0.387	0.281	0.582	1.247	0.219	0.9396	0.7703
+ CLIP (Radford et al., 2021)	0.791	0.643	0.503	0.386	0.280	0.581	1.242	0.219	0.9395	0.7703

Table 11: **Full Ablation Study**. We experimented using different information from various baselines on the Karpathy test split. Also, with the unsupervised BERT similarity, we tried with the top-2 visual context from the classifier.

C Hyperparameters and Setting

All training and the beam search are implemented with PyTorch 1.7.1 (Paszke et al., 2019). VR based $\text{BERT}_{\text{base}}$ is fine-tuned on the training dataset using the original BERT implementation, Tensorflow version 1.15 with Cuda 8 (Abadi et al., 2016) (hardware: GPU GTX 1070Ti and 32 RAM and 8-cores i7 CPU). The textual information dataset consists of 460k captions, 373k for training, and 87k for validation *i.e.* visual, caption, label ([semantically related or not related]). We use a batch size of 16 for two epochs with a learning rate $2e{-}5$, we kept the rest of hyperparameters settings as the original implementation.

D Additional Examples

We provide more comparison results with examples, including sentence-level evaluation SBERT-sts and BERTscore in Table 12. Also, in Figure 11, and Figure 12, we also evaluated our re-ranker using human subjects.

Table 12: Examples show caption re-ranked by our **V**isual **R**e-ranker with different evaluation metrics including the semantic-similarity based metrics. Note that, we only report B-1 and B-2, from BLEU, to measure the word level changes before and after re-ranking.

Model	caption	B1	B2	M	R-L	S	BERTscore	SBERT-sts
BeamS	a woman holding a tennis racquet on a tennis court	**0.54**	**0.40**	0.26	0.47	0	0.89	**0.73**
VR	a woman **standing** on a tennis court holding a racque	0.53	0.32	**0.27**	**0.49**	**0.30**	**0.93**	0.68
Refe	a woman in a short bisque skirt holding a tennis racque							
BeamS	a white train is at a train station	0.29	0.18	0.11	0.32	0.22	0.90	0.57
VR	**a train on the tracks** at a train station	**0.49**	**0.40**	**0.23**	**0.52**	**0.53**	**0.91**	**0.78**
Refe	a train that is sitting on the tracks under wires							
BeamS	a pair of black scissors on a white wall ✗	0.31	0	0.12	0.37	0	0.87	0.23
VR	a flower in a vase next to a pair of scissors ✗	**0.41**	**0.19**	**0.15**	**0.43**	0.42	**0.89**	**0.27**
Refe	a dried black flower in a long tall black and white vase							
BeamS	a man sitting on a bench	0.43	0.43	0.32	0.67	0.50	0.96	0.66
VR	a man sitting on a bench **talking on a cell phone**	**0.90**	**0.85**	**0.95**	**0.90**	**0.90**	**0.98**	**0.99**
Refe	a man sitting on a bench talking on his cell phone							
BeamS	a woman standing in an airport with luggage	0.42	0.35	0.25	0.51	0.30	**0.92**	0.67
VR	a woman standing in a luggage carousel at an airport ✗	0.54	0.40	0.25	0.56	0.50	0.89	0.68
Refe	a woman standing in front of a bench covered in luggage							
BeamS	an airplane sitting on a runway behind a fence	**0.39**	**0.29**	0.22	0.41	0.40	0.92	0.69
VR	an airplane is parked behind a fence	0.37	0.28	**0.25**	**0.45**	**0.50**	**0.94**	**0.85**
Refe	the airplane has landed behind a fence with barbed wire							
BeamS	a plate with a sandwich on a table	0.55	0.26	0.19	**0.46**	0.40	0.91	0.86
VR	**a white plate** with a sandwich on a table	**0.66**	**0.40**	**0.24**	0.44	**0.54**	**0.92**	**0.90**
Refe	a small sandwich sitting on a white china plate							
BeamS	three giraffes standing in a field under a tree	**0.26**	0	**0.14**	**0.29**	**0.26**	**0.91**	0.62
VR	a group of giraffes standing in a field	0.17	0	0.10	0.20	0	0.90	**0.64**
Refe	two tall giraffe standing next to a green leaf filled tree							
BeamS	two parking meters in front of a brick wall	**0.33**	0.20	0.15	0.22	0.25	0.47	0.87
VR	a row of parking meters in front of a building	0.30	**0.25**	**0.17**	**0.31**	0.25	**0.60**	**0.88**
Refe	different types and sizes of parking meters on display							
BeamS	a bathroom with a toilet and a mirror	0.33	0	0.10	**0.34**	0	0.89	0.69
VR	a variety of items on display in a bathroom	**0.44**	0	**0.12**	0.33	**0.14**	**0.90**	**0.70**
Refe	a view of a couple types of toilet items							
BeamS	two bulls with horns standing next to each other	**0.44**	**0.23**	0.16	**0.47**	**0.66**	**0.89**	0.76
VR	two long horn bulls standing next to each other	0.33	0	**0.18**	0.23	0.25	0.88	**0.81**
Refe	closeup of two red-haired bulls with long horns							
BeamS	a laptop computer sitting on top of a desk	0.44	0.21	0.20	0.29	0.18	0.91	0.69
VR	a desk with a laptop **and a computer monitor**	**0.53**	**0.51**	**0.31**	**0.58**	**0.46**	**0.95**	**0.77**
Refe	an office desk with a laptop and a phone on it							
BeamS	a busy highway with cars and a train	**0.33**	**0.20**	**0.12**	**0.34**	**0.18**	**0.90**	**0.43**
VR	cars are driving on a highway under a bridge ✗	0.22	0	0.04	0.22	0	0.88	0.21
Refe	a photo of a train heading down the tracks							
BeamS	a baby sitting in front of a cake	0.44	0.23	0.18	**0.46**	**0.36**	0.90	**0.81**
VR	a baby sitting in front of a **birthday cake**	0.44	0.23	0.18	0.44	0.33	0.90	0.77
Refe	a baby in high chair with bib and cake							
BeamS	a dog sitting on a bed with clothes	**0.58**	**0.44**	**0.29**	**0.65**	0.40	**0.93**	0.57
VR	a dog sitting on a bed **next to** clothes	0.49	0.33	0.25	0.52	0.40	0.91	**0.61**
Refe	a dog is sitting on an unmade bed with pillows							
BeamS	a group of boats docked in the water	0.19	0	0.07	**0.21**	0.22	0.88	0.58
VR	a group of boats **are** docked in the water	0.19	0	0.07	0.20	0.22	0.88	**0.59**
Refe	looking out over a bay with many tourist boats moored							

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 couple of airplanes that are flying in the sky	0.88	0.03	0	traffic light
VR	a group of traffic lights at an airport	**0.95**	**0.71**	**100**	
Human refe	a group of traffic lights sitting above an intersection				
BeamS	two bulls with horns standing next to each other	**0.89**	0.76	16.7	ox
VR	two long horn bulls standing next to each other	0.88	**0.81**	**83.3**	
Human refe	closeup of two red-haired bulls with long horns				
BeamS	a woman holding a tennis racquet on a tennis court	0.89	**0.73**	**85.3**	racket
VR	a woman standing on a tennis court holding a racquet	**0.93**	0.68	16.7	
Human refe	a woman in a short bisque skirt holding a tennis racquet				
BeamS	a white train is at a train station	0.90	0.57	50	⚡ locomotive
VR	a train on the tracks at a train station	**0.91**	**0.78**	50	
Human refe	a train that is sitting on the tracks under wires				
BeamS	two men cutting a cake at ceremony	**0.94**	**0.98**	**66.7**	≈ mortarboard
VR	a group of military men cutting a cake	0.80	0.91	33.3	
Human refe	two men are cutting a cake at a function				
BeamS	a little girl wearing a tie and pants	**0.94**	**0.80**	**66.7**	≈ feather boa
VR	a little girl wearing a tie standing in a room	0.93	0.77	33.3	
Human refe	a young girl wearing a tie that matches her skirt				
BeamS	a man laying on the ground with many animals	0.88	0.14	0	✗ trilobite
VR	a man kneeling down in front of a herd of sheep	**0.89**	**0.56**	**100**	
Human refe	a view of a bunch of sheep lined up with a behind them				
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				

Figure 11: Examples show caption re-ranked by our **V**isual **R**e-ranker and the original baseline Best beam. An evaluation metrics comparison between semantic-similarity based SBERT-sts, BERTscore, and the human subject.

Model	Caption	BERTscore	SBERT-sts	Human%	Visual
BeamS	a red and white boat in the water	**0.94**	**0.78**	**87.5**	fireboat
VR	a red and white boat is in the water	0.93	0.79	12.5	
Human refe	a red and white boat floating along a river				
BeamS	a dog sitting on a bed with cloths	**0.93**	0.57	50	dog/Irish terrier
VR	a dog sitting on a bed next to clothes	0.91	**0.61**	50	
Human refe	a dog is sitting on an unmade bed with pillows				
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				
BeamS	a bathroom with a toilet and a mirror	0.89	0.69	12.5	washbasin
VR	a variety of item on display in a bathroom	**0.90**	**0.70**	**83.3**	
Human refe	a view of a couple types of toilet items				
BeamS	a couple of pizzas that are on a table	**0.88**	0.75	**100**	pizza
VR	a couple of pizzas are sitting on a table	0.87	**0.77**	0	
Human refe	pizza on a table with a cup and a fork				
BeamS	a group of people in a living room playing a video game	0.93	0.47	**62.5**	television
VR	a group of people sitting in a living room playing a video game	**0.94**	**0.50**	37.5	
Human refe	a group of friends sitting inside their living room				
BeamS	a city street at night with traffic lights	**0.97**	0.81	33.3	traffic light
VR	a city street at night with cars and a traffic light	0.94	**0.85**	66.7	
Human refe	a city street at night filled with lots of traffic				
BeamS	a close up of a cat eating a doughnut	**0.83**	**0.90**	**100**	pretzel
VR	a close up of a person holding a doughnut	0.60	0.88	0	
Human refe	a cat bites into a doughnut offered by a persons hand				
BeamS	two men standing next to a group of people	**0.88**	0.20	**87.5**	groom
VR	a group of men standing next to each other	0.86	**0.30**	12.5	
Human refe	the man is holding his tie with his right hand				
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				

Figure 12: Examples show caption re-ranked by our **V**isual **R**e-ranker and the original baseline Best beam. An evaluation metrics comparison between semantic-similarity based SBERT-sts, BERTscore, and the human subject.