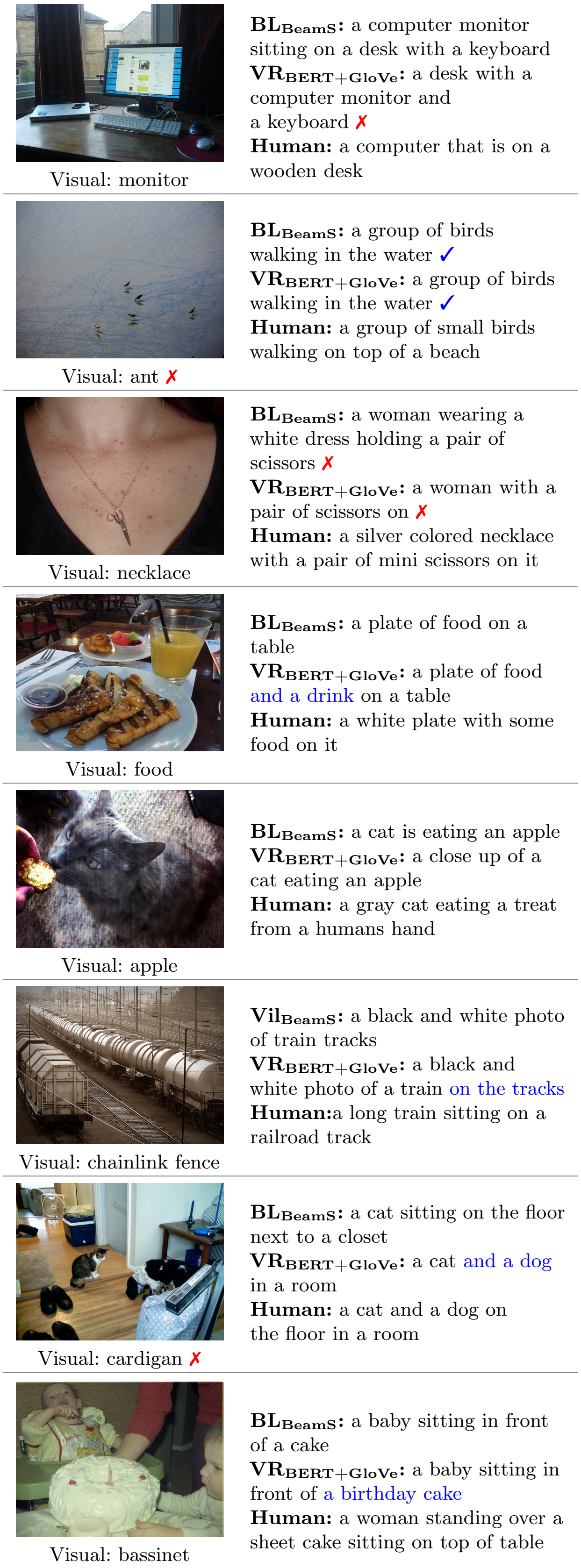
Supplementary Material: Word to Sentence Visual Semantic Similarity for Caption Generation: Lessons Learned

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



$\mathbf{BL}_{\mathbf{BeamS}}$**:** a computer monitor sitting on a desk with a keyboard
$\mathbf{VR}_{\mathbf{BERT+GloVe}}$**:** a desk with a computer monitor and a keyboard ✗
Human: a computer that is on a wooden desk
Visual: monitor

$\mathbf{BL}_{\mathbf{BeamS}}$**:** a group of birds walking in the water ✓
$\mathbf{VR}_{\mathbf{BERT+GloVe}}$**:** a group of birds walking in the water ✓
Human: a group of small birds walking on top of a beach
Visual: ant ✗

$\mathbf{BL}_{\mathbf{BeamS}}$**:** a woman wearing a white dress holding a pair of scissors ✗
$\mathbf{VR}_{\mathbf{BERT+GloVe}}$**:** a woman with a pair of scissors on ✗
Human: a silver colored necklace with a pair of mini scissors on it
Visual: necklace

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

$\mathbf{BL}_{\mathbf{BeamS}}$**:** a cat is eating an apple
$\mathbf{VR}_{\mathbf{BERT+GloVe}}$**:** a close up of a cat eating an apple
Human: a gray cat eating a treat from a humans hand
Visual: apple

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

$\mathbf{BL}_{\mathbf{BeamS}}$**:** a cat sitting on the floor next to a closet
$\mathbf{VR}_{\mathbf{BERT+GloVe}}$**:** a cat and a dog in a room
Human: a cat and a dog on the floor in a room
Visual: cardigan ✗

$\mathbf{BL}_{\mathbf{BeamS}}$**:** a baby sitting in front of a cake
$\mathbf{VR}_{\mathbf{BERT+GloVe}}$**:** a baby sitting in front of a birthday cake
Human: a woman standing over a sheet cake sitting on top of table
Visual: bassinet

Figure 1. Examples of the re-ranked captions by our visual re-ranker (VR) and the original caption (**Beam Search**) by the baseline (BL).

1 Hyperparameters and Setting

All training and the beam search are implemented in `fairseq` [6] and trained with PyTorch 1.7.1 [7] on a single K-80 GPU.

Visual Re-ranker. The only model we fine-tuned is the $\text{BERT}_{\text{base}}$ model. We fine-tuned it on the training dataset using the original BERT implementation, Tensorflow version 1.15 with Cuda 8 [1]. The textual dataset contains around 460k captions: 373k for training and 87k for validation *i.e.* visual, caption, label [semantically related or not related]). We use batch size 16 for two/three epochs with a learning rate 2e−5 and we kept the rest of hyperparameters settings as the original implementation. Note that we keep the GloVe as a static model as the model is trained on 840 billion tokens.

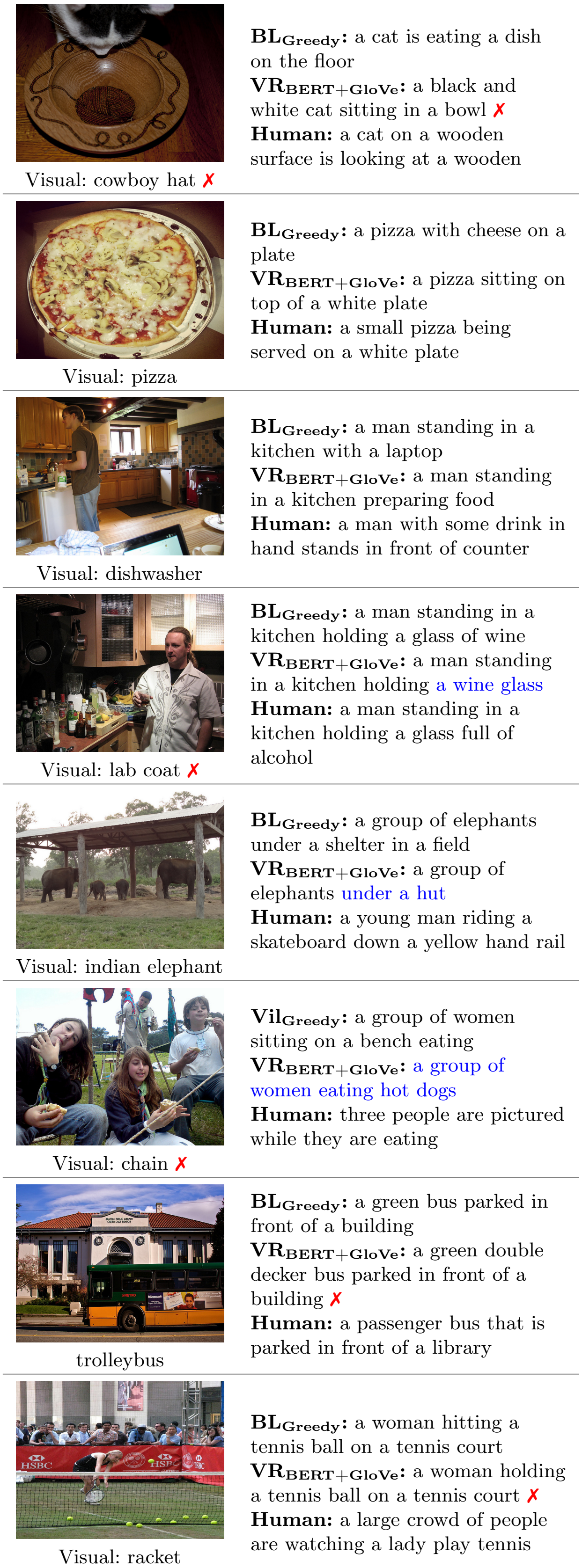
Show-and-Tell [8]. We train this shallow model from scratch on the flickr8k [4] dataset (6270 train/1730 test).

Caption Transformer [3][1]. We train the transformer from scratch with the Bottom-Up features [2]. However, unlike the original implementation by the authors, we use a full 12-layer transformer. We follow the same hyperparameters as the original implementation.

VilBERT [5]. Since VilBERT is trained on 12 datasets, we use it as an out-of-the-box model.

2 Examples of Re-ranked Captions

Best Beam. In Figure 1 we show examples of the proposed re-ranker and comparison results with the best baseline beam search ($\mathbf{BL}_{\mathbf{BeamS}}$). The model struggles to unify the information from diffident modalities, and therefore the word-level expert has a stronger influence on the final score. In addition, the visual classifier also faces difficulties with complex background images. This could be resolved in future work, by employing multiple



Visual: cowboy hat ✗

BL$_{\textbf{Greedy}}$: a cat is eating a dish on the floor
VR$_{\textbf{BERT+GloVe}}$: a black and white cat sitting in a bowl ✗
Human: a cat on a wooden surface is looking at a wooden

Visual: pizza

BL$_{\textbf{Greedy}}$: a pizza with cheese on a plate
VR$_{\textbf{BERT+GloVe}}$: a pizza sitting on top of a white plate
Human: a small pizza being served on a white plate

Visual: dishwasher

BL$_{\textbf{Greedy}}$: a man standing in a kitchen with a laptop
VR$_{\textbf{BERT+GloVe}}$: a man standing in a kitchen preparing food
Human: a man with some drink in hand stands in front of counter

Visual: lab coat ✗

BL$_{\textbf{Greedy}}$: a man standing in a kitchen holding a glass of wine
VR$_{\textbf{BERT+GloVe}}$: a man standing in a kitchen holding a wine glass
Human: a man standing in a kitchen holding a glass full of alcohol

Visual: indian elephant

BL$_{\textbf{Greedy}}$: a group of elephants under a shelter in a field
VR$_{\textbf{BERT+GloVe}}$: a group of elephants under a hut
Human: a young man riding a skateboard down a yellow hand rail

Visual: chain ✗

Vil$_{\textbf{Greedy}}$: a group of women sitting on a bench eating
VR$_{\textbf{BERT+GloVe}}$: a group of women eating hot dogs
Human: three people are pictured while they are eating

trolleybus

BL$_{\textbf{Greedy}}$: a green bus parked in front of a building
VR$_{\textbf{BERT+GloVe}}$: a green double decker bus parked in front of a building ✗
Human: a passenger bus that is parked in front of a library

Visual: racket

BL$_{\textbf{Greedy}}$: a woman hitting a tennis ball on a tennis court
VR$_{\textbf{BERT+GloVe}}$: a woman holding a tennis ball on a tennis court ✗
Human: a large crowd of people are watching a lady play tennis

Figure 2. Examples of the re-ranked captions by our visual re-ranker (VR) and the original caption (**greedy**) by the baseline (BL).

classifiers (each with multiple labels) and then using a voting technique to filter out the most probable object in the image.

Greedy. We also experiment with k-1 greedy output (**BL$_{\textbf{Greedy}}$**) as shown in Figure 2, our model suffers from the same limitation.

[1] https://github.com/aimagelab/meshed-memory-transformer

References

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

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018.

[3] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *CVPR*, 2020.

[4] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *JAIR*, 2013.

[5] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *CVPR*, 2020.

[6] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In *NAACL*, 2019.

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

[8] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015.