Appendix

A Examples of Hard Negatives

Fig. A.1 compares the outputs of VSE++ and VSE0 for a few examples.

GT: Two elephants are standing by the trees in the wild.
VSE0: [9] Three elephants kick up dust as they walk through the flat by the bushes.

GT: A large multi layered cake with candles sticking out of it.
VSE0: [1] A party decoration containing flowers, flags, and candles.

GT: The man is walking down the street with no shirt on.
VSE0: [24] A person standing on a skate board in an alley.
VSE++: [10] Two young men are skateboarding on the street.

GT: A row of motorcycles parked in front of a building.
VSE0: [2] a parking area for motorcycles and bicycles along a street
VSE++: [1] A number of motorbikes parked on an alley

GT: some skateboarders doing tricks and people watching them
VSE0: [39] Young skateboarder displaying skills on sidewalk near field.
VSE++: [3] Two young men are outside skateboarding together.

GT: a brown cake with white icing and some walnut toppings
VSE0: [6] A large slice of angel food cake sitting on top of a plate.
VSE++: [16] A baked loaf of bread is shown still in the pan.

GT: A woman holding a child and standing near a bull.
VSE0: [1] A woman holding a child and standing near a bull.
VSE++: [1] A woman holding a child looking at a cow.

GT: A woman in a short pink skirt holding a tennis racket.
VSE0: [6] A man playing tennis and holding back his racket to hit the ball.
VSE++: [1] A woman is standing while holding a tennis racket.

Figure A.1: Examples of MS-COCO test images and the top 1 retrieved captions for VSE0 and VSE++ (ResNet)-finetune. The value in brackets is the rank of the highest ranked ground-truth caption. GT is a sample from the ground-truth captions.