Supplementary Material for Neural Sequential Phrase Grounding (SeqGROUND)

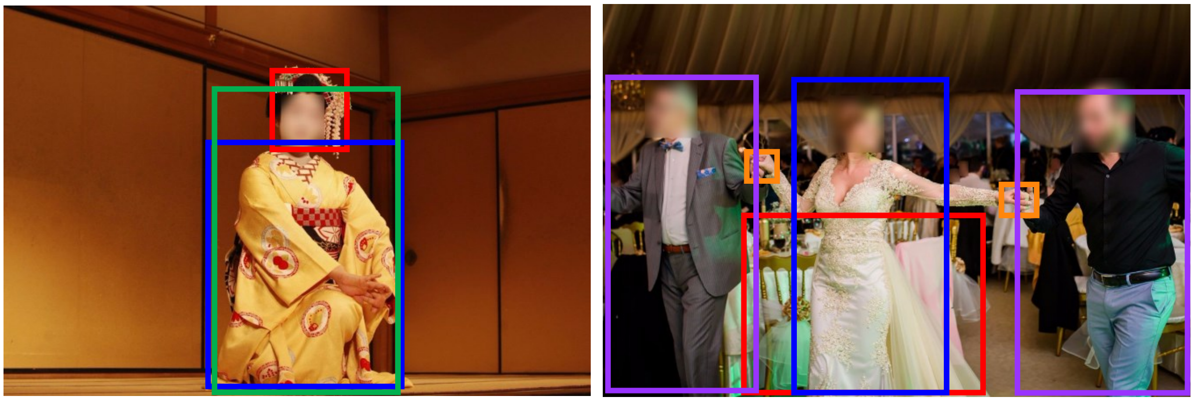
Pelin Dogan[1] Leonid Sigal[2,3] Markus Gross[1,4]
[1]ETH Zürich [2]University of British Columbia [3]Vector Institute [4]Disney Research
{pelin.dogan, grossm}@inf.ethz.ch, lsigal@cs.ubc.ca

In this supplementary material, we provide additional details on implementation of our framework, such as parameters, dimensions, etc. Furthermore, we provide more phrase grounding results computed by our approach that require many-to-many matching of phrases and bounding boxes.

1. Additional Implementation Details

The phrase and the visual encoder represent the input phrases and regions (boxes) in a 500-dimensional joint embedding space. The phrase encoder, which is shown by *3xFC* in green in Figure 2 of the main paper, has one dropout layer with a rate of 0.4 that is followed by three fully connected layers with ReLu activations, and a normalization layer. These fully connected layers have output dimensionality of 1500, 1000, and 500 respectively. The visual encoder, which is shown by *3xFC* in pink in Figure 2 of the main paper has the same layout as well. The batch size is set to 32, which provides 31 contrastive samples for every positive pair. The pre-training step is performed with the Adam optimizer using a learning rate of 10^{-4}, and a gradient clipping threshold of 2.0.

The *phrase stack*and the *history stack* are implemented with a two-layer LSTM network with an output space dimensionality of 500. Each layer of the bidirectional LSTM network of the *box stack* has the same output dimensionality, too. Between each layer in the stacks, there is a dropout layer with a rate of 0.25. Another dropput layer with the rate of 0.4 is placed after concatenating the hidden states of the stacks, which is before the fully connected layers. The last fully connected layer that outputs the decision prediction has a sigmoid activation, in order to train the model with binary cross-entropy loss. The batch size is set to 10 image-sentence pairs. The full model is trained with the Adam optimizer using a learning rate of 10^{-3} and 10^{-4}, respectively for the two stages of training mentioned in Section 4.1 of the main paper. At test time, our method performs in 1.39 seconds, on average, for an image-sentence pair. Training (including pre-training and the stages in Sec. 4.1) is close to 4 days, using one NVIDIA Titan X.



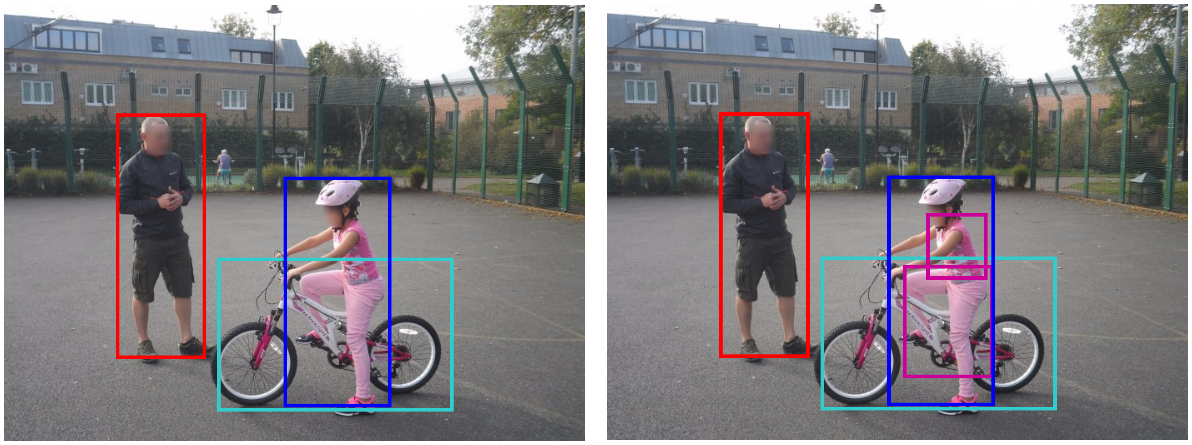
(a) A Japanese woman poses in a ceremonial clothing with an elaborate headpiece.

(b) Two men are dancing and holding hands with a bride that has an elegant wedding dress.

Figure 1: Examples of succesful results.

2. Additional Results

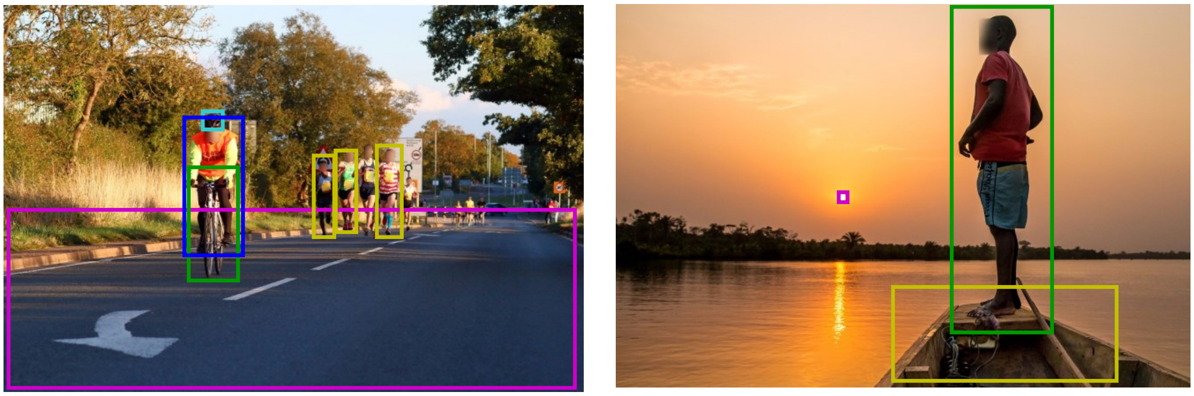
As mentioned in the main paper, we are not allowed to show images from the Flickr30K Entities Dataset due to copyright issues. However, we created similar content with images that are dedicated to public domain. For this purpose, we took sentences from the dataset and found similar images. We did minor modifications in the sentences when required, due to minor differences in the content of the new images. Figure 1- 3, shows various results computed by SeqGROUND, including some failure cases. It is important to note that given an image and its sentence, SeqGROUND decides by itself whether to ground a noun phrase, or not. In other words, it is not forced to ground any noun phrase in the examples shown in this supplementary material. Specifically, Figure 2 shows such as example case.



(a) A girl in *black clothing* is riding a bike while her father is watching her.

(b) A girl in pink clothing is riding a bike while her father is watching her.

Figure 2: Example results which are showing the efficacy of SeqGROUND. (a) Inaccurate phrase *a black clothing* is successfully ignored by SeqGROUND. (b) Succesful grounding for an accurate description.



(a) A man with a helmet is riding a bike in front of a group of running men on the road.

(b) A man is standing on a boat as the sun sets.

Figure 3: Examples of succesful results.

Acknowledgement. Pelin Dogan was partially funded by the SNF grant 200021 153307/1.