# A Supplementary Material

## A.1 Step-wise self-attention on the sample image

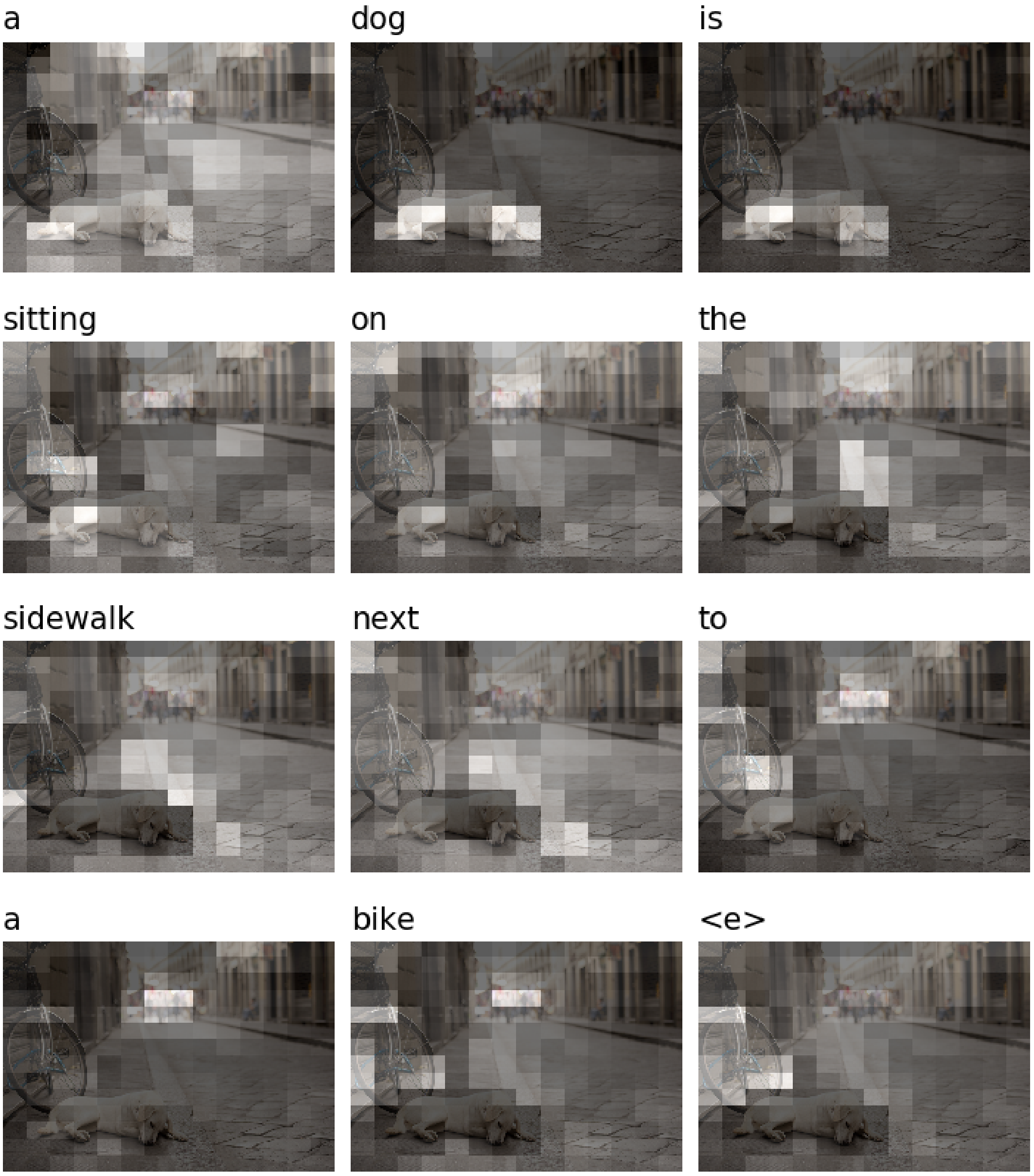
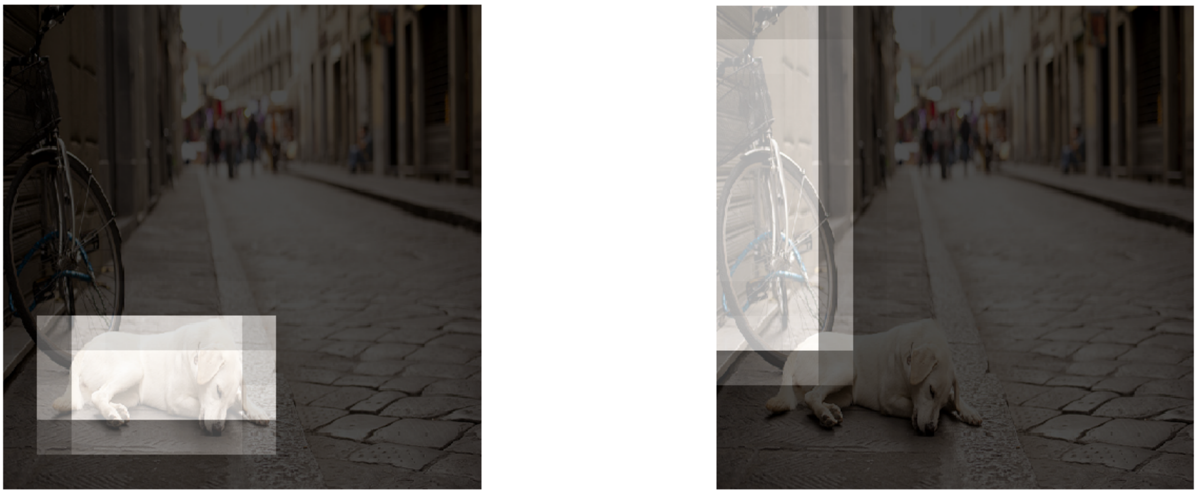


Figure 1: The step-wise alternating self-attention for the sample image with the dog and the bike. The produced caption is shown from the upper-left corner to the bottom-right corner with one word and its attention at a time-step. The models attention is shown in pixels that represent the spatial area of the visual features. The model is clearly putting higher weights on the dog, when producing the word at the second time-step. Furthermore, spatial attention is around the dog, when the resulting word is ”sidewalk”. When the ”bike” word is produced, then also some weights are put on the persons in the background and only with the end-tag the attention is clearly on the bike.

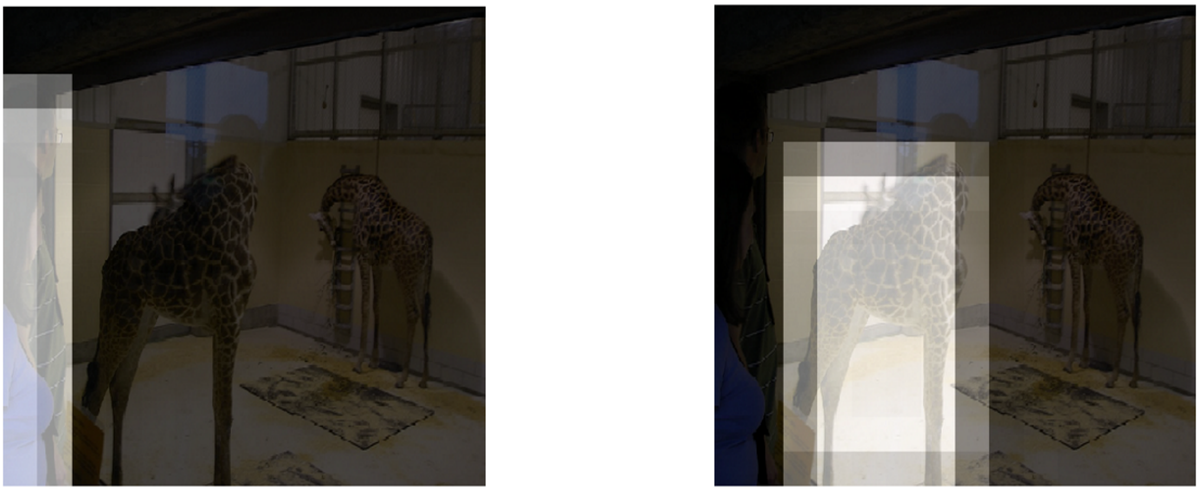
## A.2 Unlimited step-wise fixed attention on the sample image



(a) Caption: ”a dog is laying down on the street”

(b) Caption: ”a bicycle parked in front of a bicycle”

Figure 2: The box captions for the sample image with fixed attention on the dog in (a) and the bicycle in (b) The model is producing captions that are clearly focusing on either the dog in (a) or the bicycle in (b).

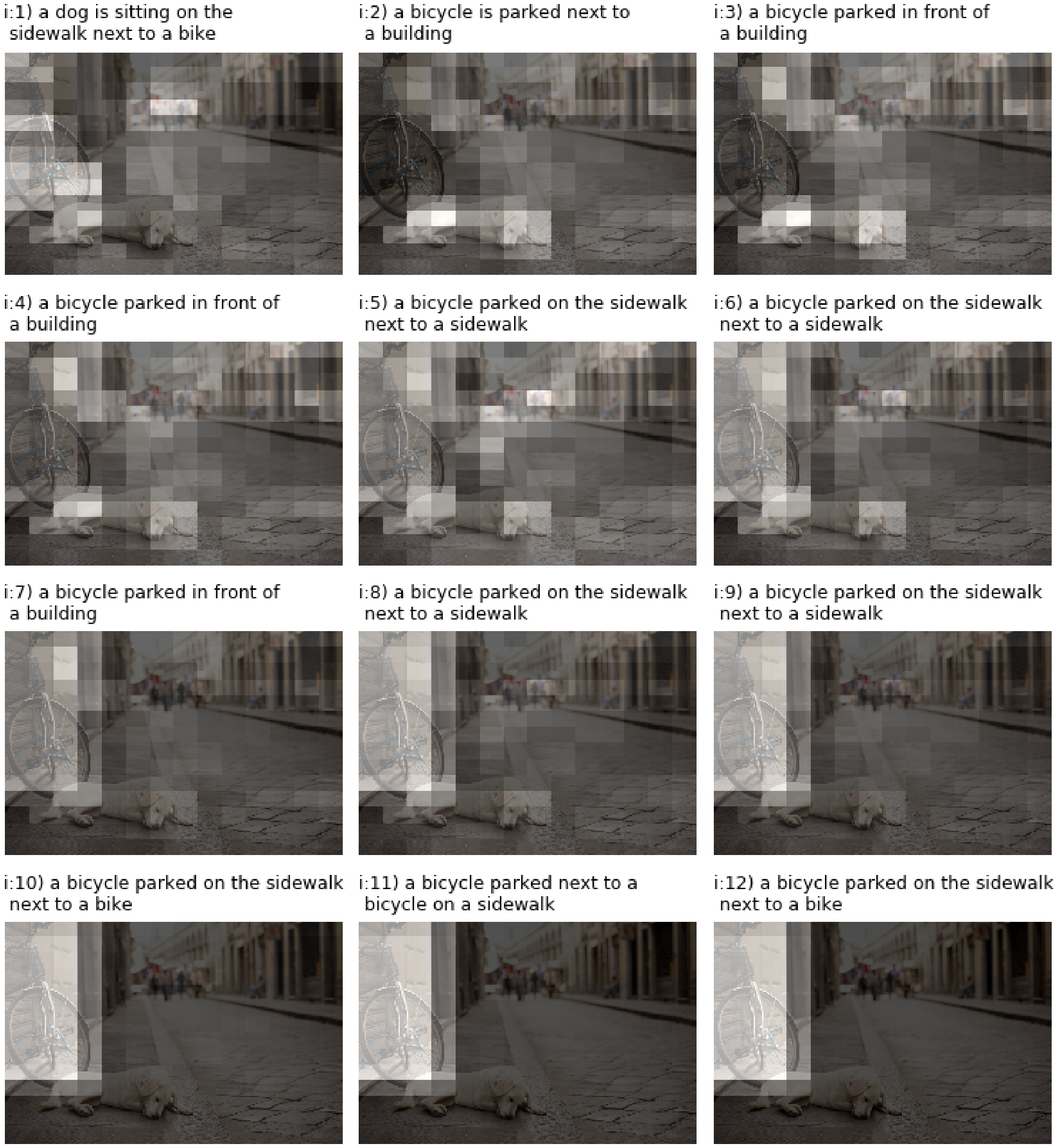


(a) Caption: ”a giraffe standing in a pen with a person standing in the background”

(b) Caption: ”a giraffe standing in a pen with a giraffe”

Figure 3: The box captions for another sample image with person watching giraffes in a pen. The attention is either fixed on the persons in (a) on the giraffe in (b). The model is producing captions that are mentioning the persons in (a) and only the giraffes in (b).

### A.3 Limited step-wise fixed attention on the sample image



i:1) a dog is sitting on the sidewalk next to a bike

i:2) a bicycle is parked next to a building

i:3) a bicycle parked in front of a building

i:4) a bicycle parked in front of a building

i:5) a bicycle parked on the sidewalk next to a sidewalk

i:6) a bicycle parked on the sidewalk next to a sidewalk

i:7) a bicycle parked in front of a building

i:8) a bicycle parked on the sidewalk next to a sidewalk

i:9) a bicycle parked on the sidewalk next to a sidewalk

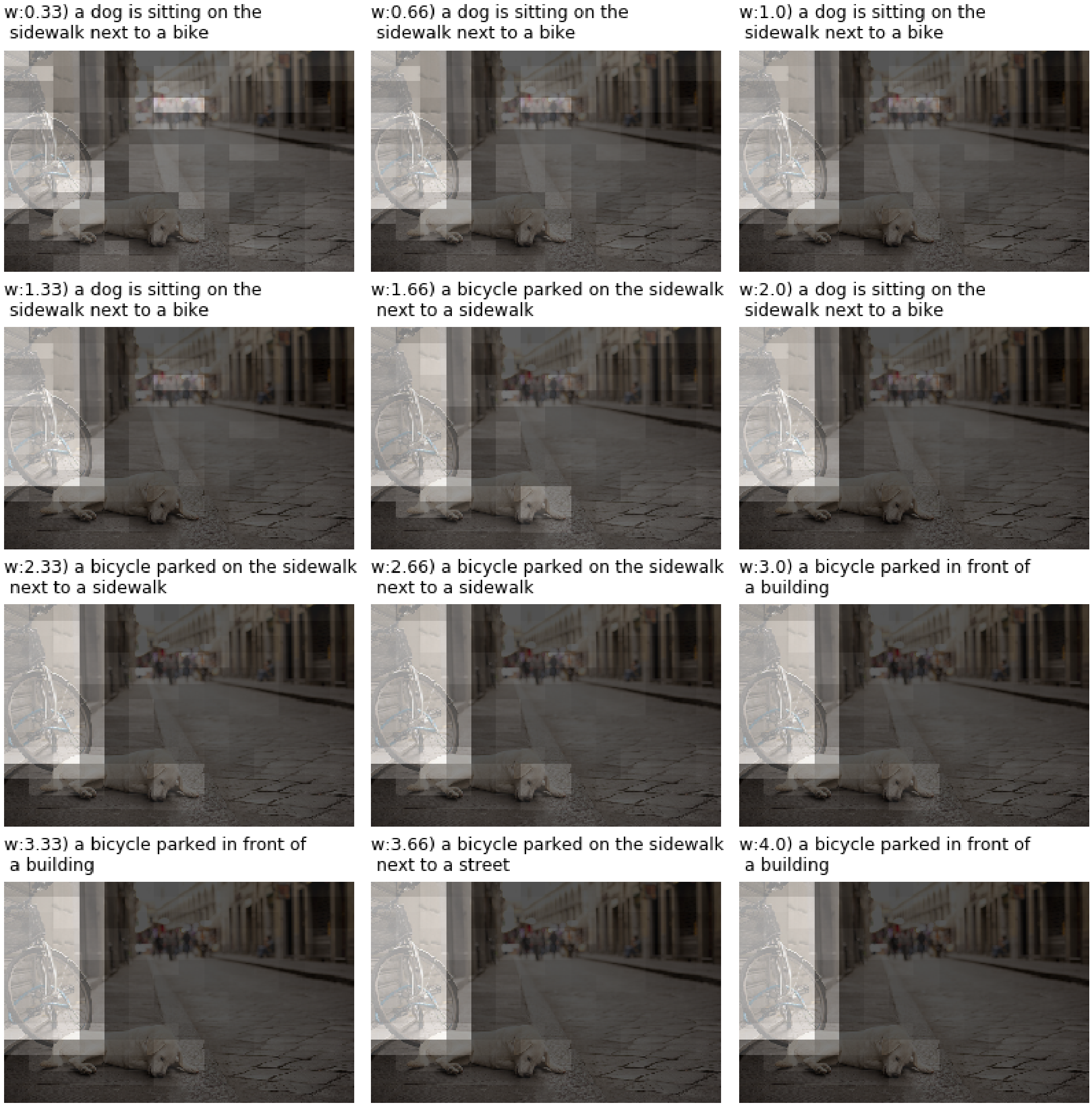
i:10) a bicycle parked on the sidewalk next to a bike

i:11) a bicycle parked next to a bicycle on a sidewalk

i:12) a bicycle parked on the sidewalk next to a bike

Figure 4: The limited step-wise fixed attention for the sample image with the dog and the bike. The number of fixed time-steps is increasing from the left-upper corner to the bottom-right corner from 1 to 12. The similarity with unlimited fixed attention is visible for more than nine fixed time-steps, while less than three time steps fixed are more similar to the alternating self-attention.

### A.4 Additive step-wise attention on the sample image



w:0.33) a dog is sitting on the sidewalk next to a bike

w:0.66) a dog is sitting on the sidewalk next to a bike

w:1.0) a dog is sitting on the sidewalk next to a bike

w:1.33) a dog is sitting on the sidewalk next to a bike

w:1.66) a bicycle parked on the sidewalk next to a sidewalk

w:2.0) a dog is sitting on the sidewalk next to a bike

w:2.33) a bicycle parked on the sidewalk next to a sidewalk

w:2.66) a bicycle parked on the sidewalk next to a sidewalk

w:3.0) a bicycle parked in front of a building

w:3.33) a bicycle parked in front of a building

w:3.66) a bicycle parked on the sidewalk next to a street

w:4.0) a bicycle parked in front of a building

Figure 5: The step-wise additive attention for the sample image with the dog and the bike. The additive weight factor is increasing from the left-upper corner to the bottom-right corner from 0.33 to 4. The similarity with unlimited fixed attention is already visible for small factors. Nevertheless, until the weight of two most captions mention the dog first and only then the bike, while starting from a weight of two, the bicycle becomes the main object in the caption.