Supplementary Material to
Normalized and Geometry-Aware Self-Attention Network for Image Captioning

1. Visualization of Geometric Weights

To gain a better insight about the effect of the relative geometry information on attention weights, we visualize the geometric weights in GSA. Specifically, we use the content-dependent version (ϕ^1) of GSA, and use a trained one-layer G-SAN model. We visualize how the geometric weight ϕ^1_{ij} between object i and j changes as the relative geometry feature $\mathbf{f}^g_{ij}$ between them changes.

Review that the relative geometry feature $\mathbf{f}^g_{ij}$ is a 4-dimensional vector:

$$\mathbf{f}^g_{ij} = \left(\log(\frac{|x_i - x_j|}{w_i}), \log(\frac{|y_i - y_j|}{h_i}), \log(\frac{w_i}{w_j}), \log(\frac{h_i}{h_j}) \right)^T, \tag{1}$$

where $(x_i, y_i), w_i, h_i$ are the center coordinate, width, and height of box i, respectively. We simplify the above equation as

$$\mathbf{f}^g_{ij} = (\log(\Delta x), \log(\Delta y), \log(\Delta w), \log(\Delta h))^T. \tag{2}$$

We then keep one of $(\Delta x, \Delta y)$ and $(\Delta w, \Delta h)$ fixed, change the other one, and see plot the values of ϕ^1_{ij}.

Figure 1 shows the cases when $\Delta w = \Delta h \in \{0.5, 1, 2\}$ are fixed, and Δx and Δy change in range of $[0, 3]$. We can observer that, basically, the geometric weight gets smaller when the relative distance between the two objects increases. Exceptions are found near $(\Delta x, \Delta y) = (0, 0)$, where the weights are relatively smaller than neighboring point. That is probably because when two boxes i and j have similar sizes, *e.g.* $\Delta w = \Delta h = 1$, and their center coordinates almost coincide, then they likely refer to the same object. Therefore, the weight of box j should be reduced to avoid repeating the object.

Figure 2 shows the cases when $\Delta x = \Delta y \in \{0.5, 1, 2\}$ are fixed, and Δw and Δh change in range of $[0, 3]$. We have the following observations. 1) The geometric weight is small when the size difference between the two boxes is too large, *i.e.* Δw or Δh is close to 0 or too large. 2) The geometric weight tends to be larger when two objects are close to each other, *e.g.* $(\Delta x, \Delta y) = (0.5, 0.5)$, than when their distance is large, *e.g.* $(\Delta x, \Delta y) = (2, 2)$.

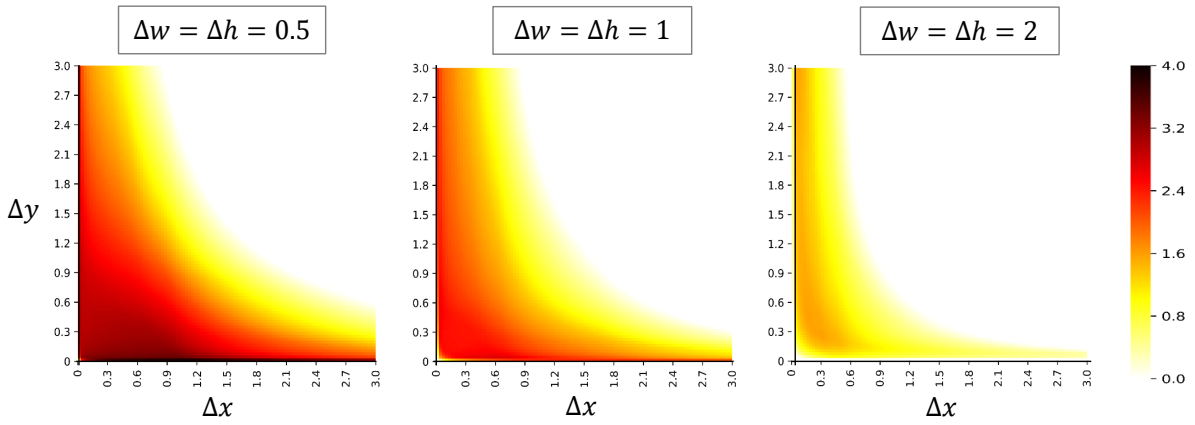

Figure 1: Visualization of the geometric weights as a function of the relative position. Each plot shows the values of ϕ^1_{ij} as Δx and Δy changing in range of $[0, 3]$, while $\Delta w = \Delta h \in \{0.5, 1, 2\}$ are kept fixed.

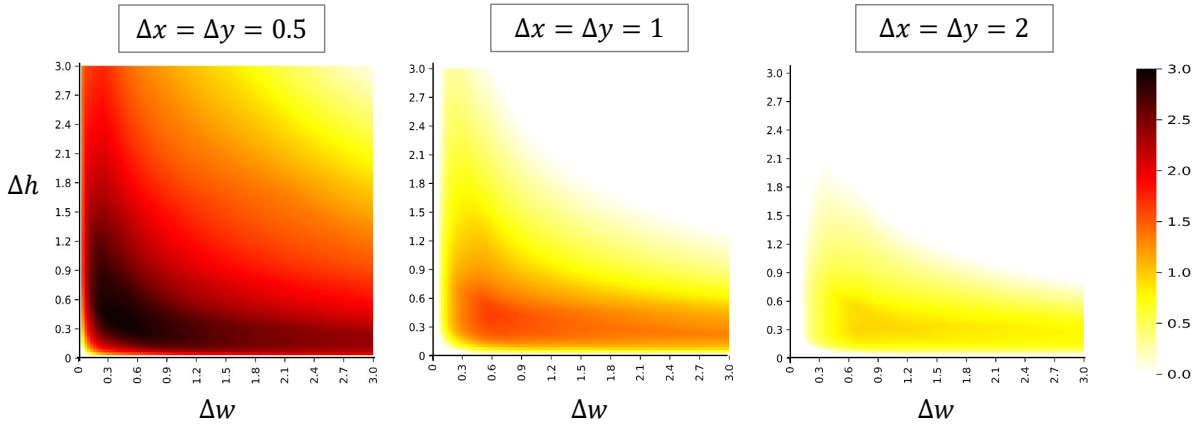

Figure 2: Visualization of the geometric weights as a function of the relative width and height. Each plot shows the values of ϕ_{ij}^1 as Δw and Δh changing in range of $[0, 3]$, while $\Delta x = \Delta y \in \{0.5, 1, 2\}$ are kept fixed.

2. Examples Results of Generated Captions

To illustrate the advantages of the G-SAN relative to the SAN, we present examples generated by the two models in Figure 3. The examples show that G-SAN is more good at determining the relationships between objects. For example, in the first example, our G-SAN generates "pots and pans *hanging on* the wall", which successfully recognizes the relationship between "pots and pans" and "wall", *i.e.* "hanging on".

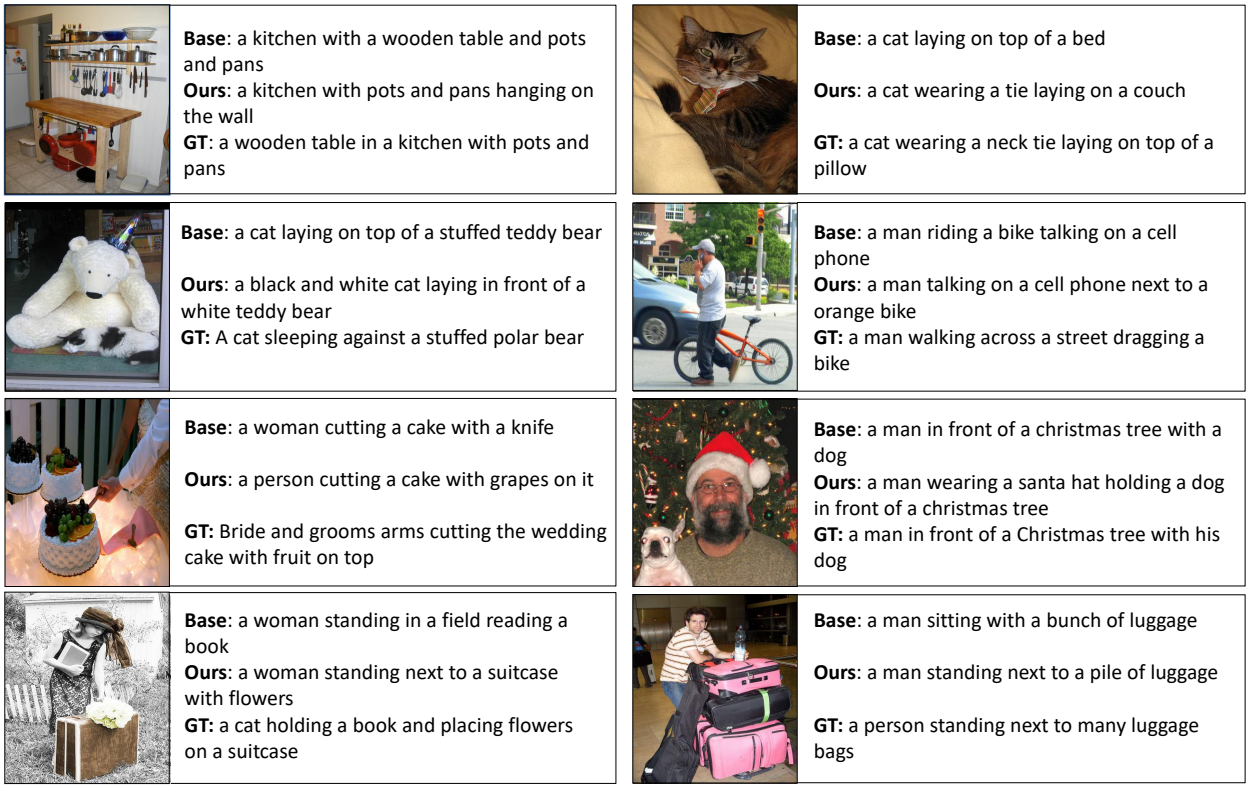

Figure 3: Example captions generated by the SAN baseline (denoted as 'Base') and our G-SAN (denoted as 'Ours') models. 'GT' denotes one of the five ground-truth captions. G-SAN shows an improvement over SAN in determining the relationships between objects.