

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_{ij}^1 between object i and j changes as the relative geometry feature \mathbf{f}_{ij}^g between them changes.

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

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

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

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

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_{ij}^1 .

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 observe 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.* $(\Delta x, \Delta y) = (2, 2)$.

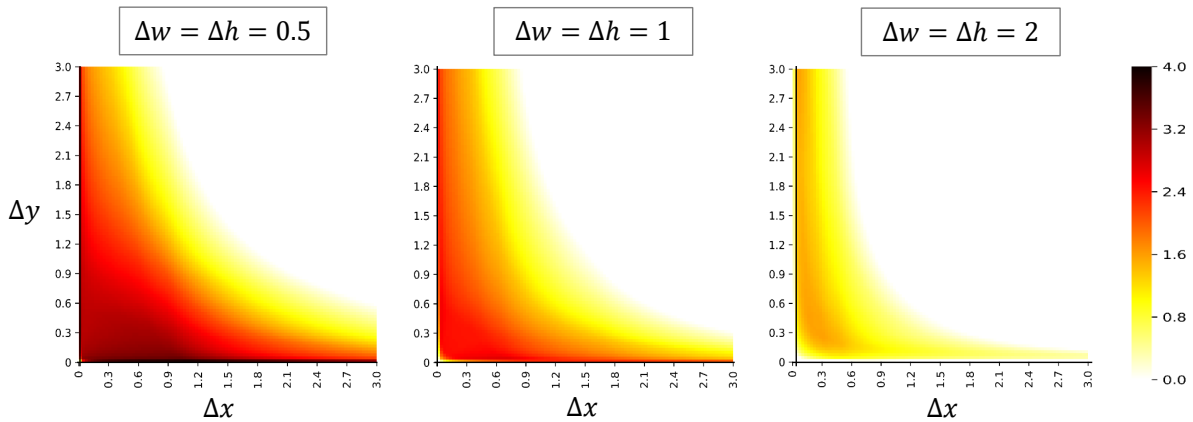


Figure 1: Visualization of the geometric weights as a function of the relative position. Each plot shows the values of ϕ_{ij}^1 as Δx and Δy changing in range of $[0, 3]$, while $\Delta w = \Delta h \in \{0.5, 1, 2\}$ are kept fixed.

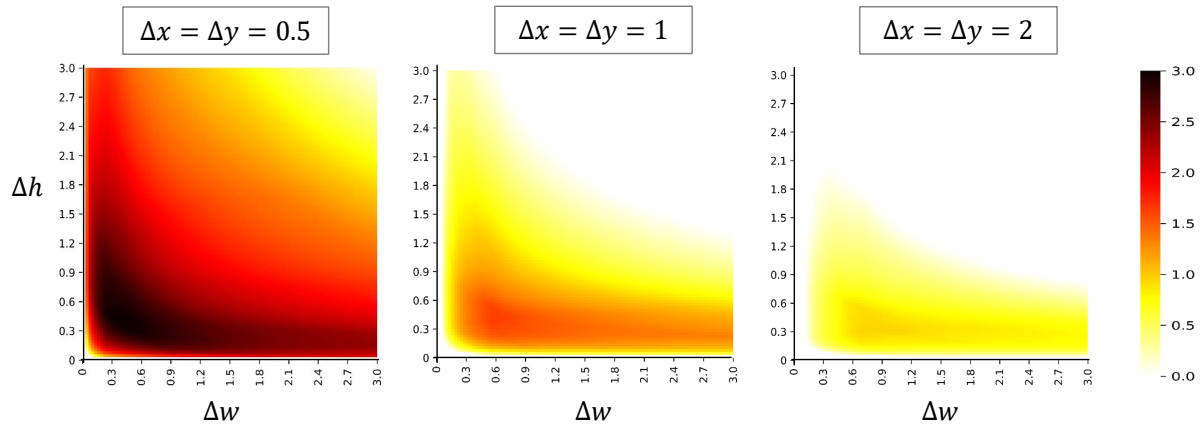


Figure 2: Visualization of the geometric weights as a function of the relative width and height. Each plot shows the values of ϕ_{ij}^1 as Δw and Δh changing in range of $[0, 3]$, while $\Delta x = \Delta y \in \{0.5, 1, 2\}$ are kept fixed.

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”.

	<p>Base: a kitchen with a wooden table and pots and pans</p> <p>Ours: a kitchen with pots and pans hanging on the wall</p> <p>GT: a wooden table in a kitchen with pots and pans</p>		<p>Base: a cat laying on top of a bed</p> <p>Ours: a cat wearing a tie laying on a couch</p> <p>GT: a cat wearing a neck tie laying on top of a pillow</p>
	<p>Base: a cat laying on top of a stuffed teddy bear</p> <p>Ours: a black and white cat laying in front of a white teddy bear</p> <p>GT: A cat sleeping against a stuffed polar bear</p>		<p>Base: a man riding a bike talking on a cell phone</p> <p>Ours: a man talking on a cell phone next to a orange bike</p> <p>GT: a man walking across a street dragging a bike</p>
	<p>Base: a woman cutting a cake with a knife</p> <p>Ours: a person cutting a cake with grapes on it</p> <p>GT: Bride and grooms arms cutting the wedding cake with fruit on top</p>		<p>Base: a man in front of a christmas tree with a dog</p> <p>Ours: a man wearing a santa hat holding a dog in front of a christmas tree</p> <p>GT: a man in front of a Christmas tree with his dog</p>
	<p>Base: a woman standing in a field reading a book</p> <p>Ours: a woman standing next to a suitcase with flowers</p> <p>GT: a cat holding a book and placing flowers on a suitcase</p>		<p>Base: a man sitting with a bunch of luggage</p> <p>Ours: a man standing next to a pile of luggage</p> <p>GT: a person standing next to many luggage bags</p>

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.