LORAS AND EQUIVARIANT ATTENTION MECHANISMS

AMELIE SCHREIBER

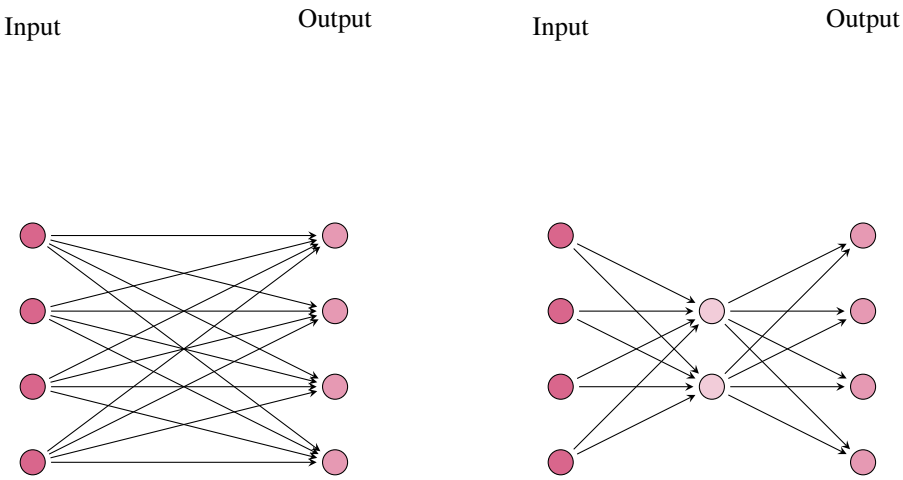
Abstract. The following is a merging of two theories, that of LoRAs or "Low Rank Adaptations" introduced in [Hu et. al.], and that of group equivariant self-attention in transformers or other neural networks that include attention. We prove three things. First, LoRAs do not disrupt translation equivariance when included in translation equivairant attention with relative positional encodings. Second, LoRAs doe not disrupt the lifting self-attention layers described in [RC]. And third, LoRAs do not disrupt equivariance of group self-attention layers as defined in [RC]. We also include a correction to the proof of equivariance of group self-attention by fixing the definition of the relative positional encodings defined in Proposition 5.2 and its proof in [RC].

Contents

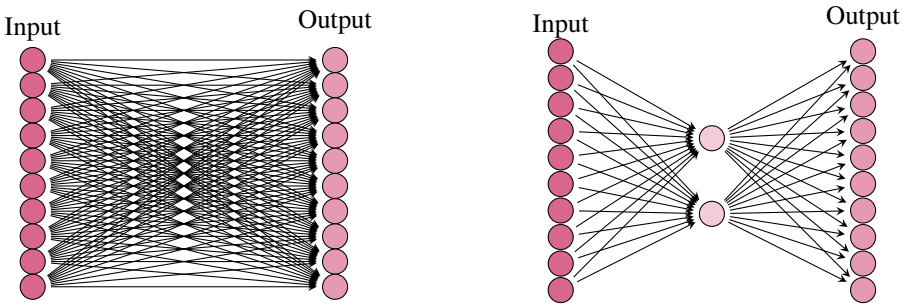
1. What is...a LoRA?

LoRAs or "Low Rank Adaptations" were introduced in the article [Hu et. al.]. They can be used in place of fine-tuning a neural network. We first freeze the original weights of a layer (learned weight matrix) in a neural network. For example, the query, key, or value weight matrix of the attention mechanism in a transformer might be frozen. We often denote these by W_Q, W_K, and W_V. We then add in a LoRA layer for one or more of these weight learned matrices. Suppose W is a frozen weight matrix. Then, a LoRA layer will have the form $W + \Delta W$, where $\Delta W = BA$ is the LoRA. Usually, these are low rank decompositions, with $A \in \mathbb{R}^{d_{in} \times r}$ and $B \in \mathbb{R}^{r \times d_{out}}$, where the original weight matrix is $W \in \mathbb{R}^{d_{out} \times d_{in}}$. Usually, we have $r << \min\{d_{in}, d_{out}\}$. If the rank is not much smaller than the input and output dimension, there is little benefit in applying a LoRA. However, we can still choose an $r << \min\{d_{in}, d_{out}\}$ and apply a LoRA in place of fine tuning. It is shown empirically that in many cases choosing $r = 4$ or 8 is sufficient even for very large weight matrices such as the query, key, and value matrices of the attention mechanism in a transformer. Let's first look at an example where there is no real benefit (in terms of reduction in parameters) in applying a LoRA:

Date: May 19, 2023.
2020 *Mathematics Subject Classification.* Primary 68T07, 68T10, 68T45, Secondary 68T50 .



Here, we see that the number of parameters for the LoRA layer $\Delta W = BA$ is the same as the original layer W, where we have $4 \times 2 \times 2 = 16$ parameters for the LoRA (on the right), and $4 \times 4 = 16$ parameters for the original frozen weight matrix (on the left). Next, let's look at an example that gives us 40% the parameters of the frozen weight matrix:



Here we see the original (frozen) weight matrix has 10^2 parameters, and the LoRA has only $10 \times 2 \times 2 = 40$ parameters. In most cases, we have that the rank (ideally this is the number of neurons in the middle layer of the LoRA) of the frozen matrix is much smaller than the input and output dimensions, and there is in fact a drastic reduction in parameter count. As an example, we might have an input and output dimension of say 100, in which case the weight matrix has $100^2 = 10,000$ parameters. However, the rank of this matrix is very often much lower than 100. In practice, it was shown that choosing $r = 4$ for the query, key, and value matrices is often more than sufficient for a LoRA as the middle dimension. In this case, we would get $100 \times 4 \times 2 = 800$ parameters in the LoRA, which is less than one tenth the original parameter count. Once we have such a LoRA in place, we can train it on some downstream task, and then add the LoRA weight matrix BA to the original (frozen) weight matrix W to obtain a model that performs well on this new task. Now, let us look at how LoRAs can be added into a translation equivariant attention mechanism without any loss of translation equivariance.

2. Incldung LoRAs Does Not Disrupt Translation Equivariance

In the following proof, we follow notation very similar to [RC]. Suppose we are given an input matrix $X \in \mathbb{R}^{d_{in} \times d}$, with columns representing the embedding vectors of d tokens. We can formulate the self-attention matrix in a transformer with relative positional encoding as

$$A_{i,j} = (W_Q X)(W_K(X + P_{x(j)-x(i)}))^T \tag{1}$$

We leave out the scaling factor $1/\sqrt{d_k}$ for simplicity, but it can easily be included without disrupting any of our arguments or proofs. It will be more beneficial for our purposes to formulate self-attention in a function theoretic way. In particular, we can view the input matrix X as a vector valued function $f : S \to \mathbb{R}^{d_{in}}$, that is $f \in L_{\mathbb{R}^{d_{in}}}(S)$, for the index set $S = \{1, 2, ..., d\}$. We then view the query and key matrices as maps $\varphi_{qry} : L_{\mathbb{R}^{d_{in}}}(S) \to L_{\mathbb{R}^{d_k}}(S)$ and $\varphi_{key} : L_{\mathbb{R}^{d_{in}}}(S) \to L_{\mathbb{R}^{d_k}}(S)$. There is also a value function $\varphi_{val} : L_{\mathbb{R}^{d_{in}}}(S) \to L_{\mathbb{R}^{d_v}}(S)$. With these in hand, we can express the attention map with positional encoding as

$$A_{i,j} = \alpha[f](i, j) = \langle \varphi_{qry}(f(i)), \varphi_{key}(f(j) + \rho(i, j)) \rangle \tag{2}$$

Here, we have written $\rho(i, j)$ for the positional encoding. The map $\alpha[f] : S \times S \to \mathbb{R}$ maps pairs of elements $i, j \in S$ to the attention score of j relative to i. We can then write the attention mechanism as

$$\zeta[f](i) = \sum_{j \in S} \sigma_j(\alpha[f](i, j))\, \varphi_{val}(f(j)) \tag{3}$$

Next, we would like to include a LoRA for the query, key, and value maps. We will formulate this as

$$\Delta\varphi_{qry}(f(i)) = (\varphi^A_{qry} \circ \varphi^B_{qry})(f(i)) = \varphi^B_{qry}(\varphi^A_{qry}(f(i))) \tag{4}$$

$$\Delta\varphi_{key}(f(i)) = (\varphi^A_{key} \circ \varphi^B_{key})(f(i)) = \varphi^B_{key}(\varphi^A_{key}(f(i))) \tag{5}$$

$$\Delta\varphi_{val}(f(i)) = (\varphi^A_{val} \circ \varphi^B_{val})(f(i)) = \varphi^B_{val}(\varphi^A_{val}(f(i))) \tag{6}$$

Here, we have

$$\varphi^A_{qry} : L_{\mathbb{R}^{d_k}}(S) \to L_{\mathbb{R}^{r(A)}}(S) \qquad \varphi^B_{qry} : L_{\mathbb{R}^{r(A)}}(S) \to L_{\mathbb{R}^{d_k}}(S) \tag{7}$$

$$\varphi^A_{key} : L_{\mathbb{R}^{d_k}}(S) \to L_{\mathbb{R}^{r(A)}}(S) \qquad \varphi^B_{key} : L_{\mathbb{R}^{r(A)}}(S) \to L_{\mathbb{R}^{d_k}}(S) \tag{8}$$

$$\varphi^A_{val} : L_{\mathbb{R}^{d_v}}(S) \to L_{\mathbb{R}^{r(A)}}(S) \qquad \varphi^B_{val} : L_{\mathbb{R}^{r(A)}}(S) \to L_{\mathbb{R}^{d_v}}(S) \tag{9}$$

Next, including this in the attention mechanism, we get

$$\alpha^{LoRA}[f](i, j) = \langle \varphi_{qry}(f(i)) + \Delta\varphi_{qry}(f(i)), \varphi_{key}(f(j) + \rho(i, j)) + \Delta\varphi_{key}(f(j) + \rho(i, j)) \rangle \tag{10}$$

and then

$$\zeta^{LoRA}[f](i) = \sum_{j \in S} \sigma_j\left(\alpha^{LoRA}[f](i, j)\right)(\varphi_{val}(f(j)) + \Delta\varphi_{val}(f(j))) \tag{11}$$

Now, in order to have translation equivariance of a LoRA multihead self-attention with relative positional encoding, we need the following equation to hold,

$$m^r_{LoRA}[L_y[f],\rho](i) = L_y[m^r_{LoRA}[f,\rho]](i) \tag{12}$$

where $L_y[f](i) = f(x^{-1}(x(i) - y))$. The LoRA multihead self-attention with relative positional encodings on $L_y[f]$ is given by

$$\begin{aligned}
&m^r_{LoRA}[L_y[f],\rho](i)\\
&= \varphi_{out}\Bigg(\bigcup_{h\in[H]}\sum_{j\in N(i)}\sigma_j\Bigg(\Big\langle\varphi^{(h)}_{qry}(L_y[f](i)) + \Delta\varphi^{(h)}_{qry}(f(i)),\\
&\quad\varphi^{(h)}_{key}(L_y[f](j) + \rho(i,j)) + \Delta\varphi^{(h)}_{key}(f(j) + \rho(i,j))\Big\rangle\Bigg)\Bigg(\varphi^{(h)}_{val}(L_y[f](j)) + \Delta\varphi^{(h)}_{val}(f(j))\Bigg)\Bigg)\\
&= \varphi_{out}\Bigg(\bigcup_{h\in[H]}\sum_{j\in N(i)}\sigma_j\Bigg(\Big\langle\varphi^{(h)}_{qry}(f(x^{-1}(x(i) - y))) + \Delta\varphi^{(h)}_{qry}(f(x^{-1}(x(i) - y))),\\
&\quad\varphi^{(h)}_{key}(f(x^{-1}(x(j) - y)) + \rho(i,j)) + \Delta\varphi^{(h)}_{key}(f(x^{-1}(x(j) - y)) + \rho(i,j))\Big\rangle\Bigg)\\
&\quad\times\Bigg(\varphi^{(h)}_{val}(f(x^{-1}(x(j) - y))) + \Delta\varphi^{(h)}_{val}(f(x^{-1}(x(j) - y)))\Bigg)\Bigg)\\
&= \varphi_{out}\Bigg(\bigcup_{h\in[H]}\sum_{x^{-1}(x(\bar{j})+y)\in N(x^{-1}(x(\bar{i})+y))}\sigma_{x^{-1}(x(\bar{j})+y)}\Bigg(\Big\langle\varphi^{(h)}_{qry}(f(\bar{i})) + \Delta\varphi^{(h)}_{qry}(f(\bar{i})),\\
&\quad\varphi^{(h)}_{key}(f(\bar{j}) + \rho(x^{-1}(x(\bar{i}) + y), x^{-1}(x(\bar{j}) + y))) + \Delta\varphi^{(h)}_{key}(f(\bar{j}) + \rho(x^{-1}(x(\bar{i}) + y), x^{-1}(x(\bar{j}) + y)))\Big\rangle\Bigg)\\
&\quad\times\Bigg(\varphi^{(h)}_{val}(f(\bar{j})) + \Delta\varphi^{(h)}_{val}(f(\bar{j}))\Bigg)\Bigg)
\end{aligned}$$

Here we have used the substitution $\bar{i} = x^{-1}(x(i)-y) \implies i = x^{-1}(x(\bar{i})+y)$ and $\bar{j} = x^{-1}(x(j)-y) \implies j = x^{-1}(x(\bar{j})+y)$. We can further reduce the equations using the defintion of $\rho(i,j) = \rho^P(x(j) - x(i))$:

$$= \varphi_{out}\Bigg(\bigcup_{h\in[H]} \sum_{x^{-1}(x(\bar{j})+y)\in N(x^{-1}(x(\bar{i})+y))} \sigma_{x^{-1}(x(\bar{j})+y)}\Big(\Big\langle \varphi^{(h)}_{qry}(f(\bar{i})) + \Delta\varphi^{(h)}_{qry}(f(\bar{i})),$$

$$\varphi^{(h)}_{key}(f(\bar{j}) + \rho^{P}(x(\bar{j}) + y - (x(\bar{i}) + y))) + \Delta\varphi^{(h)}_{key}(f(\bar{j}) + \rho^{P}(x(\bar{j}) + y - (x(\bar{i}) + y)))\Big\rangle\Big)\Big(\varphi^{(h)}_{val}(f(\bar{j})) + \Delta\varphi^{(h)}_{val}(f(\bar{j}))\Big)\Bigg)$$

$$= \varphi_{out}\Bigg(\bigcup_{h\in[H]} \sum_{x^{-1}(x(\bar{j})+y)\in N(x^{-1}(x(\bar{i})+y))} \sigma_{x^{-1}(x(\bar{j})+y)}\Big(\Big\langle \varphi^{(h)}_{qry}(f(\bar{i})) + \Delta\varphi^{(h)}_{qry}(f(\bar{i})),$$

$$\varphi^{(h)}_{key}(f(\bar{j}) + \rho^{P}(x(\bar{j}) - x(\bar{i}))) + \Delta\varphi^{(h)}_{key}(f(\bar{j}) + \rho^{P}(x(\bar{j}) - x(\bar{i})))\Big\rangle\Big)\Big(\varphi^{(h)}_{val}(f(\bar{j})) + \Delta\varphi^{(h)}_{val}(f(\bar{j}))\Big)\Bigg)$$

$$= \varphi_{out}\Bigg(\bigcup_{h\in[H]} \sum_{x^{-1}(x(\bar{j})+y)\in N(x^{-1}(x(\bar{i})+y))} \sigma_{x^{-1}(x(\bar{j})+y)}\Big(\Big\langle \varphi^{(h)}_{qry}(f(\bar{i})) + \Delta\varphi^{(h)}_{qry}(f(\bar{i})),$$

$$\varphi^{(h)}_{key}(f(\bar{j}) + \rho(\bar{i},\bar{j})) + \Delta\varphi^{(h)}_{key}(f(\bar{j}) + \rho(\bar{i},\bar{j}))\Big\rangle\Big)\Big(\varphi^{(h)}_{val}(f(\bar{j})) + \Delta\varphi^{(h)}_{val}(f(\bar{j}))\Big)\Bigg)$$

For any translation $y \in \mathbb{R}^{d_{in}}$, where d_{in} is the dimension of $f(i)$ and $f(j)$, the summation remains the same, so we have:

$$\sum_{x^{-1}(x(\bar{j})+y)\in N(x^{-1}(x(\bar{i})+y))} [\bullet] = \sum_{x^{-1}(x(\bar{j}))\in N(x^{-1}(x(\bar{i})))} [\bullet] = \sum_{\bar{j}\in N(\bar{i})} [\bullet] \tag{13}$$

As we can see, $m^{r}_{LoRA}[L_y[f],\rho](i) = L_y[m^{r}_{LoRA}[f,\rho]](i)$. We can thus conclude that relative positional encodings, coupled with LoRAs for the query, key, and value weight matrices gives a translation equivariant multihead self-attention mechanism. In particular, addition of LoRAs in a translation equivariant model with relative positional encodings does not disrupt the translation equivariance, which means we can further reduce the expression as

$$\begin{aligned} m^{r}_{LoRA}[L_y[f],\rho](i) &= \varphi_{out}\Bigg(\bigcup_{h\in[H]} \sum_{x^{-1}(x(\bar{j})+y)\in N(x^{-1}(x(\bar{i})+y))} \sigma_{x^{-1}(x(\bar{j})+y)}\Big(\Big\langle \varphi^{(h)}_{qry}(f(\bar{i})) + \Delta\varphi^{(h)}_{qry}(f(\bar{i})),\\ &\qquad \varphi^{(h)}_{key}(f(\bar{j}) + \rho(\bar{i},\bar{j})) + \Delta\varphi^{(h)}_{key}(f(\bar{j}) + \rho(\bar{i},\bar{j}))\Big\rangle\Big)\Big(\varphi^{(h)}_{val}(f(\bar{j})) + \Delta\varphi^{(h)}_{val}(f(\bar{j}))\Big)\Bigg)\\ &= m^{r}_{LoRA}[f,\rho](\bar{i})\\ &= m^{r}_{LoRA}[f,\rho](x^{-1}(x(i) - y))\\ &= L_y[m^{r}_{LoRA}[f,\rho]](i) \end{aligned}$$

To reiterate in slightly different notation, for translation equivariance of a LoRA multihead self-attention with relative positional encoding, the following equation needs to hold,

$$m^{r}_{LoRA}[L_y[f],\rho](i) = L_y[m^{r}_{LoRA}[f,\rho]](i) \tag{14}$$

where $L_y[f](i) = f(x^{-1}(x(i) - y))$. The LoRA multihead self-attention with relative positional encodings on $L_y[f]$ is:

$$
\begin{aligned}
&m^r_{LoRA}[L_y[f],\rho](i)\\
&= \varphi_{out}\Bigg(\bigcup_{h\in[H]}\sum_{j\in N(i)}\frac{\exp\left\langle \varphi^{(h)}_{qry}(L_y[f](i)) + \Delta\varphi^{(h)}_{qry}(f(i)), \varphi^{(h)}_{key}(L_y[f](j)+\rho(i,j)) + \Delta\varphi^{(h)}_{key}(f(j)+\rho(i,j))\right\rangle}{\sum_{k\in N(i)}\exp\left\langle \varphi^{(h)}_{qry}(L_y[f](k)) + \Delta\varphi^{(h)}_{qry}(f(k)), \varphi^{(h)}_{key}(L_y[f](k)+\rho(i,k)) + \Delta\varphi^{(h)}_{key}(f(k)+\rho(i,k))\right\rangle}\\
&\quad\times\left(\varphi^{(h)}_{val}(L_y[f](j)) + \Delta\varphi^{(h)}_{val}(f(j))\right)\Bigg)
\end{aligned}
$$

We use the substitution $\bar{i} = x^{-1}(x(i)-y) \implies i = x^{-1}(x(\bar{i})+y)$ and $\bar{j} = x^{-1}(x(j)-y) \implies j = x^{-1}(x(\bar{j})+y)$, and the definition of $\rho(i,j) = \rho^P(x(j)-x(i))$:

$$
\begin{aligned}
&m^r_{LoRA}[L_y[f],\rho](i)\\
&= \varphi_{out}\Bigg(\bigcup_{h\in[H]}\sum_{\bar{j}\in N(\bar{i})}\frac{\exp\left\langle \varphi^{(h)}_{qry}(f(\bar{i})) + \Delta\varphi^{(h)}_{qry}(f(\bar{i})), \varphi^{(h)}_{key}(f(\bar{j})+\rho(\bar{i},\bar{j})) + \Delta\varphi^{(h)}_{key}(f(\bar{j})+\rho(\bar{i},\bar{j}))\right\rangle}{\sum_{\bar{k}\in N(\bar{i})}\exp\left\langle \varphi^{(h)}_{qry}(f(\bar{k})) + \Delta\varphi^{(h)}_{qry}(f(\bar{k})), \varphi^{(h)}_{key}(f(\bar{k})+\rho(\bar{i},\bar{k})) + \Delta\varphi^{(h)}_{key}(f(\bar{k})+\rho(\bar{i},\bar{k}))\right\rangle}\\
&\quad\times\left(\varphi^{(h)}_{val}(f(\bar{j})) + \Delta\varphi^{(h)}_{val}(f(\bar{j}))\right)\Bigg)\\
&= m^r_{LoRA}[f,\rho](\bar{i})\\
&= m^r_{LoRA}[f,\rho](x^{-1}(x(i)-y))\\
&= L_y[m^r_{LoRA}[f,\rho]](i)
\end{aligned}
$$

The last equations conclude that relative positional encodings, coupled with LoRAs for the query, key, and value weight matrices, result in a translation equivariant multihead self-attention mechanism. More specifically, addition of LoRAs in a translation equivariant model with relative positional encodings does not disrupt the translation equivariance.

3. LoRAs and Lifting Self-Attention

For the lifting self-attention layer to be G-equivariant, we must have that $m^r_{G\uparrow}[L_g[f],\rho](i,h) = [m^r_{G\uparrow}[f,\rho]](i,h)$. Consider a g-transformed input signal $L_g[f](i) = L_y L_{h_3}[f](i) = f(x^{-1}(h_3^{-1}(x(i)-y)))$. Here $g = (y,h_3) \in \mathbb{R}^{d_{in}} \rtimes \mathcal{H}$. The lifting self-attention layer with LoRAs, applied to $L_g[f]$ is:

$$
\begin{aligned}
&m^{r,LoRA}_{G\uparrow}[L_y L_{h_3}[f],\rho](i,h)\\
&= \varphi_{out}\Bigg(\bigcup_{head\in[H]}\sum_{j\in N(i)}\sigma_j\Big(\langle\varphi^{head}_{qry}(L_y L_{h_3}[f](i)) + \Delta\varphi^{head}_{qry}(L_y L_{h_3}[f](i)), \varphi^{head}_{key}(L_y L_{h_3}[f](j)\\
&\quad + L_h[\rho](i,j)) + \Delta\varphi^{head}_{key}(L_y L_{h_3}[f](j) + L_h[\rho](i,j))\rangle\Big)(\varphi^{head}_{val}(L_y L_{h_3}[f](j)) + \Delta\varphi^{head}_{val}(L_y L_{h_3}[f](j)))\Bigg)\\
&= \varphi_{out}\Bigg(\bigcup_{head\in[H]}\sum_{j\in N(i)}\sigma_j\Big(\langle\varphi^{head}_{qry}(f(x^{-1}(h_3^{-1}(x(i)-y))) + \Delta\varphi^{head}_{qry}(f(x^{-1}(h_3^{-1}(x(i)-y))), \varphi^{head}_{key}(f(x^{-1}(h_3^{-1}(x(j)-y)))\\
&\quad + L_h[\rho](i,j) + \Delta\varphi^{head}_{key}(f(x^{-1}(h_3^{-1}(x(j)-y))) + L_h[\rho](i,j))\rangle\Big)\\
&\quad \times(\varphi^{head}_{val}(f(x^{-1}(h_3^{-1}(x(j)-y)))) + \Delta\varphi^{head}_{val}(f(x^{-1}(h_3^{-1}(x(j)-y))))\Bigg)\\
&= \varphi_{out}\Bigg(\bigcup_{head\in[H]}\sum_{x^{-1}(h_3^{-1}x(\bar{j})+y)\in N(x^{-1}(h_3^{-1}x(\bar{i})+y))}\sigma_j\Big(\langle\varphi^{head}_{qry}(f(\bar{i})) + \Delta\varphi^{head}_{qry}(f(\bar{i})), \varphi^{head}_{key}(f(\bar{j}))\\
&\quad + L_h[\rho](x^{-1}(h_3x(\bar{i})+y), x^{-1}(h_3x(\bar{j})+y)) + \Delta\varphi^{head}_{key}(f(\bar{j}) + L_h[\rho](x^{-1}(h_3x(\bar{i})+y), x^{-1}(h_3x(\bar{j})+y)))\rangle\Big)\\
&\quad \times(\varphi^{head}_{val}(f(\bar{j})) + \Delta\varphi^{head}_{val}(f(\bar{j})))\Bigg)
\end{aligned}
$$

Here we have used $\bar{i} = x^{-1}(h_3^{-1}(x(i)-y)) \implies i = x^{-1}(h_3x(\bar{i})+y)$ and $\bar{j} = x^{-1}(h_3^{-1}(x(j)-y)) \implies j = x^{-1}(h_3x(\bar{j})+y)$. By using the defintion of $\rho(i,j)$ we can further reduce the expression:

$$
\begin{aligned}
&= \varphi_{out}\Bigg(\bigcup_{h\in[H]} \sum_{x^{-1}(h_3^{-1}x(\bar{j})+y)\in N(x^{-1}(h_3^{-1}x(\bar{i})+y))} \sigma_{x^{-1}(h_3^{-1}x(\bar{j})+y)}\Big\langle \varphi_{qry}^{(h)} f(\bar{i}) + \Delta\varphi_{qry}^{(h)} f(\bar{i}), \varphi_{key}^{(h)}(f(\bar{j}) \\
&\quad + \rho^P(h^{-1}(h_3^{-1}x(\bar{j}) + y) - h^{-1}(h_3 x(\bar{i}) + y))) + \Delta\varphi_{key}^{(h)}(f(\bar{j}) + \rho^P(h^{-1}(h_3^{-1}x(\bar{j}) + y) - h^{-1}(h_3 x(\bar{i}) + y)))\Big\rangle\Bigg) \\
&\quad \times (\varphi_{val}^{(h)}(f(\bar{j})) + \Delta\varphi_{val}^{(h)}(f(\bar{j})))\Bigg) \\
&= \varphi_{out}\Bigg(\bigcup_{h\in[H]} \sum_{x^{-1}(h_3^{-1}x(\bar{j})+y)\in N(x^{-1}(h_3^{-1}x(\bar{i})+y))} \sigma_{x^{-1}(h_3^{-1}x(\bar{j})+y)}\Big\langle \varphi_{qry}^{(h)} f(\bar{i}) + \Delta\varphi_{qry}^{(h)} f(\bar{i}), \varphi_{key}^{(h)}(f(\bar{j}) \\
&\quad + \rho^P(h^{-1}h_3(x(\bar{j}) - x(\bar{i}))) + \Delta\varphi_{key}^{(h)}(f(\bar{j}) + \rho^P(h^{-1}h_3(x(\bar{j}) - x(\bar{i})))))\Big\rangle\Bigg) \\
&\quad \times (\varphi_{val}^{(h)}(f(\bar{j})) + \Delta\varphi_{val}^{(h)}(f(\bar{j})))\Bigg) \\
&= \varphi_{out}\Bigg(\bigcup_{h\in[H]} \sum_{x^{-1}(h_3^{-1}x(\bar{j})+y)\in N(x^{-1}(h_3^{-1}x(\bar{i})+y))} \sigma_{x^{-1}(h_3^{-1}x(\bar{j})+y)}\Big\langle \varphi_{qry}^{(h)} f(\bar{i}), \varphi_{key}^{(h)}(f(\bar{j}) \\
&\quad + L_{h_3^{-1}h}(\bar{i},\bar{j}) + \Delta\varphi_{key}^{(h)}(f(\bar{j}) + L_{h_3^{-1}h}(\bar{i},\bar{j}))\Big\rangle\Bigg) \\
&\quad \times (\varphi_{val}^{(h)}(f(\bar{j})) + \Delta\varphi_{val}^{(h)}(f(\bar{j})))\Bigg)
\end{aligned}
$$

Since unimodular (or compact) groups the area of summation remains equal for any $g \in G$, we have:

$$
\sum_{x^{-1}(h_3 x(\bar{j})+y)\in N(x^{-1}(h_3 x(\bar{i})+y))} [\bullet] = \sum_{x^{-1}(h_3 x(\bar{j}))\in N(x^{-1}(h_3 x(\bar{i})))} [\bullet] = \sum_{x^{-1}(x(\bar{j}))\in N(x(\bar{i})))} [\bullet] = \sum_{\bar{j}\in N(\bar{i})} [\bullet] \tag{15}
$$

Resultantly, we can further reduce the expression above as:

$$
\begin{aligned}
&m_{G\uparrow}^{r,LoRA}[L_y L_{h_3}[f],\rho](i,h) \\
&= \varphi_{out}\Bigg(\bigcup_{h\in[H]} \sum_{\bar{j}\in N(\bar{i})} \sigma_{\bar{j}}\Big\langle \varphi_{qry}^{(h)}(f(\bar{i})) + \Delta\varphi_{qry}^{(h)}(f(\bar{i})), \varphi_{key}^{(h)}(f(\bar{j}) + L_{h_3^{-1}h}[\rho](\bar{i},\bar{j})) + \Delta\varphi_{key}^{(h)}(f(\bar{j}) + L_{h_3^{-1}h}[\rho](\bar{i},\bar{j}))\Big\rangle\Bigg) \\
&\quad \times (\varphi_{val}^{(h)}(f(\bar{j})) + \Delta\varphi_{val}^{(h)}(f(\bar{j})))\Bigg) \\
&= m_{G\uparrow}^{r,LoRA}[r,\rho](\bar{i}, h_3^{-1}h) \\
&= m_{G\uparrow}^{r,LoRA}(x^{-1}(h_3^{-1}(x(i) - y)), h_3^{-1}h) \\
&= L_y L_{h_3}[m_{G\uparrow}^{r,LoRA}[f,\rho]](i,h)
\end{aligned}
$$

So, we see that adding LoRAs does not disrupt equivariance of the lifting self-attention layer.

4. LoRAs and Group Self-Attention

Next, we can also show that the inclusion of LoRAs does not disrupt equivariance.

$$
\begin{aligned}
&m^r_{G,LoRA}[L_y L_{h_3}[f],\rho](i,h_1)\\
&= \varphi_{out}\Bigg(\bigcup_{head\in[H]} \sum_{h\in\mathcal{H}} \sum_{(j,h_2)\in N(i,h_1)} \sigma_{(j,h_2)}\Bigg(\langle \varphi^{head}_{qry} L_y L_{h_3}[f](i,h_1) + \Delta\varphi^{head}_{qry} L_y L_{h_3}[f](i,h_1), \varphi^{head}_{key}(L_y L_{h_3}[f](j,h_2)\\
&\quad + L_h[\rho]((i,h_1),(j,h_2))) + \Delta\varphi^{head}_{key}(L_y L_{h_3}[f](j,h_2) + L_h[\rho]((i,h_1),(j,h_2)))\rangle\Bigg)\\
&\quad \times (\varphi^{head}_{val}(L_y L_{h_3}[f](j,h_2)) + \Delta\varphi^{head}_{val}(L_y L_{h_3}[f](j,h_2)))\Bigg)\\
&= \varphi_{out}\Bigg(\bigcup_{head\in[H]} \sum_{h\in\mathcal{H}} \sum_{(j,h_2)\in N(i,h_1)} \sigma_{(j,h_2)}\Bigg(\langle \varphi^{head}_{qry} f(x^{-1}(h_3^{-1}(x(i)-y)), h_3^{-1}h_1) + \Delta\varphi^{head}_{qry} f(x^{-1}(h_3^{-1}(x(i)-y)), h_3^{-1}h_1),\\
&\quad \varphi^{head}_{key}(f(x^{-1}(h_3^{-1}(x(i)-y)), h_3^{-1}h_2)\\
&\quad + L_h[\rho]((i,h_1),(j,h_2))) + \Delta\varphi^{head}_{key}(f(x^{-1}(h_3^{-1}(x(i)-y)), h_3^{-1}h_2) + L_h[\rho]((i,h_1),(j,h_2)))\rangle\Bigg)\\
&\quad (\varphi^{head}_{val}(f(x^{-1}(h_3^{-1}(x(i)-y)), h_3^{-1}h_2)) + \Delta\varphi^{head}_{val}(f(x^{-1}(h_3^{-1}(x(i)-y)), h_3^{-1}h_2)))\Bigg)\\
&= \varphi_{out}\Bigg(\bigcup_{head\in[H]} \sum_{h_3 h\in\mathcal{H}} \sum_{(x^{-1}(h_3 x(\bar{j})+y), h_3 h_2')\in N((x^{-1}(h_3 x(\bar{i})+y), h_3 h_1'))} \sigma_{(x^{-1}(h_3 x(\bar{j})+y), h_3 h_2')}\Bigg(\langle \varphi^{head}_{qry} f(\bar{i}, h_1') + \Delta\varphi^{head}_{qry} f(\bar{i}, h_1'), \varphi^{head}_{key}(f(\bar{j}, h_2')\\
&\quad + L_h[\rho]((x^{-1}(h_3^{-1}x(\bar{i})+y), h_3 h_1'), (x^{-1}(h_3^{-1}x(\bar{j})+y), h_3 h_2')))\\
&\quad + \Delta\varphi^{head}_{key}(f(\bar{j}, h_2') + L_h[\rho]((x^{-1}(h_3^{-1}x(\bar{i})+y), h_3 h_1'), (x^{-1}(h_3^{-1}x(\bar{j})+y), h_3 h_2')))\rangle\Bigg)\\
&\quad \times (\varphi^{head}_{val}(f(\bar{j}, h_2')) + \Delta\varphi^{head}_{val}(f(\bar{j}, h_2')))\Bigg)
\end{aligned}
$$

Next, we use the definition of $L_h[\rho]((i,h_1),(j,h_2)) = \rho^P(h^{-1}(x(j)-x(i)), h_1^{-1}h_2)$ to reduce the equations further:

$$= \varphi_{out}\Bigg(\bigcup_{head\in[H]} \sum_{h_3 h\in\mathcal{H}} \sum_{(x^{-1}(h_3x(\bar{j})+y),h_3h'_2)\in N((x^{-1}(h_3x(\bar{i})+y),h_3h'_1))} \sigma_{(x^{-1}(h_3x(\bar{j})+y),h_3h'_2)}\Bigg(\langle \varphi^{head}_{qry} f(\bar{i},h'_1) + \Delta\varphi^{head}_{qry} f(\bar{i},h'_1), \varphi^{head}_{key}(f(\bar{j},h'_2)$$
$$+ \rho^P(h^{-1}(h_3x(\bar{j}) + y - (h_3x(\bar{i}) + y)), (h_3h'_1)^{-1}(h_3h'_2))$$
$$+ \Delta\varphi^{head}_{key}(f(\bar{j},h'_2) + \rho^P(h^{-1}(h_3x(\bar{j}) + y - (h_3x(\bar{i}) + y)), (h_3h'_1)^{-1}(h_3h'_2)))\rangle\Bigg)$$
$$\times (\varphi^{head}_{val}(f(\bar{j},h'_2)) + \Delta\varphi^{head}_{val}(f(\bar{j},h'_2)))\Bigg)$$
$$= \varphi_{out}\Bigg(\bigcup_{head\in[H]} \sum_{h_3 h\in\mathcal{H}} \sum_{(x^{-1}(h_3x(\bar{j})+y),h_3h'_2)\in N((x^{-1}(h_3x(\bar{i})+y),h_3h'_1))} \sigma_{(x^{-1}(h_3x(\bar{j})+y),h_3h'_2)}\Bigg(\langle \varphi^{head}_{qry} f(\bar{i},h'_1) + \Delta\varphi^{head}_{qry} f(\bar{i},h'_1), \varphi^{head}_{key}(f(\bar{j},h'_2)$$
$$+ \rho^P(h^{-1}(h_3x(\bar{j}) - h_3x(\bar{i})), {h'_1}^{-1}h'_2) + \Delta\varphi^{head}_{key}(f(\bar{j},h'_2) + \rho^P(h^{-1}(h_3x(\bar{j}) - h_3x(\bar{i})), {h'_1}^{-1}h'_2))\rangle\Bigg)$$
$$\times (\varphi^{head}_{val}(f(\bar{j},h'_2)) + \Delta\varphi^{head}_{val}(f(\bar{j},h'_2)))\Bigg)$$
$$= \varphi_{out}\Bigg(\bigcup_{head\in[H]} \sum_{h_3 h\in\mathcal{H}} \sum_{(x^{-1}(h_3x(\bar{j})+y),h_3h'_2)\in N((x^{-1}(h_3x(\bar{i})+y),h_3h'_1))} \sigma_{(x^{-1}(h_3x(\bar{j})+y),h_3h'_2)}\Bigg(\langle \varphi^{head}_{qry} f(\bar{i},h'_1) + \Delta\varphi^{head}_{qry} f(\bar{i},h'_1), \varphi^{head}_{key}(f(\bar{j},h'_2)$$
$$+ L_{h_3^{-1}h}[\rho](\bar{i},h'_1),(\bar{j},h'_2)) + \Delta\varphi^{head}_{key}(f(\bar{j},h'_2) + L_{h_3^{-1}h}[\rho](\bar{i},h'_1),(\bar{j},h'_2))\rangle\Bigg)(\varphi^{head}_{val}(f(\bar{j},h'_2)) + \Delta\varphi^{head}_{val}(f(\bar{j},h'_2)))\Bigg)$$

5. Appendix

In [RC] there are a few calculation errors in the proof of the equivariance of group self-attention layers. Below is a and explanation of why the proof of the equivariance property does not work. Let $L_g[f](i,h_1) = L_yL_{h_3}[f](i,h_1) = f(x^{-1}(h_3^{-1}(x(i)-y)), h_3^{-1}h_1)$ be a g-transformed input signal, where $g = (y,h_3) \in G = \mathbb{R}^{d_{in}} \rtimes \mathcal{H}$. The group self-attention operation on $L_g[f]$ is given by:

$$m^r_G[L_yL_{h_3}[f],\rho](i,h_1) = \varphi_{out}\Bigg(\bigcup_{head\in[H]} \sum_{h\in\mathcal{H}} \sum_{(j,h_2)\in N(i,h_1)} \sigma_{(j,h_2)}\Bigg(\langle \varphi^{head}_{qry}(L_yL_{h_3}[f](i,h_1)), \varphi^{head}_{key}(L_yL_{h_3}[f](j,h_2)$$
$$+ L_h[\rho]((i,h_1),(j,h_2)))\rangle\Bigg)\varphi^{head}_{val}(L_yL_{h_3}[f](j,h_2))\Bigg)$$
$$= \varphi_{out}\Bigg(\bigcup_{head\in[H]} \sum_{h\in\mathcal{H}} \sum_{(j,h_2)\in N(i,h_1)} \sigma_{(j,h_2)}\Bigg(\langle \varphi^{head}_{qry}(f(x^{-1}(h_3^{-1}(x(i)-y))), h_3^{-1}h_1),$$
$$\varphi^{head}_{key}(f(x^{-1}(h_3^{-1}(x(i)-y)), h_3^{-1}h_2)$$
$$+ L_h[\rho]((i,h_1),(j,h_2)))\rangle\Bigg)\varphi^{head}_{val}(f(x^{-1}(h_3^{-1}(x(i)-y)), h_3^{-1}h_2))\Bigg)$$
$$= \varphi_{out}\Bigg(\bigcup_{head\in[H]} \sum_{h_3h\in\mathcal{H}} \sum_{(x^{-1}(h_3x(\bar{j})+y),h_3h'_2)\in N((x^{-1}(h_3x(\bar{i})+y),h_3h'_1))} \sigma_{(x^{-1}(h_3x(\bar{j})+y),h_3h'_2)}\Bigg(\langle \varphi^{head}_{qry}(f(\bar{i},h'_1)), \varphi^{head}_{key}(f(\bar{j},h'_2)$$
$$+ L_h[\rho]((x^{-1}(h_3^{-1}x(\bar{i})+y),h_3h'_1),(x^{-1}(h_3^{-1}x(\bar{j})+y),h_3h'_2)))\rangle\Bigg)\varphi^{head}_{val}(f(\bar{j},h'_2))\Bigg)$$

Here, we have used $\bar{i} = x^{-1}(h_3^{-1}(x(i) - y)) \implies i = x^{-1}(h_3^{-1}x(\bar{i}) + y)$, and $\bar{j} = x^{-1}(h_3^{-1}(x(j) - y)) \implies j = x^{-1}(h_3^{-1}x(\bar{j}) + y)$, and $h'_1 = h_3^{-1}h_1$ and $h'_2 = h_3^{-1}h_2$. By using the definition of $L_h\rho((i,h_1),(j,h_2)) = \rho^P(h^{-1}(x(j) - x(i)), h^{-1}h_1^{-1}h_2)$ we can further reduce the equations:

$$
\begin{aligned}
&= \varphi_{out}\Bigg(\bigcup_{head\in[H]} \sum_{h_3h\in\mathcal{H}} \sum_{(x^{-1}(h_3x(\bar{j})+y),h_3h'_2)\in N((x^{-1}(h_3x(\bar{i})+y),h_3h'_1))} \sigma_{(x^{-1}(h_3x(\bar{j})+y),h_3h'_2)}\Big(\langle \varphi_{qry}^{head}(f(\bar{i},h'_1)), \varphi_{key}^{head}(f(\bar{j},h'_2)) \\
&\quad + \rho^P(h^{-1}(h_3x(\bar{j}) + y - (h_3x(\bar{i}) + y)), h^{-1}(h_3h'_1)^{-1}(h_3h'_2)) \rangle \Big) \varphi_{val}^{head}(f(\bar{j},h'_2)) \Bigg) \\
&= \varphi_{out}\Bigg(\bigcup_{head\in[H]} \sum_{h_3h\in\mathcal{H}} \sum_{(x^{-1}(h_3x(\bar{j})+y),h_3h'_2)\in N((x^{-1}(h_3x(\bar{i})+y),h_3h'_1))} \sigma_{(x^{-1}(h_3x(\bar{j})+y),h_3h'_2)}\Big(\langle \varphi_{qry}^{head}(f(\bar{i},h'_1)), \varphi_{key}^{head}(f(\bar{j},h'_2)) \\
&\quad + \rho^P(h^{-1}(h_3x(\bar{j}) - h_3x(\bar{i})), h^{-1}h'^{-1}_1h'_2) \rangle \Big) \varphi_{val}^{head}(f(\bar{j},h'_2)) \Bigg) \quad \text{notice the lack of a factor of } h_3^{-1} \text{ in front of } h'^{-1}_1h'_2 \text{ here} \\
&\neq \varphi_{out}\Bigg(\bigcup_{head\in[H]} \sum_{h_3h\in\mathcal{H}} \sum_{(x^{-1}(h_3x(\bar{j})+y),h_3h'_2)\in N((x^{-1}(h_3x(\bar{i})+y),h_3h'_1))} \sigma_{(x^{-1}(h_3x(\bar{j})+y),h_3h'_2)}\Big(\langle \varphi_{qry}^{head}(f(\bar{i},h'_1)), \varphi_{key}^{head}(f(\bar{j},h'_2)) \\
&\quad + L_{h_3^{-1}h}[\rho](\bar{i},h'_1),(\bar{j},h'_2)) \rangle \Big) \varphi_{val}^{head}(f(\bar{j},h'_2)) \Bigg)
\end{aligned}
$$

So, we see that $m'_G[L_yL_{h_3}[f],\rho](i,h) \neq L_yL_{h_3}[m'_G[f,\rho]](i,h)$. It is also said in [RC] that the equivariance of group self-attention is due to the relative positional encoding being **invariant** to the group action. That is, it is claimed to follow from $L_g[\rho](i,j) = \rho(i,j)$. This however is false. In particular, the positional encoding $\rho(i,j)$ is not used in the proof, it is the positional encoding $\rho((i,h_1),(j,h_2))$ that is used. Furthermore, we see that it is not G-invariant unless $x(j) - x(i)$ is an H-invariant vector.

$$
\begin{aligned}
L_g[\rho]((i,h_1),(j,h_2)) &= L_yL_h[\rho]((i,h_1),(j,h_2)) \\
&= L_h[L_y[\rho]]((i,h_1),(j,h_2)) \\
&= L_h[\rho]((x^{-1}(x(i) - y),h_1),(x^{-1}(x(j) - y),h_2)) \\
&= \rho((x^{-1}(h^{-1}(x(i) - y)),h^{-1}h_1),(x^{-1}(h^{-1}(x(j) - y)),h^{-1}h_2)) \\
&= \rho^P(h^{-1}(x(j) - y) - h^{-1}(x(i) - y),(h^{-1}h_1)^{-1}(h^{-1}h_2)) \\
&= \rho^P(h^{-1}(x(j) - x(i)),(h^{-1}h_1)^{-1}(h^{-1}h_2)) \\
&= \rho^P(h^{-1}(x(j) - x(i)),h_1^{-1}hh^{-1}h_2) \\
&= \rho^P(h^{-1}(x(j) - x(i)),h_1^{-1}h_2) \\
&\neq \rho((i,h_1),(j,h_2)) \quad \text{unless } h^{-1}(x(j) - x(i)) = x(j) - x(i)
\end{aligned}
$$

This means that group self-attention is not group equivariant, as is. The other proofs in the paper do in fact work, however. If we redefine $L_h[\rho]((i,h_1),(j,h_2))$ as $\rho^P(h^{-1}(x(j) - x(i)),h_1^{-1}h_2)$ we see that equivariance does in fact hold, and that the positional encoding is $\mathcal{H}$-invariant in the group component.

References

[Hu et. al.] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, *LoRA: Low-Rank Adaptation of Large Language Models*, https://arxiv.org/abs/2106.09685

[RC] David W. Romero, Jean-Baptiste Cordonnier, *Group Equivariant Stand-Alone Self-Attention For Vision*, https://arxiv.org/abs/2010.00977

Email address: amelie.schreiber.math@gmail.com