Appendix

A. Proof of Permutation Equivariance of Group Shuffle Attention

Lemma 1 (Permutation matrix and permutation function)**.** $\forall \Lambda \in \mathbb{R}^{N\times N}$, $\forall$ *permutation matrix* P *of size* N, $\exists p : \{1, 2, ..., N\} \to \{1, 2, ..., N\}$ *is a permutation function:*

$$\Lambda_{ij} = (P \cdot \Lambda)_{p(i)j} = (\Lambda \cdot P^T)_{ip(j)} = (P \cdot \Lambda \cdot P^T)_{p(i)p(j)}. \quad (15)$$

Lemma 2. *Given* $A \in \mathbb{R}^{N\times N}$, $\forall$ *permutation matrix* P *of size* N,

$$softmax(PAP^T) = Psoftmax(A)P^T. \quad (16)$$

Proof. $softmax(A)_{ij} = \frac{e^{A_{ij}}}{\sum_{n=1}^{N} e^{A_{ij}}}$.

Consider a permutation function p for P, using Eq. 15, we get:

$$\begin{aligned}(P \cdot softmax(A) \cdot P^T)_{p(i)p(j)} &= softmax(A)_{ij} \\ &= \frac{e^{A_{ij}}}{\sum_{n=1}^{N} e^{A_{ij}}} \\ &= \frac{e^{(PAP^T)_{p(i)p(j)}}}{\sum_{n=1}^{N} e^{(PAP^T)_{p(i)p(j)}}} \\ &= softmax(P \cdot A \cdot P^T)_{p(i)p(j)}.\end{aligned}$$

which implies $P \cdot softmax(A) \cdot P^T = softmax(P \cdot A \cdot P^T)$. □

Lemma 3. *Non-linear self-attention* $Attn_\sigma(Q, X) = S(Q, X) \cdot \sigma(X)$*, with* $S(Q, X) = softmax(QX^T/\sqrt{c})$ *is permutation-equivariant.*

Proof. For a self-attention, $Q = X$.

σ is an element-wise function, thus $\sigma(P \cdot X) = P \cdot \sigma(X)$.

$$\begin{aligned}Attn_\sigma(PX, PX) &= S(PX, PX) \cdot \sigma(PX) \\ &= softmax(P \cdot XX^T \cdot P^T/\sqrt{c}) \cdot P \cdot \sigma(X) \\ &= P \cdot softmax(XX^T/\sqrt{c}) \cdot P^T \cdot P \cdot \sigma(X) \\ &= P \cdot softmax(XX^T/\sqrt{c}) \cdot \cdot \sigma(X) \\ &= P \cdot Attn_\sigma(X, X).\end{aligned}$$

which implies non-linear self-attention is permutation-equivariant. □

Proposition 1. *The Group Shuffle Attention operation is permutation-equivariant,* i.e.*, given input* $X \in \mathbb{R}^{N\times c}$*,* $\forall$ *permutation matrix* P *of size* N*,*

$$GSA(P \cdot X) = P \cdot GSA(X).$$

Proof.

$$GSA(X) = \mathcal{GN}(\psi(GroupAttn(X)) + X), \quad (11)$$

where,

$$GroupAttn(X) = concat\{Attn_\sigma(X_i, X_i) \mid X_i = X^{(i)}W_i\}_{i=1,...,g}. \quad (8)$$

GSA only introduces element-wise operations, which does not change the permutation-equivariance of $Attn_\sigma$. □

B. Proof of Permutation Invariance of Gumbel Subset Sampling

Proposition 1. *The Gumbel Subset Sampling operation is permutation-invariant,* i.e.*, given input* $X \in \mathbb{R}^{N\times c}$*,* $\forall$ *permutation matrix* P *of size* N*,*

$$GSS(P \cdot X) = GSS(X).$$

Proof. According to Eq. 14,

$$GSS(X) = gumbel\_softmax(WX^T) \cdot X, \quad W \in \mathbb{R}^{N\times c}.$$

Similar to Lemma 2,

$$softmax(AP^T) = softmax(A)P^T. \quad (17)$$

Since *gumbel_softmax* add element-wise operations on *softmax*, it does not change the permutation property. In this way,

$$\begin{aligned}GSS(P \cdot X) &= gumbel\_softmax(WX^TP^T) \cdot PX \\ &= gumbel\_softmax(WX^T)P^T \cdot PX \\ &= gumbel\_softmax(WX^T) \cdot X \\ &= GSS(X).\end{aligned}$$

□