Set Transformer: A Framework for Attention-based Permutation-Invariant Neural Networks

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam R. Kosiorak, Seungjin Choi, and Yee Whye Teh
Set-input problems and Deep Sets [Zaheer et al., 2017]

- Take sets (variable lengths, order does not matter) as inputs

- Application includes multiple instance learning, point-cloud classification, few-shot image classification, etc.

- Deep Sets: a simple way to construct permutation invariant set-input neural networks, but does not effectively modeling interactions between elements in sets.

\[ f(X) = \rho \left( \sum_{x \in X} \phi(x) \right). \]
Attention based set operations

• Use multihead self-attention [Vaswani et al., 2017] to encode interactions between elements in a set.

\[
\begin{align*}
Q &= X W_q \\
K &= Y W_k \\
V &= Y W_v
\end{align*}
\]

\[
\begin{align*}
X &\rightarrow q_1^{T} \\
&\rightarrow q_2^{T} \\
&\rightarrow \vdots \\
&\rightarrow q_n^{T}
\end{align*}
\]

\[
\begin{align*}
Y &\rightarrow k_1^{T} \\
&\rightarrow k_2^{T} \\
&\rightarrow \vdots \\
&\rightarrow k_m^{T}
\end{align*}
\]

\[
\begin{align*}
X &\rightarrow v_1^{T} \\
&\rightarrow v_2^{T} \\
&\rightarrow \vdots \\
&\rightarrow v_m^{T}
\end{align*}
\]

\[
\begin{align*}
\text{SelfAtt}(X) &= \text{Att}(X, X) \\
\text{Att}(X, Y) &= \text{softmax}\left(\frac{X W_q W_k^{T} Y^{T}}{\sqrt{d}}\right) Y W_v
\end{align*}
\]

• Note that a self-attention is permutation equivariant,

\[
\text{SelfAtt}(\pi \cdot X) = \pi \cdot \text{SelfAtt}(X)
\]
Set transformer - building blocks

- Multihead attention block (MAB): residual connection + multihead QKV attention followed by a feed-forward layer
  \[ \text{MAB}(X, Y) = \text{FFN}(WX + \text{Att}(X, Y)) \cdot \]
- Self attention block (SAB): MAB applied in self-attention way, \( O(n^2) \)
  \[ \text{SAB}(X) = \text{MAB}(X, X) \cdot \]
- Induced self-attention block (ISAB): introduce a set of trainable inducing points to simulate self-attention, \( O(nm) \) with \( m \) inducing points.
  \[ \text{ISAB}(X) = \text{MAB}(X, \text{MAB}(I, X)) \cdot \]
Set transformer - building blocks

- Pooling by multihead attention (PMA): instead of a simple sum/max/min aggregation, use multihead attention to aggregate features into a single vector.

- Introduce a trainable seed vector, and use it to produce one output vector.

\[ o = \text{PMA}_1(Z) = \text{MAB}(s, Z) \]

- Use multiple seed vectors and apply self-attention to produce multiple interacting outputs (e.g., explaining away)

\[ O = \text{SelfAtt}(\text{PMA}_k(Z)) = \text{SelfAtt}(\text{MAB}(S, Z)) \quad S = [s_1^T, \ldots, s_k^T] \]
Set transformer - architecture

- **Encoder**: a stack of permutation-equivariant ISABs.

```
X
\[ x_1^T \]
\[ x_2^T \]
\[ \vdots \]
\[ x_n^T \]

\[ \rightarrow \] ISAB_1 \rightarrow ISAB_2 \rightarrow \cdots \rightarrow ISAB_L \rightarrow Z
\[ z_1^T \]
\[ z_2^T \]
\[ \vdots \]
\[ z_n^T \]
```

- **Decoder**: PMA followed by self-attention to produce outputs.

```
s_1^T
\[ \vdots \]
\[ s_k^T \]

\[ \rightarrow \] MAB(S, X) \rightarrow SAB_1 \rightarrow \cdots \rightarrow SAB_L \rightarrow \[ o_1^T \]
\[ \vdots \]
\[ o_k^T \]
```
Experiments

- Amortized clustering - learn a mapping from dataset to clustering

Deep Sets

Set transformer
Experiments

- Works well for various tasks such as unique character counting, amortized clustering, point cloud classification, and anomaly detection

- Generalize well with small number of inducing points

- Attentions both in encoder (ISAB) and decoder (PMA + SAB) are important for the performance.
Conclusion

• New set-input neural network architecture

• Can efficiently model pairwise/higher order interactions between elements in sets

• Demonstrated to work well for various set-input tasks

• Code available at https://github.com/juho-lee/set_transformer
References