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

Juho Lee, Yoonho Lee, **Jungtaek Kim**, Adam R. Kosiorek, Seungjin Choi, and Yee Whye Teh

AITRICS

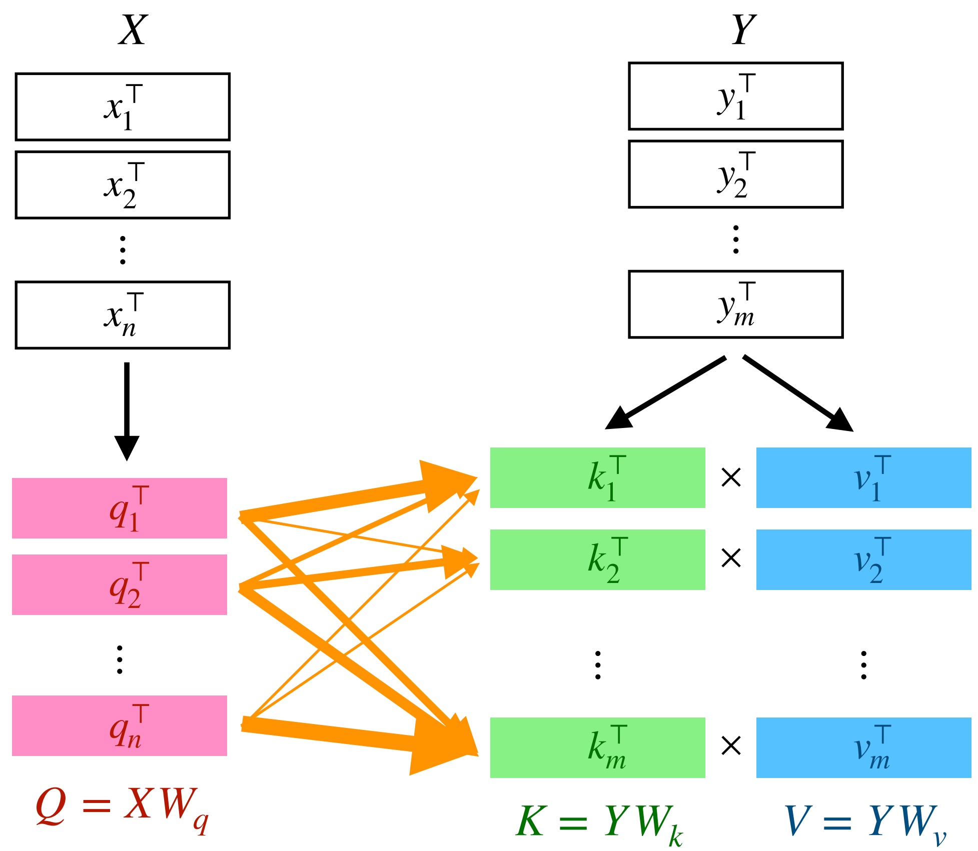
Set-input problems and Deep Sets [Zaheer et al., 2017]

- 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 nerual 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.



$$\mathrm{Att}(X, Y) = \mathrm{softmax}\left(\frac{XW_qW_k^\top Y^\top}{\sqrt{d}}\right)YW_v\,.$$

$$\mathrm{SelfAtt}(X) = \mathrm{Att}(X, X)\,.$$

- Note that a self-attention is **permutation equivariant**,

$$\mathrm{SelfAtt}(\pi \cdot X) = \pi \cdot \mathrm{SelfAtt}(X)$$

Set transformer - building blocks

- Multihead attention block (MAB): residual connection + multihead QKV attention followed by a feed-forward layer

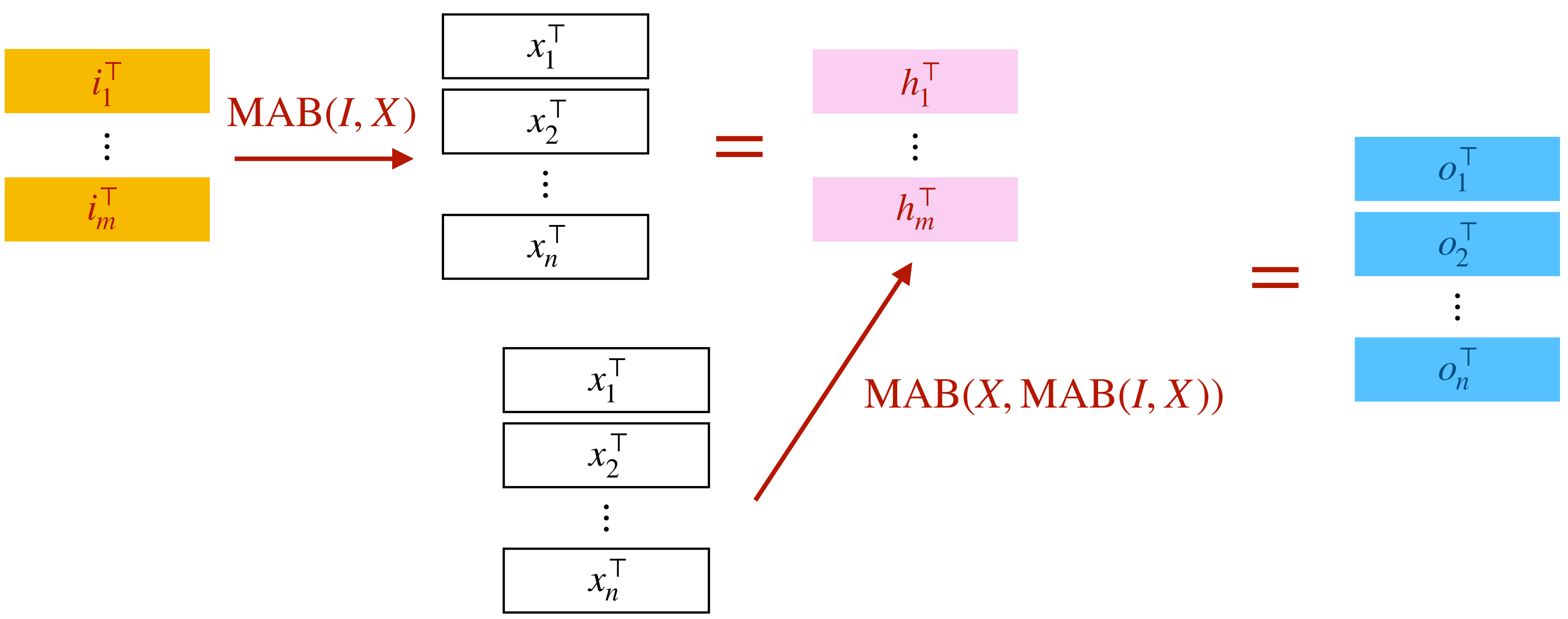
$$\mathrm{MAB}(X, Y) = \mathrm{FFN}(WX + \mathrm{Att}(X, Y)) \, .$$

- Self attention block (SAB): MAB applied in self-attention way, $O(n^2)$

$$\mathrm{SAB}(X) = \mathrm{MAB}(X, X) \, .$$

- Induced self-attention block (ISAB): introduce a set of trainable inducing points to simulate self-attention, $O(nm)$ with m inducing points.

$$\mathrm{ISAB}(X) = \mathrm{MAB}(X, \mathrm{MAB}(I, X)) \, .$$



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.

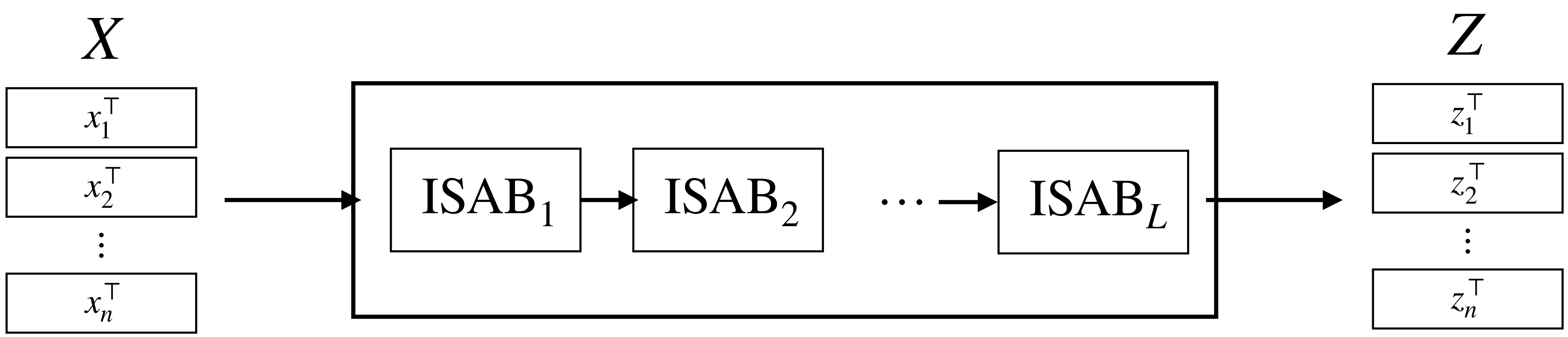
$$o = \mathrm{PMA}_1(Z) = \mathrm{MAB}(s, Z)$$

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

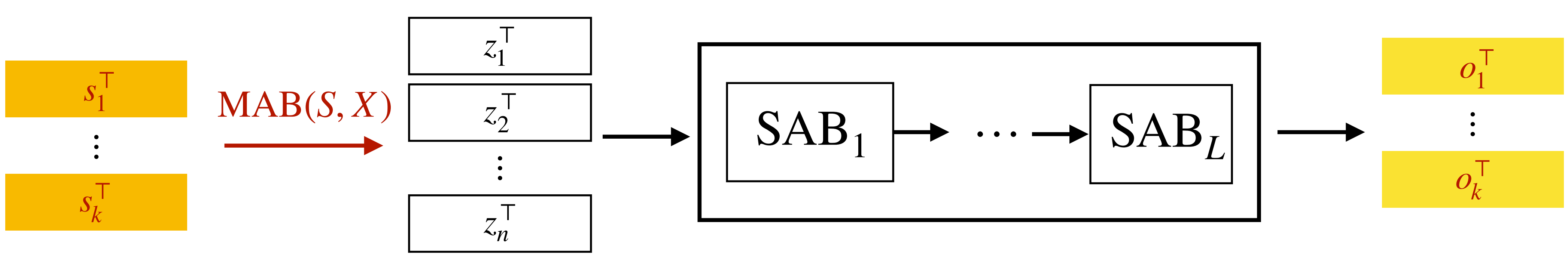
$$O = \mathrm{SelfAtt}(\mathrm{PMA}_k(Z)) = \mathrm{SelfAtt}(\mathrm{MAB}(S, Z)) \quad S = [s_1^\top, \ldots, s_k^\top] \, .$$

Set transformer - architecture

- Encoder: a stack of permutation-equivarinat ISABs.

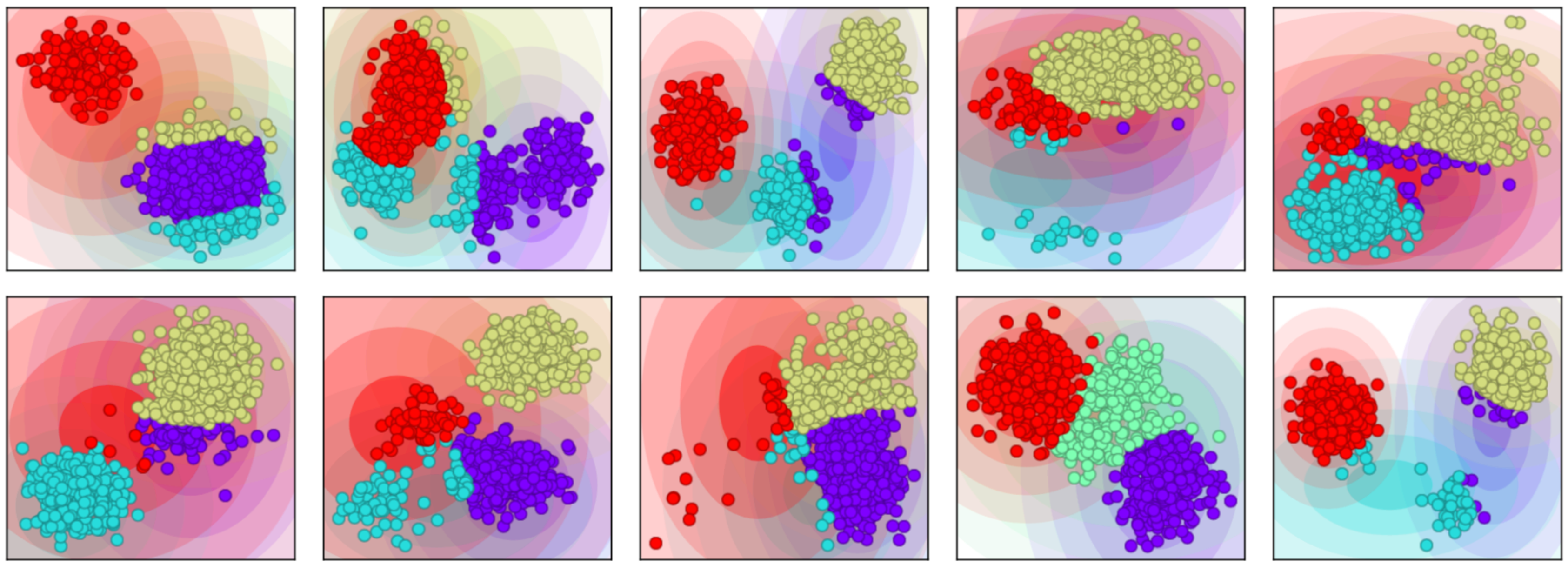


- Decoder: PMA followed by self-attention to produce outputs.

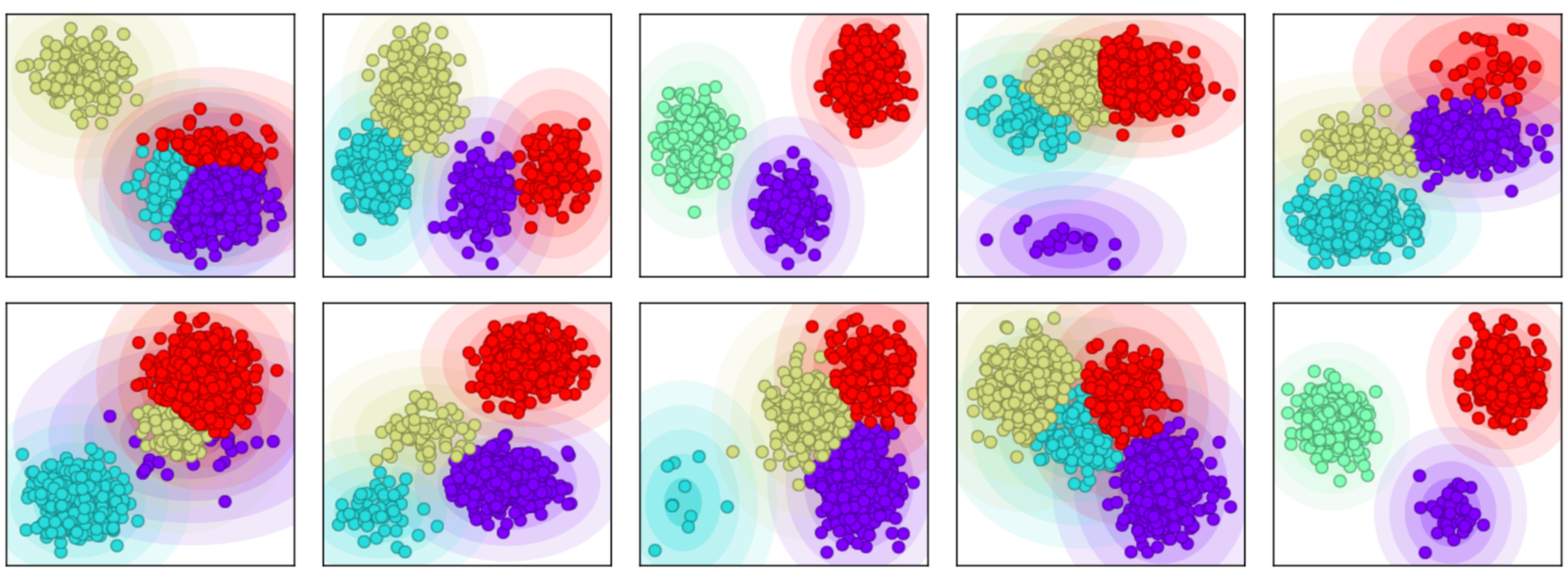


Experiments

- 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

[Qi et al., 2017] Qi, R. C., Su, H., Mo, K., and Guibas, J. L. PointNet: Deep learning on point sets for 3D classification and segmentation. CVPR, 2017.
[Vinyals et al., 2016] Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., and Wierstra, D. Matching networks for one shot learning. NIPS, 2016.
[Zaheer et al., 2017] Zaheer, M., Kottur, S., Ravanbhakhsh, S., Póczos, B., Salakhutdinov, R., and Smola, A. J. Deep sets. NIPS, 2017.
[Wagstaff et al, 2019] Wagstaff, E., Fuchs, F. B., Engelcke, M., Posner, I., and Osborne, M. On the limitations of representing functions on sets. arXiv:1901.09006, 2019.
[Cybenko 1989] Cybenko, G. Approximation by superpositions of sigmoidal functions. Mathematics of Control, Signals, and Systems, 2(4), 303314, 1989.
[Shi et al., 2015] Shi, B., Bai, S., Zhou, Z., and Bai, X. DeepPano: deep panoramic representation for 3-D shape recognition. IEEE Signal Processing Letters, 22(12):2339–2343, 2015.
[Su et al., 2015] Su, H., Maji, S., Kalogerakis, E., and Learned-Miller, E. Multi-view convolutional neural networks for 3d shape recognition. ICCV, 2015.
[Snell et al., 2017] Snell, J., Swersky, K., and Zemel, R. Prototypical networks for few-shot learning. NIPS, 2017.
[Ilse et al., 2018] Ilse, M., Tomczak, J. M., and Welling, M. Attention-based deep multiple instance learning. ICML, 2018.
[Garnelo et al., 2018] Garnelo, M., Rosenbaum, D., Maddison, C. J., Ramalho, T., Saxton, D., Shanahan, M., Teh, Y. W., Rezende, D. J., and Eslami, S. M. A. ICML, 2018.
[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. NIPS, 2017.