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

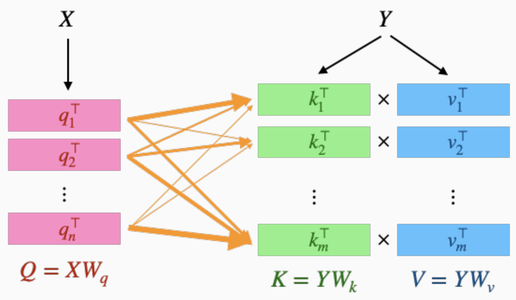
Juho Lee[1], Yoonho Lee[2], Jungtaek Kim[3], Adam R. Kosiorek[4], Seungjin Choi[3], and Yee Whye Teh[4,5]

[1]AITRICS, [2]KaKao Corporation, [3]Pohang University of Science and Technology, [4]University of Oxford [5]Deepmind juho@aitrics.com

Permutation invariant neural networks

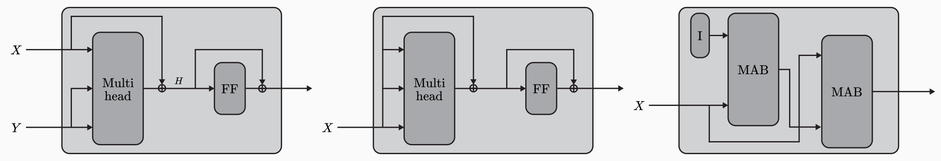
- A function $f : 2^{\mathcal{X}} \to \mathcal{Y}$ is permutation invariant if
$$f(x_1, \ldots, x_n) = f(x_{\pi(1)}, \ldots, x_{\pi(n)}),$$
for any permutation π of $\{1, \ldots, n\}$.
- Deep Sets [1]
$$f(x_1, \ldots, x_n) = \rho(\mathrm{pool}(\phi(x_1), \ldots, \phi(x_n))),$$
where ϕ is an encoder, pool is an aggregation operation such as sum, and ρ is a decoder.
- Limitation of the Deep Sets: not effective in modeling interactions between elements.

QKV Attention mechanism [2]



- Easy and efficient way of computing interaction between sets
- Attending $X = [x_1^\top, \ldots, x_n^\top]$ to $Y = [y_1^\top, \ldots, y_m^\top]$
$$\mathrm{Att}(X, Y) = \mathrm{softmax}\Big((XW_q)(YW_k)^\top/\sqrt{d}\Big)YW_v.$$
- Multihead attention: apply parallel attentions and concatenate.
$$\mathrm{mAtt}_k(X, Y) = [\mathrm{Att}(X_1, Y_1), \ldots, \mathrm{Att}(X_k, Y_k)].$$
- Self-attention: encode pairwise interactions via $\mathrm{Att}(X, X)$.

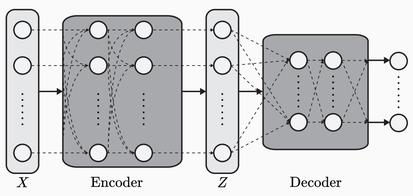
Set transformer - building blocks



(a) MAB (b) SAB (c) ISAB

- Multihead Attention Block (MAB) and Self Attention Block (SAB), permutation equivariant set-operations.
$$\mathrm{MAB}(X, Y) = \mathrm{LayerNorm}(H + \mathrm{rFF}(H)),$$
$$\text{where } H = \mathrm{LayerNorm}(X + \mathrm{mAtt}_k(X, Y))$$
$$\mathrm{SAB}(X) = \mathrm{MAB}(X, X),$$
- Induced Self-Attention Block (ISAB)
  - Inspired by sparse Gaussian process [4], i ntroduce a set of trainable m inducing points I to define
$$\mathrm{ISAB}_m(X) = \mathrm{MAB}(X, \mathrm{MAB}(I, X)).$$
  - I can be interpreted as a memory cell to write about X.
  - Reduced computation cost $O(n^2)$ to $O(nm)$.
- Pooling by Multihead Attention (PMA)
$$\mathrm{PMA}_k(Z) = \mathrm{MAB}(S, \mathrm{rFF}(Z)).$$
  - Attentional aggregation of encoded features
  - Attend k seed points to the features Z to produce k outputs

Set transformer - architecture



- Encoder: stack of (induced) SABs, encode pairwise and even higher-order interactions between elements in sets.
$$\mathrm{Encoder}(X) = \mathrm{ISAB}_m(\ldots(\mathrm{ISAB}_m(X)\ldots).$$
- Decoder: aggregate features via PMA, and apply self-attentions to outputs to model dependencies between multiple outputs.
$$\mathrm{Decoder}(Z) = \mathrm{SAB}(\ldots\mathrm{SAB}(\mathrm{PMA}_k(Z))\ldots).$$

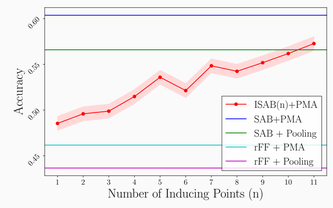
Toy experiments

- Max regression: finding maximum number in a given set.

Architecture	MAE
rFF + Pooling (mean)	2.133 ± 0.190
rFF + Pooling (sum)	1.902 ± 0.137
rFF + Pooling (max)	**0.1355** ± 0.0074
SAB + PMA (ours)	0.2085 ± 0.0127

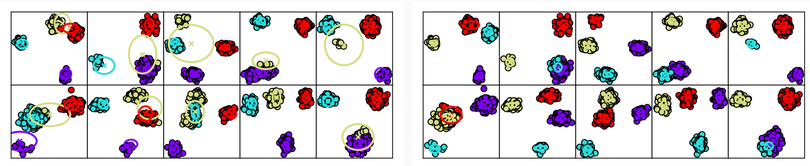
- Counting unique characters in given set of Omniglot images.

Architecture	Accuracy
rFF + Pooling	0.4382 ± 0.0072
rFFp-mean + Pooling	0.4617 ± 0.0076
rFFp-max + Pooling	0.4359 ± 0.0077
rFF + Dotprod	0.4471 ± 0.0076
rFF + PMA (ours)	0.4572 ± 0.0076
SAB + Pooling (ours)	0.5659 ± 0.0077
SAB + PMA (ours)	**0.6037** ± 0.0075



Amortized clustering via mixture of Gaussians

- Learn a set-input neural network f that maximizes
$$\mathbb{E}_X\left[\sum_{x\in X}\sum_{j=1}^{k}\pi_j(X)\mathcal{N}(x_i; \mu_j(X), \mathrm{diag}(\sigma_j^2(X)))\right].$$
- Attentions in both encoder and decoder are crucial.



(d) DeepSets (e) ours

Architecture	Synthetic		CIFAR-100	
	LL0/data	LL1/data	ARI0	ARI1
Oracle	-1.4726		0.9150	
rFF + Pooling	-2.0006 ± 0.0123	-1.6186 ± 0.0042	0.5593 ± 0.0149	0.5693 ± 0.0171
rFFp-mean + Pooling	-1.7606 ± 0.0213	-1.5191 ± 0.0026	0.5673 ± 0.0053	0.5798 ± 0.0058
rFFp-max + Pooling	-1.7692 ± 0.0130	-1.5103 ± 0.0035	0.5369 ± 0.0154	0.5536 ± 0.0186
rFF + Dotprod	-1.8549 ± 0.0128	-1.5621 ± 0.0046	0.5666 ± 0.0221	0.5763 ± 0.0212
SAB + Pooling (ours)	-1.6772 ± 0.0066	-1.5070 ± 0.0115	0.5831 ± 0.0341	0.5943 ± 0.0337
ISAB (16) + Pooling (ours)	-1.6955 ± 0.0730	-1.4742 ± 0.0158	0.5672 ± 0.0124	0.5805 ± 0.0122
rFF + PMA (ours)	-1.6680 ± 0.0040	-1.5409 ± 0.0037	0.7612 ± 0.0237	0.7670 ± 0.0231
SAB + PMA (ours)	-1.5145 ± 0.0046	-1.4619 ± 0.0048	0.9015 ± 0.0097	0.9024 ± 0.0097
ISAB (16) + PMA (ours)	**-1.5009** ± 0.0068	**-1.4530** ± 0.0037	**0.9210** ± 0.0055	**0.9223** ± 0.0056

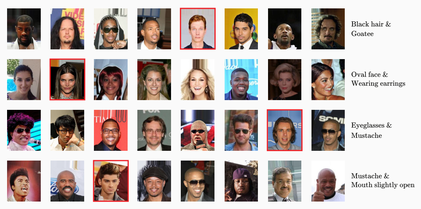
Point-cloud classification

- Classifying a set of points (3D coordinates that describes objects) into one of 40 categories.

Architecture	100 pts	1000 pts	5000 pts
rFF + Pooling [1]	-	0.83 ± 0.01	-
rFFp-max + Pooling [1]	0.82 ± 0.02	0.87 ± 0.01	**0.90** ± 0.003
rFF + Pooling	0.7951 ± 0.0166	0.8551 ± 0.0142	0.8933 ± 0.0156
rFF + PMA (ours)	0.8076 ± 0.0160	0.8534 ± 0.0152	0.8628 ± 0.0136
ISAB (16) + Pooling (ours)	0.8273 ± 0.0159	**0.8915** ± 0.0144	**0.9040** ± 0.0173
ISAB (16) + PMA (ours)	**0.8454** ± 0.0144	0.8662 ± 0.0149	0.8779 ± 0.0122

Set-anomaly detection

- Detecting anomalies in given sets.



Architecture	Test AUROC	Test AUPR
Random guess	0.5	0.125
rFF + Pooling	0.5643 ± 0.0139	0.4126 ± 0.0108
rFFp-mean + Pooling	0.5687 ± 0.0061	0.4125 ± 0.0127
rFFp-max + Pooling	0.5717 ± 0.0117	0.4135 ± 0.0162
rFF + Dotprod	0.5671 ± 0.0139	0.4155 ± 0.0115
SAB + Pooling (ours)	0.5757 ± 0.0143	0.4189 ± 0.0167
rFF + PMA (ours)	0.5756 ± 0.0130	0.4227 ± 0.0127
SAB + PMA (ours)	**0.5941** ± 0.0170	**0.4386** ± 0.0089

Conclusion

- Attention-based permutation invariant architecture.
- Modeling interactions between input elements and multiple outputs is important.
- ISAB scales well, and works well with small number of inducing points.
- https://github.com/juho-lee/set_transformer.

References

[1] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan R. Salakhutdinov, and Alexander J. Smola. Deep sets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv e-prints*, arXiv:1607.06450, 2016.

[4] Edward Snelson and Zoubin Ghahramani. Sparse Gaussian processes using pseudo-inputs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2005.