A Trainable Optimal Transport Embedding for Feature Aggregation

Grégoire Mialon*, Dexiong Chen*, Alexandre d'Aspremont and Julien Mairal

INVENTEURS DU MONDE NUMÉRIQUE

NEURAL INFORMATION PROCESSING SYSTEMS

Motivation: biological sequences

Handling datasets of sets with positional information (*e.g*, biological sequences or sentences in natural language) requires addressing different problems:

Motivation: biological sequences

Handling datasets of sets with positional information (*e.g*, biological sequences or sentences in natural language) requires addressing different problems:

- Long-range and potentially complex dependencies between elements of the set.

Motivation: biological sequences

Handling datasets of sets with positional information (*e.g*, biological sequences or sentences in natural language) requires addressing different problems:

- Long-range and potentially complex dependencies between elements of the set.
- Varying size of the sequences.

Motivation: biological sequences

Handling datasets of sets with positional information (*e.g*, biological sequences or sentences in natural language) requires addressing different problems:

- Long-range and potentially complex dependencies between elements of the set.
- Varying size of the sequences.

We are mainly interested in biological sequences, which often pose two more problems:

`ATTKLGPKDLVCIQHNPTDRAGQGEQFQLH`

Figure 1: Protein sequence (each symbol represents an amino-acid).

Motivation: biological sequences

Handling datasets of sets with positional information (*e.g*, biological sequences or sentences in natural language) requires addressing different problems:

- Long-range and potentially complex dependencies between elements of the set.
- Varying size of the sequences.

We are mainly interested in biological sequences, which often pose two more problems:

- Long sequences (1000+ base pairs).

`ATTKLGPKDLVCIQHNPTDRAGQGEQFQLH`

Figure 1: Protein sequence (each symbol represents an amino-acid).

Motivation: biological sequences

Handling datasets of sets with positional information (*e.g*, biological sequences or sentences in natural language) requires addressing different problems:

- Long-range and potentially complex dependencies between elements of the set.
- Varying size of the sequences.

We are mainly interested in biological sequences, which often pose two more problems:

- Long sequences (1000+ base pairs).
- Few labeled data (e.g, 20 labels per class for SCOP1.75).

`ATTKLGPKDLVCIQHNPTDRAGQGEQFQLH`

Figure 1: Protein sequence (each symbol represents an amino-acid).

Current models are not adapted

Popular families of models for sets:

Current models are not adapted

Popular families of models for sets:

- Standard kernel methods [Lyu, 2004]: hand-crafted, lack of adaptivity.

Current models are not adapted

Popular families of models for sets:

- Standard kernel methods [Lyu, 2004]: hand-crafted, lack of adaptivity.
- Neural networks with attention mechanism [Bahdanau et al., 2015, Vaswani et al., 2017], and/or designed for sets [Lee et al., 2019]: possible memory issues with long sequences, performance drop when trained on few data.

Current models are not adapted

Popular families of models for sets:

- Standard kernel methods [Lyu, 2004]: hand-crafted, lack of adaptivity.
- Neural networks with attention mechanism [Bahdanau et al., 2015, Vaswani et al., 2017], and/or designed for sets [Lee et al., 2019]: possible memory issues with long sequences, performance drop when trained on few data.

We need a **trainable** embedding for sets with **lower memory/sample requirements**.

Idea: attention with optimal transport and kernel methods

We provide an embedding with an inductive bias akin to that of self-attention. Two steps:

Figure 2: The input point cloud x is transported onto the reference $\mathrm{z} = (\mathrm{z}_1, \dots, \mathrm{z}_p)$ (left), yielding the optimal transport plan $\mathrm{P}_\kappa(\mathrm{x}, \mathrm{z})$ used to aggregate the embedded features and form $\Phi_\mathrm{z}(\mathrm{x})$ (right).

Idea: attention with optimal transport and kernel methods

We provide an embedding with an inductive bias akin to that of self-attention. Two steps:

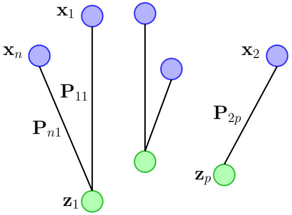
1. Non-linear layer: we use a parametrized kernel embedding in the fashion of [Chen et al., 2019a].

Figure 2: The input point cloud x is transported onto the reference $z = (z_1, \dots, z_p)$ (left), yielding the optimal transport plan $P_\kappa(x, z)$ used to aggregate the embedded features and form $\Phi_z(x)$ (right).

Idea: attention with optimal transport and kernel methods

We provide an embedding with an inductive bias akin to that of self-attention. Two steps:

1. Non-linear layer: we use a parametrized kernel embedding in the fashion of [Chen et al., 2019a].
2. Pooling: similar elements are **pooled** together. The measure of similarity is the optimal **transport plan** between the input set $x \in \mathbb{R}^{n \times d}$ and a **learned reference** $z \in \mathbb{R}^{p \times d}$.



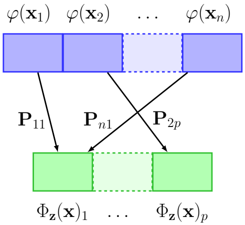


Figure 2: The input point cloud x is transported onto the reference $z = (z_1, \ldots, z_p)$ (left), yielding the optimal transport plan $P_\kappa(x, z)$ used to aggregate the embedded features and form $\Phi_z(x)$ (right).

Results

The resulting (non-standard) kernel formulation provides a rich representation for sequences with **relatively few parameters** that can be trained end-to-end or without supervision.

Table 1: Classification accuracy (top 1/5/10) on test set for SCOP 1.75 for different unsupervised and supervised baselines, averaged from 10 different runs. (q references $\times$ p supports).

Method	Unsupervised	Supervised
DeepSF [Hou et al., 2019]	Not available.	73.0/90.3/94.5
CKN [Chen et al., 2019a]	81.8±0.8/92.8±0.2/95.0±0.2	84.1±0.1/94.3±0.2/96.4±0.1
RKN [Chen et al., 2019b]	Not available.	85.3±0.3/95.0±0.2/96.5±0.1
Set Transformer [Lee et al., 2019]	Not available.	79.2±4.6/91.5±1.4/94.3±0.6
Approximate Rep the Set [Skianis et al., 2020]	Not available.	84.5±0.6/94.0±0.4/95.7±0.4
Ours (dot-product instead of OT)	78.2±1.9/93.1±0.7/96.0±0.4	87.5±0.3/95.5±0.2/96.9±0.1
Ours (Unsup.: 1×100 / Sup.: 5×10)	**85.8±0.2/95.3±0.1/96.8±0.1**	**88.7±0.3/95.9±0.2/97.3±0.1**

Conclusion

- Relationship to self-attention.

Conclusion

- Relationship to self-attention.
- Results for other bioinformatics tasks, natural language processing and computer vision can be found in the longer version of our paper `https://arxiv.org/abs/2006.12065`.

Conclusion

- Relationship to self-attention.
- Results for other bioinformatics tasks, natural language processing and computer vision can be found in the longer version of our paper `https://arxiv.org/abs/2006.12065`.
- Our code is freely available at `https://github.com/claying/OTK`.

Conclusion

- Relationship to self-attention.
- Results for other bioinformatics tasks, natural language processing and computer vision can be found in the longer version of our paper `https://arxiv.org/abs/2006.12065`.
- Our code is freely available at `https://github.com/claying/OTK`.

Thank you!

References I

Bahdanau, D., Cho, K., and Bengio, J. (2015). Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations (ICLR)*.

Chen, D., Jacob, L., and Mairal, J. (2019a). Biological sequence modeling with convolutional kernel networks. *Bioinformatics*, pages 35(18):3294–3302.

Chen, D., Jacob, L., and Mairal, J. (2019b). Recurrent kernel networks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Hou, J., Adhikari, B., and Cheng, J. (2019). Deepsf: deep convolutional neural network for mappingprotein sequences to folds. *Bioinformatics*, pages 34(8):1295–1303.

Lee, J., Lee, Y., Kim, J., Kosiorek, A. R., Choi, S., and Teh, Y. W. (2019). Set transformer: A framework for attention-based permutation invariant neural networks. In *International Conference on Machine Learning (ICML)*.

References II

Lyu, S. (2004). Mercer kernels for object recognition with local features. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Skianis, K., Nikolentzos, G., Limnios, S., and Vazirgiannis, M. (2020). Rep the set: Neural networks for learning set representations. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*.