

A Trainable Optimal Transport Embedding for Feature Aggregation and its Relationship to Attention

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



ICLR

Motivation: representation learning for biological sequences

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

Motivation: representation learning for 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: representation learning for 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: representation learning for 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:

LDKVEAEVQIDRLITG

Figure 1: Short part of mRNA sequence for the SARS-Cov-2 spike protein (each symbol represents an amino-acid).

Motivation: representation learning for 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).

LDKVEAEVQIDRLITG

Figure 1: Short part of mRNA sequence for the SARS-Cov-2 spike protein (each symbol represents an amino-acid).

Motivation: representation learning for 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).

LDKVEAEVQIDRLITG

Figure 1: Short part of mRNA sequence for the SARS-Cov-2 spike protein (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 embed a sequence (e.g point cloud of n-gram features in the bioinformatics case) $\mathbf{x} \in \mathbb{R}^{n \times d}$ in the following way:

$$\Phi_{\mathbf{z}}(\mathbf{x}) := \sqrt{p} \times P(\mathbf{x}, \mathbf{z})^{\top} \varphi(\mathbf{x}),$$

- $\varphi(\mathbf{x}) := [\varphi(x_1), \dots, \varphi(x_n)]^{\top} \in \mathbb{R}^{n \times d}$ is a non-linear embedding.
- $P(\mathbf{x}, \mathbf{z}) \in \mathbb{R}^{n \times p}$ is the optimal transport plan between \mathbf{x} and \mathbf{z} . We will use the opposite of the p.d. kernel κ associated to φ .

Idea: attention with optimal transport and kernel methods

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



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

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



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

Connection to Linear Optimal Transport and embedding learning

- Our references z can be learned without supervision, e.g via K-means, or with supervision, as Sinkhorn's algorithm allows back-propagation.

Connection to Linear Optimal Transport and embedding learning

- Our references z can be learned without supervision, *e.g.* via K-means, or with supervision, as Sinkhorn's algorithm allows back-propagation.
- Our pooling method is tightly connected to the Linear Optimal Transport Embedding method (LOT), see, *e.g.* [Moosmüller and Cloninger, 2020].

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 \pm 0.8/92.8 \pm 0.2/95.0 \pm 0.2	84.1 \pm 0.1/94.3 \pm 0.2/96.4 \pm 0.1
RKN [Chen et al., 2019b]	Not available.	85.3 \pm 0.3/95.0 \pm 0.2/96.5 \pm 0.1
Set Transformer [Lee et al., 2019]	Not available.	79.2 \pm 4.6/91.5 \pm 1.4/94.3 \pm 0.6
Approximate Rep the Set [Skianis et al., 2020]	Not available.	84.5 \pm 0.6/94.0 \pm 0.4/95.7 \pm 0.4
Ours (dot-product instead of OT)	78.2 \pm 1.9/93.1 \pm 0.7/96.0 \pm 0.4	87.5 \pm 0.3/95.5 \pm 0.2/96.9 \pm 0.1
Ours (Unsup.: 1 \times 100 / Sup.: 5 \times 10)	85.8\pm0.2/95.3\pm0.1/96.8\pm0.1	88.7\pm0.3/95.9\pm0.2/97.3\pm0.1

Conclusion

- Relationship to self-attention.

Conclusion

- Relationship to self-attention.
- Multi-head extension.

Conclusion

- Relationship to self-attention.
- Multi-head extension.
- Results for other bioinformatics tasks, natural language processing and computer vision.

Conclusion

- Relationship to self-attention.
- Multi-head extension.
- Results for other bioinformatics tasks, natural language processing and computer vision.
- 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). Deepssf: deep convolutional neural network for mapping protein 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)*.
- Moosmüller, C. and Cloninger, A. (2020). Linear optimal transport embedding: Provable fast wasserstein distance computation and classification for nonlinear problems.
- 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)*.