Hindawi
Shock and Vibration
Volume 2021, Article ID 3272119, 10 pages
https://doi.org/10.1155/2021/3272119

Research Article

The Novel Sequence Distance Measuring Algorithm Based on Optimal Transport and Cross-Attention Mechanism

Yanmin Yu,[1,2] Yongcai Lai ,[1] Ping Yan,[2] and Haiying Liu[1,2]

[1]*Heilongjiang Academy of Agricultural Sciences Postdoctoral Programme, Harbin 150086, China*
[2]*Biotechnology Institute of Heilongjiang Academy of Agricultural Sciences, Harbin 150028, China*

Correspondence should be addressed to Yongcai Lai; sws@haas.cn

Received 9 July 2021; Revised 3 August 2021; Accepted 11 August 2021; Published 31 August 2021

Academic Editor: Chaoqun Duan

In this paper, we propose a novel sequence distance measuring algorithm based on optimal transport (OT) and cross-attention mechanism. Given a source sequence and a target sequence, we first calculate the ground distance between each pair of source and target terms of the two sequences. The ground distance is calculated over the subsequences around the two terms. We firstly pay attention from each the source terms to each target terms with attention weights, so that we have a representative source subsequence vector regarding each term in the target subsequence. Then, we pay attention from each representative vector of the term of the target subsequence to the entire source subsequence. In this way, we construct the cross-attention weights and use them to calculate the pairwise ground distances. With the ground distances, we derive the OT distance between the two sequences and train the attention parameters and ground distance metric parameters together. The training process is conducted with training triplets of sequences, where each triplet is composed of an anchor sequence, a must-link sequence, and a cannot-link sequence. The corresponding hinge loss function of each triplet is minimized, and we develop an iterative algorithm to solve the optimal transport problem and the attention/ground distance metric parameters in an alternate way. The experiments over sequence similarity search benchmark datasets, including text, video, and rice smut protein sequence data, are conducted. The experimental results show the algorithm is effective.

1. Introduction

1.1. Background. Sequence data is one of the most popular data type in real-world applications of machine learning and data mining [1–6]. For example, in natural language processing, a sentence is a sequence of words, and in computer vision, a video is a sequence of frames, while in bioinformatics, a protein structure is a sequence of amino acids in a polypeptide chain. Unlike the flat vector data of most machine learning problems, sequence data has the following inherent features:

(1) Sequence data is varying at the number of items. The flat feature is usually given at a fixed size, while the length of the sequences could be different, due to the sampling process to form the sequence.

(2) Sequence data has a temporal and relational nature. The order of the items in the sequence plays an important role in the understanding of the sequence. Given two sequences of the same items but with different orders, their meaning could be completely different. This is a critical, different nature different from the flat vector data, where the items of the vector are considered to be independent of each other and their orders are not important for the learning problem.

Given these two natures of the sequence data, the common machine learning methods are not necessarily applicable to the sequence data, such as the classification, similarity comparison, representation, and regression models. The most popular way to handle sequence data is to map a sequence to a flat vector and then apply the conventional methods. However, this methodology usually cannot capture the sequential feature of the data; thus, the results are not satisfying [4, 7–12].Comparing the similarity/

dissimilarity of a pair of sequences is a fundamental problem of sequence data analysis and understanding. The applications include the similarity search [13–16] and nearest neighbor-based classification [17, 18]. However, the similarity of the two sequences has an essential difference compared with the distance/similarity metrics of flat vectors, such as Euclidean distance, ℓp-norm distances, correlation, Mahalanobis distance, and all kinds of learned metrics. The calculation of the distance between a pair of sequences is more difficult than that of the flat vectors, due to the complex nature of the sequences as mentioned above. Two similar sequences may have different lengths because they are generated with different sampling rates, and encoding the temporal patterns and sequential relations of the items of sequences to distance measures is also difficult. To tackle these challenges, various solutions are proposed, such as dynamic time warping (DTW) [19–22] and optimal transport (OT) [23–26]. Most of the methods are based on the item-to-item ground distances of the item pairs of the two sequences and matching them accordingly. The ground distance is extremely important for these methods, but ground distance learning does not receive enough attention from the previous researches. In this paper, we study the problem of learning effective ground distance between the items of the two sequences for the purpose of sequence distance comparison.

1.2. Existing Works. In this section, we reviewed a few ground distance-based sequence distance learning methods.

(1) Villani [23] proposed to compare the distance between two sequences by OT. OT treats one sequence as a set of mass, while the other sequence as a set of demands. The effort to move one unit of mass from the ith item of the source sequence to the jth item of the target sequence is treated as the ground distance between the pairs (i, j). The purpose of OT is to move all the masses from the source sequence to the target sequence, with the minimum amount of effort. To this end, OT minimizes the overall effort of mass moving with regard to the amounts of mass moved from the ith source item to the jth target item for all pairs of (i, j). With the solution of the moved amounts, the overall effort is the distance between the sequences of OT.

(2) Su and Hua [4] improved the OT method to consider the positions of items of both source and target sequences. The thought behind this method is that the moved amount of mass from a source item to a neighboring target item should be larger than the other items. To this end, the two regularization objectives are imposed on the learning process of the moved amounts. The first one calculates a position similarity between each pair of source and target items and impose the corresponding moved amount to be large if the similarity is large. The second one firstly constructs a position distance between the pair of source and target items, converts it to the probability of positions being nearby, and finally minimizes the Kullback–Leibler (KL) divergence between the probability and moved amount of each pair.

(3) Su and Wu [7] developed a novel ground distance metric learning algorithm by firstly combining a sequence with its label to form a metasequence and then learn the ground distance to compare the sequence to the metasequence. A linear transformation function is designed to map the sequence to a new space, where the sequence items are calculated. With the ground distances of pairs, the OT method is applied to compare the sequence to the meta sequence. The linear transformation parameters and the transportation amount jointly in a minimization problem where a training set of sequences.

(4) Su et al. [5] designed a novel sequence representation and similarity learning method by using dimensionality reduction to the feature vectors of the items of sequences. It firstly maps the features of the items to a low-dimensional space so that the sequence classes are separated as much as possible. The class separability is measured by the sequence statistics, and different forms of statistics lead to different dimensional reduction methods. Two statistics are considered, which are model-based and distance-based. The model-based method explores the dynamical structure of the sequences, while the distance-based one explores the similarity of pairs of sequences.

1.3. Our Contribution. It has been proven that the OT-based sequence distance comparison is the most powerful method for the classification and retrieval of sequence data. The most critical factor of the OT-based method is actually the ground distance measure between the items. Although there are many studies on how to improve the OT-based sequence distance learning, however, most of them are focusing on learning the optimal parameters of OT, with a given ground distance metric. However, the ground distance is a critical component of the OT, and the quality of the ground distance directly affects the quality of OT-based distance methods. In this paper, we proposed a novel ground distance metric learning method, which employs the cross-attention mechanism [27–31]. To calculate the ground distance between two items from two sequences, respectively, we firstly represent each item by paying attention from itself to the neighbors of the other item and then paying attention back. The representation vector of one item is the linear combination of its neighbors weighted by the attention scores. The attention scores are the normalized similarity between a neighboring item and the target item. To learn the parameters, we build a unified learning framework to optimize the attention layers and the OT parameters, which are transported amounts. Our contributions of this work are listed as follows:

(1) We proposed a novel learning framework to learn the ground distance and OT parameters jointly. In this framework, the ground distance model is composed of cross-attention layers, and OT-based sequence distance is parameterized by the transport amounts. The learning framework allows the attention layers and the transport mounts to regularize the learning of each other. This is the first learning framework to guide the learning of attention layers by OT.

(2) We model the learning framework as a minimization problem and develop an iterative algorithm to solve it. In this algorithm, the attention weight parameters and the transport amounts are updated alternately until the algorithm converges. In each iterative step, we consider the optimization of the parameters one by one, while fixing the other parameters, by solving the suboptimization problems.

(3) We conducted extensive experiments over four benchmark datasets to compare our algorithm against the other sequence distance comparison algorithms. Experimental results show the advantage of the attention-based OT algorithm, and we also show the stable property of the algorithm regarding the change of the trade-off parameter and the iteration number.

1.4. Paper Organization. We organize the following parts of this paper as follows: in Section 2, we introduce the proposed algorithm of sequence distance comparison, in Section 3, we conduct experiments to compare our algorithm against the other popular sequence distance methods and also study the properties of the algorithm, and in Section 4, we give the conclusion of this paper and some future works of attention-based sequence distance learning.

2. Proposed Method

2.1. Problem Modeling. Suppose we have two sequences of items, denoted as $X = \{x1, \ldots, xn\}$ and $YY = \{y1, \ldots, yn\}$, where $xi \in R^{dx}$ is the vector of the ith item of X, and $yj \in R^{dy}$ is the vector of the jth item of Y. To calculate the distance between them, we firstly define an attention-based ground distance metric and then measure the optimal transport distance according to the ground metric.

2.1.1. Cross-Attention-Based Ground Distance. To calculate the ground distance between the ith item of X, xi, and the jth item of Y, yi, we firstly explore their neighboring items. For xi, we collect the h items in X before it, $x_{i-h}, \ldots, x_{i-1}$, and h items after it, $x_{i+1}, \ldots, x_{i+h}$, to form a subsequence around xi, denoted as Ni as the contextual sequence of xi. Similarly, we have the h items before and after yj from Y as its contextual sequence, Mj:

$$\begin{aligned} N_i &= \{x_{i-h}, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_{i+h}\}, \\ M_j &= \{y_{j-g}, \ldots, y_{j-1}, y_j, y_{j+1}, \ldots, y_{j+g}\}. \end{aligned} \tag{1}$$

In this way, the contextual and temporal information of each item is effectively encoded in the subsequences Ni and Mj. To compare the dissimilarity between xi and yj, we compare the two subsequences by the cross-attention mechanism.

2.1.2. Attention from Items of Ni to yj. Firstly, to compare the dissimilarity between the Ni and yj, we calculate the attention from items of Ni to yi. To estimate the attention weight, we firstly calculate the affinity between $xl \in Ni$ and $yk \in Mj$:

$$\omega_{lk} = f\left(\theta^{\top}\begin{bmatrix} xl \\ yk \end{bmatrix}\right), \quad \forall l: xl \in Ni, k: yk \in Mi, \tag{2}$$

where $f(\cdot)$ is a nonlinear activation function, such as hyperbolic tangent transformation, and $\theta \in R^{(dx+dy)}$ is the parameter of the affinity function. The attention weights are obtained by softmax normalization over the items of Ni:

$$\alpha_{lk} = \frac{\exp(\omega_{lk})}{\sum_{l':\, xl' \in Ni} \exp(\omega_{l'k})}. \tag{3}$$

With the attention weights from xl to yk, we calculate a representative vector of Ni with attention to yk:

$$z_k^i = \sum_{l:\, xl \in Ni} \alpha_{lk} x_l \in R^{dx}, \tag{4}$$

as the weighted sum of the items $xl \in Ni$.

2.1.3. Attention from z_k^i of Mj to Ni. We represent the subsequence Ni by averaging the vectors of the items as

$$\overline{x}_i = \frac{1}{|Ni|} \sum_{l:\, xl \in Ni} xl \in R^{dx}. \tag{5}$$

Again, we would like to pay attention from each z_k^i to xi. Similarly, we first calculate the affinity between them as follows:

$$\varpi_k^i = f\left(\phi^{\top}\begin{bmatrix} z_k^i \\ \overline{x}i \end{bmatrix}\right), \tag{6}$$

where $\phi \in R^{2dx}$ is the affinity function parameter. From the affinities between $\overline{x}i$ and $z_k^i | k: yk \in Mj$, we calculate the attention weights from xi to $yk \in Mj$ with a softmax function,

$$\beta_k^i = \frac{\exp\left(\varpi_k^i\right)}{\sum_{k':\, yk' \in Mj} \exp\left(\varpi_{k'}^i\right)}. \tag{7}$$

2.1.4. Cross-Attention-Based Ground Distance. To compare the distance between the representative vector z_k^i to and yk, we first perform a linear transformation over z_k^i by

$$W^{\top} z_k^i \in R^{dy}, \tag{8}$$

with $W \in R^{dx \times dy}$ as parameter. This transformation is to map z_k^i to the same space as yk. Then, we compare their distance by the squared Euclidean distance:

$$d\left(z_k^i, y_k\right) = \left\|W^\top z_k^i - yk\right\|_F^2. \tag{9}$$

The final distance between Ni and Mj is the attention-weighted sum of distances $d(z_k^i, yk)|k\colon k \in Mj$. The attention is calculated in equation (7), and the distance $d(Ni, Mj)$ is

$$d(Ni, Mj) = \sum_{k\colon yk \in Mj} \beta_k^i d\left(z_k^i, yk\right). \tag{10}$$

2.1.5. Optimal Transport Distance. With the ground distance, $d(Ni, Mj)$, between each pair of items $(xi, yj)|xl \in X, yk \in Y$ of two sequences, we can compute the transport distance. The ground distance between xi and yj is viewed as the effort to move one unit of mass from xi to yj. We define a variable, ηij, to denote the mount of mass moved from xi to yj, then the total effort to move the mass from X to Y is calculated as

$$\sum_{i,j\colon xi \in X, yj \in Y} \eta ij\, d(Ni, Mj). \tag{11}$$

Moreover, we define an amount of mass for each item xi of X to be moved out, γi. Thus, the constraint of the amounts moved out of xi is applied as

$$\sum_{j\colon yj \in Y} \eta ij = \gamma i \tag{12}$$

We also define an amount of mass to be received by each item yi of Y, δj, and accordingly

$$\sum_{i\colon xi \in X} \eta ij = \delta j. \tag{13}$$

The optimal transport distance between X and Y is achieved by solving the moved amounts to minimize the moving efforts with the above constraints:

$$\begin{aligned} d(X, Y) = \begin{cases} \min\limits_{tij\colon xi \in X, yj \in Y} & \sum\limits_{i,j\colon xi \in X, yj \in Y} \eta ij\, d(Ni, Mj) \\ \text{s.t.} & \eta ij \geq 0,\ \forall i, j\colon xi \in X,\ yj \in Y, \end{cases} \\ \sum_{j\colon yj \in Y} \eta ij = \gamma i, \quad \forall i\colon xi \in X, \\ \sum_{i\colon xi \in X} \eta ij = \delta j, \quad \forall j\colon yj \in Y. \end{aligned} \tag{14}$$

We rewrite the optimal transport distance as matrix form by defining the following matrices and vectors:

$$\begin{aligned} T &= [\eta ij] \in R_+^{n \times m}, \\ D &= [d(Ni, Mj)] \in R_+^{n \times m}, \\ \gamma &= [\gamma 1, \ldots, \gamma n]\top \in R_+^n, \\ \delta &= [\delta 1, \ldots, \delta m]\top \in R_+^m. \end{aligned} \tag{15}$$

We rewrite equation (14) as

$$\begin{aligned} &d(X, Y) = \min_{T \in \Theta} \operatorname{tr}\left(T^\top D\right), \\ &\text{where } \Theta = \left\{T | T \in R_+^{n \times m}, T1_n = \gamma, T^\top 1_m = \delta\right\}, \end{aligned} \tag{16}$$

where $tr(\cdot)$ is the trace of a matrix and 1_n is a vector of n ones.

2.1.6. Supervised Learning of Attention Parameters and Ground Distance Metric. In the distance measure of optimal transport, we need to learn the parameters of the two attention layers, θ and ϕ, and the parameter of the ground distance W. To learn these parameters, we have a training set of T triplets of sequences:

$$T = \left\{\left(X_t, Y_t^+, Y_t^-\right)\right\}\Big|_{t=1}^T. \tag{17}$$

The tth triplet is composed of an anchor sequence, Xt, a must-link sequence Y_t^+, and a cannot-link sequence Y_t^-. The must-link sequence is supposed to have a short distance to anchor sequence, while the cannot-link sequence is supposed to have a long distance to the anchor. In our scenario, we impose the cannot-link sequence has a longer distance to the anchor than the must-link one, with a margin of ε:

$$d\left(X_t, Y_t^-\right) > d\left(X_t, Y_t^+\right) + \varepsilon. \tag{18}$$

Accordingly, we define the hinge loss function as follows:

$$\begin{aligned} L\left(X_t, Y_t^+, Y_t^-; W, \theta, \phi\right) &= \max\left(0, d\left(X_t, Y_t^+\right) - d\left(X_t, Y_t^-\right) + \varepsilon\right) \\ &= \tau t \times \left(d\left(X_t, Y_t^+\right) - d\left(X_t, Y_t^-\right) + \varepsilon\right), \\ \text{where } \tau t &= \begin{cases} 1, & \text{if } d\left(X_t, Y_t^+\right) + \varepsilon > d\left(X_t, Y_t^-\right), \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \tag{19}$$

The corresponding minimization problem is modeled to learn the parameters:

$$\min_{W,\theta,\phi} \left\{\frac{1}{T}\sum_{t=1}^T L\left(X^t, Y_t^+, Y_t^-; W, \theta, \phi\right) + C\left(\|W\|_F^2 + \|\theta\|_F^2 + \|\phi\|_F^2\right)\right\}. \tag{20}$$

In the objective, the first term is the average of the hinge losses over the training triplets. The second term is the squared $\ell 2$-norms of the parameters to reduce the complexity of the model. C is the trade-off parameter.

2.2. Problem Optimization. To solve the problem of equation (20), we substitute equations (16) and (19) into equation (20):

$$\begin{aligned} \min_{W,\theta,\phi} \Bigg\{ o(W, \theta, \phi) &= \frac{1}{T}\sum_{t=1}^T \tau t \\ &\times \left(\min_{T_t^+ \in \Theta} \operatorname{tr}\left(T_t^{+\top} D_t^+\right) - \min_{T_t^- \in \Theta} \operatorname{tr}\left(T_t^{-\top} D_t^-\right) + \varepsilon\right) \\ &+ C(\|W\|2F + \|\theta\|2F + \|\phi\|2F)\Big\}, \end{aligned} \tag{21}$$

where T_t^+ and T_t^- are the transport amount matrix of the positive and negative pairs of the tth training triplet, and D_t^+ and D_t^- are the corresponding ground distance matrix. In this optimization problem, the optimization of the transport amounts is coupled with the optimization of the parameters of the attention layer and the ground distance metric. The optimizations of (W, θ, ϕ) and $(T_t^+, T_t^-)|_{t=1}^T$ are dependent on each other, making the problem difficult to be solved directly. Instead of seeking the close solution of equation (21), we propose to solve the attention and ground distance metric parameters and the optimal transport variables jointly in an unified minimization problem:

$$\begin{cases} \min\limits_{W,\theta,\phi,(T_t^+,T_t^-)|_{t=1}^T} \left\{ o\left(W,\theta,\phi,(T+t,T-t)|_{t=1}^T\right) = \dfrac{1}{T}\sum\limits_{t=1}^{T} \tau t \times \left(\operatorname{tr}\left(T_t^{+\top} D + t\right) - \operatorname{tr}\left(T_t^{-\top} D_t^-\right) + \varepsilon\right) + C\left(\|W\|_F^2 + \|\theta\|_F^2 + \|\phi\|_F^2\right) \right\} \\ \text{s.t. } T + t \in \ominus,\ T - t \in \ominus,\ t = 1, \ldots, T. \end{cases} \tag{22}$$

In this optimization problem, both W, θ, ϕ and $(T_t^+, T_t^-)|_{t=1}^T$ are the variables of a joint objective function. The optimization of both variables are conducted simultaneously. To solve this problem, we use the alternate optimization method. In an iteration of an iterative algorithm, to optimize one parameter, we firstly fix the other parameters and then solve the suboptimization problem with regard to this parameter. The optimizations of these parameters are introduced as follows.

2.2.1. Optimization of W. By fixing the other parameters and only considering W, we have the following suboptimization problem:

$$\min_W \left\{ o1(W) = \frac{1}{T}\sum_{t=1}^{T} \tau t \times \left(\operatorname{tr}\left(T_t^{+\top} D_t^+\right) - \operatorname{tr}\left(T_t^{-\top} D_t^-\right)\right) + C\|W\|_F^2 \right\}. \tag{23}$$

We substitute equations (9) and (10) into equation (16) and rewrite the optimal transport distance between two sequences X and Y as

$$\begin{aligned} \operatorname{tr}\left(T^\top D\right) &= \sum_{i,j:\ xi\in X, yj\in Y} \eta i j\, d(Ni, Mj) \\ &= \sum_{i,j:\ xi\in X, yj\in Y} \eta i j \left(\sum_{k:\ yk\in Mj} \beta i k \|W\top z i k - \ yk\|_F^2 \right) \\ &= \sum_{i,j:\ xi\in X, yj\in Y} \eta i j \left(\sum_{k:\ yk\in Mj} \beta i k \times \left(\operatorname{tr}\left(W^\top z_k^i z_k^{i\top} W\right) \right.\right. \\ &\quad \left.\left. - 2\operatorname{tr}\left(W^\top z_k^i y_k^\top\right) + \ y_k^\top y_k \right)\right) \\ &= \operatorname{tr}\left(W^\top A_{X,Y,T} W\right) \\ &\quad - 2\operatorname{tr}\left(W^\top B_{X,Y,T}\right) + c_{X,Y,T}, \end{aligned} \tag{24}$$

where

$$\begin{aligned} A_{X,Y,T} &= \sum_{i,j:\ xi\in X, yj\in Y} \eta i j \sum_{k:\ yk\in Mj} \beta i k \left(z_k^i z_k^{i\top}\right), \\ B_{X,Y,T} &= \sum_{i,j:\ xi\in X, yj\in Y} \eta i j \sum_{k:\ yk\in Mj} \beta i k \left(z_k^i y_k^\top\right). \end{aligned} \tag{25}$$

Substituting equation (24) into equation (23), we rewrite the objective function as

$$\begin{aligned} o1(W) = \frac{1}{T}\sum_{t=1}^{T} \tau t \times \Big(& \left(\operatorname{tr}\left(W^\top A_{Xt,Y_t^+,Tt} W\right)\right. \\ & -2\operatorname{tr}\left(W^\top B_{Xt,Y_t^+,Tt}\right) + c_{Xt,Y_t^+,Tt}\Big) - \Big(\operatorname{tr}\left(W^\top A_{Xt,Y_t^-,Tt} W\right) \\ & \left. -2\operatorname{tr}\left(W^\top B_{Xt,Y_t^-,Tt}\right) + c_{Xt,Y_t^-,Tt}\right)\Big) + C \times \operatorname{tr}\left(W^\top W\right) \\ = \operatorname{tr}\left(W^\top E W\right) & - 2\operatorname{tr}\left(W^\top F\right) + c, \end{aligned} \tag{26}$$

where

$$\begin{aligned} E &= \frac{1}{T}\sum_{t=1}^{T} \tau t \times \left(A_{Xt,Y_t^+,Tt} - A_{Xt,Y_t^-,Tt}\right) + C \times I, \\ F &= \frac{1}{T}\sum_{t=1}^{T} \tau t \times \left(B_{Xt,Y_t^+,Tt} - B_{Xt,Y_t^-,Tt}\right), \\ c &= \frac{1}{T}\sum_{t=1}^{T} \tau t \times \left(c_{Xt,Y_t^+,Tt} - c_{Xt,Y_t^-,Tt}\right). \end{aligned} \tag{27}$$

The problem of minimizing $o1(W)$ of equation (23) has a closed-form solution. It is obtained by setting the derivative of $o1(W)$ with regard to W to zero:

$$\nabla W o1(W) = 2EW - 2F = 0 \Longrightarrow W^* = FE - 1. \tag{28}$$

2.2.2. Optimization of θ. We optimize the attention parameter of equation (2), θ. Fixing the other parameters and removing the irrelevant terms from the objective function, we have the following suboptimization problem for θ:

$$\min_{\theta}\left\{ o2(\theta) = \frac{1}{T}\sum_{t=1}^{T} \tau t \times \left(\mathrm{tr}\left(T_t^{+\top} D_t^{+}\right) - \mathrm{tr}\left(T_t^{-\top} D_t^{-}\right)\right) + C\|\theta\|_F^2 = \frac{1}{T}\sum_{t=1}^{T} \tau t \times \sum_{i,j:\ xi\in Xt, yj\in Y_t^{+}} \eta_{ijt}^{+} + d(Ni, Mj) - \sum_{i,j:\ xi\in Xt, yj\in Y_t^{-}} \eta_{ijt}^{-} d(Ni, Mj) + C\|\theta\|_F^2 \right\}. \tag{29}$$

To solve this problem, we use the gradient descent algorithm as

$$\theta \leftarrow \theta - v\nabla_{\theta} o2(\theta), \tag{30}$$

where v is the descent step and $\nabla\theta o2(\theta)$ is the gradient function. To this end, we calculate the gradient of $o2(\theta)$ with regard to θ by the chain rule:

$$\nabla\theta o2(\theta) = \frac{1}{T}\sum_{t=1}^{T} \tau t \times \left(\sum_{i,j:\ xi\in Xt, yj\in Y_t^{+}} \eta_{ijt}^{+} \nabla_{\theta} d(\theta; Ni, Mj) - \sum_{i,j:\ xi\in Xt, yj\in Y_t^{-}} \eta_{ijt}^{-} \nabla_{\theta} d(\theta; Ni, Mj) \right) + 2C\theta, \tag{31}$$

where $\nabla\theta d(\theta; Ni, Mj)$ is the gradient of ground distance between Ni and Mj regarding θ. We substitute equation (9) into equation (10), and meanwhile rewrite the variables are function of θ, we have

$$\begin{aligned} d(\theta; Ni, Mj) &= \sum_{k:\ y_k\in Mj} \beta_k^i \left\| W\top z_k^i(\theta) - y_k \right\|_F^2, \\ \nabla_{\theta} d(\theta; Ni, Mj) &= 2\sum_{k:\ yk\in Mj} \beta_k^i W\left(W\top z_k^i(\theta) - y_k\right)\nabla_{\theta} z_k^i(\theta). \end{aligned} \tag{32}$$

Moreover, the derivatives of the functions of θ are

$$\nabla_{\theta} z_k^i(\theta) = \sum_{l:\ xl\in Ni} \nabla_{\theta}\alpha_{lk}(\theta) x_l. \tag{33}$$

2.2.3. Optimization of ϕ. To optimize ϕ, we have the following suboptimization problem:

$$\min\phi\left\{ o3(\phi) = \frac{1}{T}\sum_{t=1}^{T} \tau t \times \left(\mathrm{tr}\left(T_t^{+\top} D_t^{+}\right) - \mathrm{tr}\left(T_t^{-\top} D_t^{-}\right)\right) + C\|\phi\|_2^F \right\}. \tag{34}$$

Again, we use the gradient descent algorithm to update ϕ as

$$\phi\leftarrow\phi - v\nabla_{\phi} o3(\phi), \tag{35}$$

where $\nabla_{\phi} o3(\phi)$ is the gradient function of $o3(\phi)$ with regard to ϕ. According to the chain rule, we have

$$\nabla_{\phi} o3(\phi) = \frac{1}{T}\sum_{t=1}^{T} \tau t \times \left(\sum_{i,j:\ xi\in Xt, yj\in Y_t^{+}} \eta_{ijt}^{+} \nabla_{\phi} d(\phi; Ni, Mj) - \sum_{i,j:\ xi\in Xt, yj\in Y_t^{-}} \eta_{ijt}^{-} \nabla_{\phi} d(\phi; Ni, Mj) \right) + 2C\phi. \tag{36}$$

2.2.4. Optimization of $(T_t^{+}, T_t^{-})|_{t=1}^{T}$. To optimize the transport amounts, we have the simplified suboptimization problem as

$$\begin{cases} \min\limits_{(T_t^{+}, T_t^{-})|_{t=1}^{T}} & \left\{ o4\left((T_t^{+}, T_t^{-})|_{t=1}^{T}\right) = \frac{1}{T}\sum_{t=1}^{T} \tau t \times \left(\mathrm{tr}\left(T_t^{+\top} D_t^{+}\right) - \mathrm{tr}\left(T - t\top D - t\right) + \varepsilon\right) \right\} \\ \text{s.t.} & T + t \in \ominus,\ T - t \in \ominus,\ t = 1, \ldots, T. \end{cases} \tag{37}$$

According to the objective, the transport amount matrices $(T_t^{+}, T_t^{-})|_{t=1}^{T}$ are in $2T$ independent objectives, so their solutions are also independent to each other. Thus, we can decompose the optimization problem to $2T$ optimal transport problems. For the tth training triplet, we have the following two minimization problems of optimal transport:

$$\begin{cases} \min\limits_{T_t^{+}} & \frac{\tau_t}{T} \times \mathrm{tr}\left(T_t^{+\top} D_t^{+}\right) \\ \text{s.t.} & T_t^{-} \in \ominus, \end{cases} \qquad \begin{cases} \min\limits_{T_t^{-}} & -\frac{\tau_t}{T} \times \mathrm{tr}\left(T_t^{+\top} D_t^{-}\right) \\ \text{s.t.} & T_t^{-} \in \ominus. \end{cases} \tag{38}$$

Each one of the above problems can be solved as a line programming problem (LP).

2.3. Iterative Algorithm. With these optimization results, we design an iterative algorithm to update the parameters. In this algorithm, the parameters are firstly initialized as random variables. Then in a while loop, they are updated sequentially until a maximum iteration number is reached, or the objective value change is smaller than a given threshold. The algorithm is summarized in Algorithm 1.

3. Experiments

We conduct experiments over four benchmark sequence datasets to verify the performance of the proposed AGD algorithm. The experiments are performed from three aspects:

Input: training set of sequence triplets, $T = \{(Xt, Y_t^+, Y_t^-)\}|_{t=1}^T$, maximum iteration number κ, and objective difference threshold δ.
Initialization Initializing parameters (W, θ, φ) as random variables, iteration number $t = 0$.
While $t \leq \kappa$ or $\|o_t - o_{t-1}\|_1 \geq \delta$:
Repeat
(1) Update t according to equation (19) for each training triplet.
(2) Update W according to equation (28).
(3) Update θ by repeating the updating step in equation (30).
(4) Update φ by repeating the updating step in equation (35).
End repeat
Output: (W, θ, φ).

Algorithm 1: Iterative learning algorithm of attention-based ground distance (AGD).

(1) compare with the other sequence similarity/distance methods, (2) study the impacts of the trade-off parameter C, and (3) study the convergence of the iterative algorithm.

3.1. Datasets. In our experiments, we used four benchmark datasets of sequences:

(1) *Spoken Arabic Digits (SAD).* This dataset has 8,800 sequences [32]. Each sequence is a series of speech frames of a wave of a spoken Arabic digit. The vector of each item is the 13-dimensional Mel-frequency cepstrum coefficients feature vector. These sequences belong to 10 classes, and each class is a digit. Each class has 880 sequences. The number of items in each sequence is from 4 to 93.

(2) *NTU RGB + D (NTU).* This dataset has 56,880 sequences [33]. Each sequence is a Kinect video, and each item is a frame of the video. The sequences belong to 60 action classes. The feature vector of each item is constructed by combining the joint locations and the skeleton-based frame wide features.

(3) *Rice Blast Sequence (RBS).* This dataset has 66,153 protein sequences of rice genome proteins, collected from the MSU Rice Genome Database [34]. Each sequence is a sequence of amino acids, and each amino acid is represented by amino acid embedding. The embedding vectors are also learned as a parameter of the model. The sequences are tagged by rice blast disease or not.

(4) *Australian Sign Language (ASL) Signs.* This dataset is composed of 2,565 sequences of sign language signs [32]. The sequences are from 95 classes, and each class has 27 sequences. Each item of a sequence is presented by a 22-dimensional feature vector.

The summary of the statistics of the benchmark datasets is listed in Table 1.

3.2. Experimental Setting

(1) *Training.* To measure the quality of a distance/similarity measure of sequence, we perform the nearest neighbor classification over the sequence data. Given a dataset of sequences with their class labels, we first split the entire dataset by a 10-fold cross-validation protocol. Each fold is used as a test set, while the other folds are used as training folds. Within the training set, we use each sequence as an anchor sequence and randomly pick up another sequence of the same class as its must-link sequence, meanwhile pick up a sequence of a different class as its cannot-link sequence. In this way, we construct the training set of triplets of sequences. The model parameters are trained by the training set and then tested over the test set.

(2) *Testing.* With the trained sequence distance metric, we calculate the distance between each test sequence and each training sequence. The class label of the training sequence with the shortest distance to a test set is assigned to the test, as the classification result of the test sequence.

(3) *Performance Measure.* The accuracy of the test sequences is calculated as the performance. The accuracy rate is the percentage of the correctly classified test sequences over the total number of test sequences.

3.3. Experimental Results

3.3.1. Comparison to Other Methods. We compare the proposed AGD algorithm against the most popular sequence distance learning methods, including the optimal transport (OT) [23], the Order-Reserving Optimal Transport (OPOT) [4], the Regressive Virtual Sequence Metric Learning (RVSML) [7], and the Linear Sequence Discriminant Analysis (LSDA) [5]. The accuracy is reported in Figure 1. From this figure, we can observe that in all the benchmark datasets, the proposed AGD method always has the best performances. The differences between AGD and other methods vary from datasets. For example, in the NTU dataset, the AGD has much better accuracy than the others, while in the RBS dataset, it is only slightly better than the second-best method, LSDA. The main factor behind this phenomenon is the power of the attention mechanism, which embeds each item with its attention to the neighboring items from both the source and target sequences. In most cases, the LSDA is the second-best method, while the original OT method is the worst.

Table 1: Summary of datasets.

Dataset	Number of sequences	Number of classes
SAD	8,800	10
NTU	56,880	60
RBS	66,153	2
ASL	2,565	95

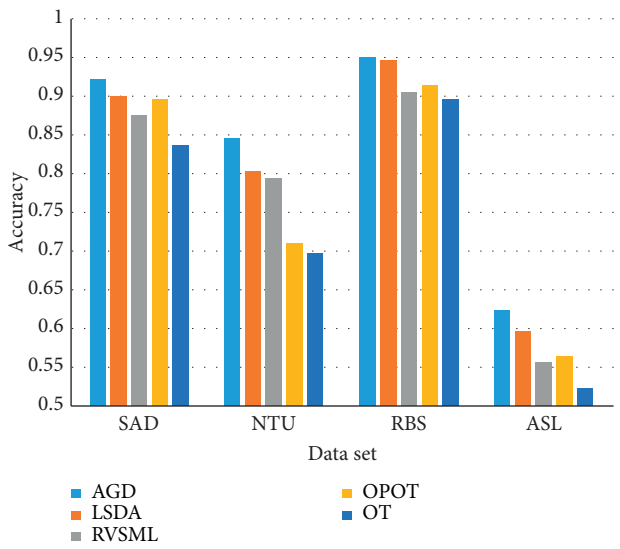

Figure 1: Accuracy of compared methods.

3.3.2. Sensitivity to Trade-Off Parameter. In the objective function of our method, there is only one trade-off parameter, C. It controls the regularization term's importance. We perform experiments with varying values of C and the results are shown in Figure 2. From the curves in the figure, we can see that the proposed AGD algorithm is stable to the changes of the trade-off parameter in most cases. The only exception is the results of the dataset NTU. But the change of the accuracy over the change of the value of C is acceptable. The overall conclusion is that AGD is not sensitive to C. Thus, the parameter tuning of C is easy for the users. One more observation is with the value of C increasing, the accuracy is slightly improving. This also verifies that the regularization term is also beneficial to the model.

3.3.3. Convergence Study. Since our algorithm is an iterative algorithm, we are also interested in the convergence of the algorithm. Thus, we plot the curve of accuracy versus the number of iterations. The curves are given in Figure 3. From this figure, we can see that with the iteration number increasing, the accuracy keeps improving until converge. The number of iterations for the convergence is around 50. The convergence of the algorithm is experimentally verified, and for the size of datasets comparable to our benchmark, the convergence iteration number is acceptable.

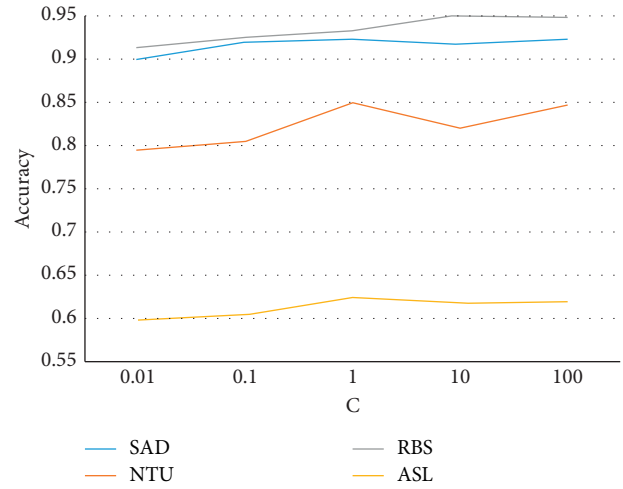

Figure 2: Sensitivity to the trade-off parameter C.

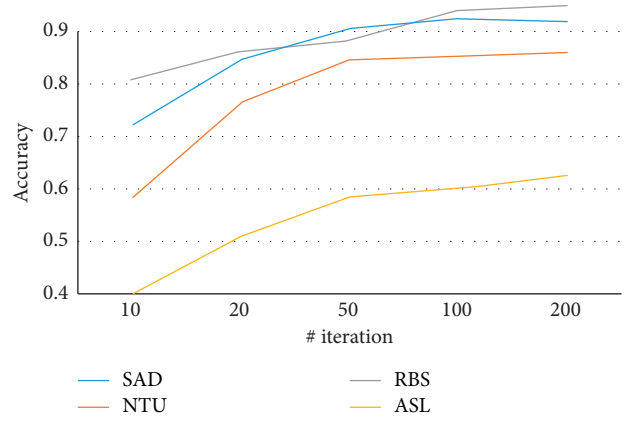

Figure 3: Convergence curves.

We test the significance of the convergence of the accuracy by the Ratio test, and the r values are reported in Table 2. According to these r values, all of them are smaller than 1, meaning all the curves are significantly converged.

3.3.4. Running Time. We also compare the running time of the proposed method. The running times over the four benchmark datasets are shown in Figure 4. From this figure, we have the following conclusions:

(1) Running time and data size are positively correlated. The largest dataset has the longest running time while the smaller one has shorter running time. This is natural since both the training and test processes scan the data points one by one, and more data points means more scanning time.

(2) Our algorithm is faster than the LSDA and RCSML algorithms, while it is slower than the OPOT and OT algorithms. This is acceptable given the significant improvement of the accuracy.

Table 2: r values of ratio test of the convergence of the accuracy.

Dataset	SAD	NTU	RBS	ASL
r value	0.701	0.674	0.833	0.891

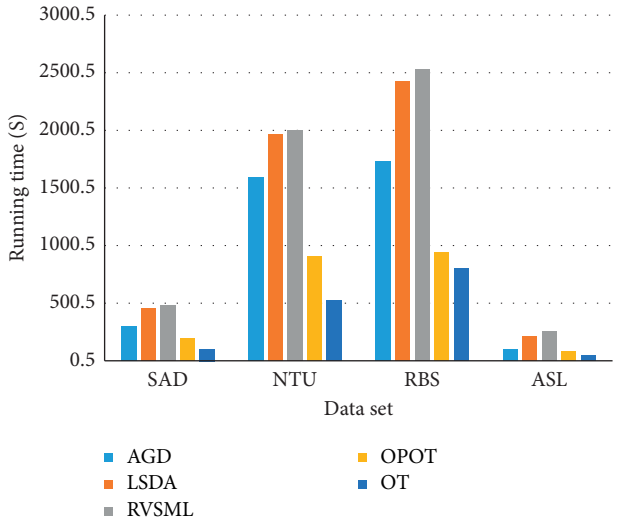

Figure 4: Running time analysis.

4. Conclusion

In this paper, we proposed a novel sequence distance measuring algorithm. This algorithm is based on OT, but its main focus is how to learn an effective ground distance measure for the two sequences. The ground distance learning falls to the framework of the cross-attention mechanism, and the attention layer parameters and the OT parameters are learned jointly. This design can use the OT to guild the learning of the attention layers. Thus, this framework can provide representation of the two sequences, the ground distance, and the OT simultaneously to optimize the model. The learning is also guided by the supervisor of the must-link and cannot-link triplets of the sequences. The parameters are optimized in an iterative algorithm, and the algorithm is tested over four sequence datasets. The experimental results show its advantage over the sequence comparison algorithms.

Data Availability

All the datasets used in this paper to produce the experimental results are publicly accessed online.

Conflicts of Interest

The authors declare that there are no potential conflicts of interest regarding the publication of this study.

Acknowledgments

This work was supported by the Project of "Breeding of New Varieties of High Quality and Anti-Resistant Rice" (Grant no. 2020ZX16B01013), Agricultural Science and Technology Innovation Spanning Project of Heilongjiang Academy of Agricultural Sciences (Grant no. HNK2019CX02), and the National Technology System for Modern Agricultural Industry "Wuchang Integrated Test Station" (Grant no. CARS-01-54).

References

[1] G. Dong and J. Pei, Sequence Data Mining, vol. 33, Springer Science & Business Media, Berlin, Germany.

[2] Yu Chung-Ching Yu and Y.-L. Yen-Liang Chen, "Mining sequential patterns from multidimensional sequence data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 1, pp. 136–140, 2005.

[3] B. Su and Y. Wu, "Learning meta-distance for sequences by learning a ground metricvia virtual sequence regression," *IEEE Annals of the History of Computing*, vol. 1, no. 1, p. 1, 2020.

[4] B. Su and G. Hua, "Order-preserving optimal transport for distances between sequences," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 12, pp. 2961–2974, 2018.

[5] B. Su, X. Ding, H. Wang, and Y. Wu, "Discriminative dimensionality reduction for multi-dimensional sequences," *IEEE Transactions on Pattern Analysis Andmachine Intelligence*, vol. 40, no. 1, pp. 77–91, 2017.

[6] C. Wang and H. Mo, "Learning deep attention network from incremental and decremental features for evolving features," *Scientific Programming*, vol. 2021, Article ID 1492828, 8 pages, 2021.

[7] B. Su and Y. Wu, "Learning meta-distance for sequences by learning a ground metricvia virtual sequence regression," *IEEE Transactions on Pattern Analysis and MachineIntelligence*, vol. 1, p. 1, 2020.

[8] B. Liu, C.-C. Li, and K. Yan, "DeepSVM-fold: protein fold recognition by combining support vector machines and pairwise sequence similarity scores generated by deep learning networks," *Briefings in Bioinformatics*, vol. 21, no. 5, pp. 1733–1741, 2020.

[9] K. James, M. Taylor, A. D. Steen, and A. Sadovnik, "Unaligned sequence similarity search using deep learning," in *Proceedings of the 2019 IEEE International Conferenceon Bioinformatics and Biomedicine (BIBM)*, pp. 1892–1899, IEEE, San Diego, CA, USA, November 2019.

[10] N. Paul, M. Versteegh, and M. Rotaru, "Learning text similarity with siamese recurrent networks," in *Proceedings of the 1st Workshop on Representation Learning for NLP*, pp. 148–157, Berlin, Germany, August 2016.

[11] G. Liang, H. Mo, Z. Wang, C.-Q. Dong, and J.-Y. Wang, "Joint deep recurrent network embedding and edge flow estimation," in *Proceedings of the International Conference on Intelligent Computing*, pp. 467–475, Springer, Shenzhen, China, August 2020.

[12] L. Yu, H. Wang, and H. Mo, "Estimating network flowing over edges byrecursive network embedding," *Shock and Vibration*, vol. 2020, Article ID 8893381, 7 pages, 2020.

[13] R. D. Finn, J. Clements, and S. R. Eddy, "HMMER web server: interactive sequence similarity searching," *Nucleic Acids Research*, vol. 39, no. 2, pp. W29–W37, 2011.

[14] W. R. Pearson, "Flexible sequence similarity searching with the fasta3 program package," in *Bioinformatics Methods and Protocols*, Springer, Berlin, Germany, 2000.

[15] R. Agrawal, C. Faloutsos, and A. Swami, "Efficient similarity search in sequence databases," in *Proceedings of the International Conference On Foundations Of Data*

Organizationand Algorithms, pp. 69–84, Springer, Chicago, IL, USA, October 1993.
[16] S. Wiseman and A. M. Rush, "Sequence-to-sequence learning as beam-search optimization," 2016, https://arxiv.org/abs/1606.02960.
[17] Z. Xing, J. Pei, and E. Keogh, "A brief survey on sequence classification," *ACM Sigkdd Explorations Newsletter*, vol. 12, no. 1, pp. 40–48, 2010.
[18] L. Neal, M. J. Zaki, and M. Ogihara, "Mining features for sequence classification," in *Proceedings of the 5th ACM SIGKDD international conference onKnowledge discovery and data mining*, pp. 342–346, San Diego, CA, USA, August 1999.
[19] M. Müller, *Dynamic Time Warping. Information Retrieval for Music and Motion*, Springer, Berlin, Germany, 2007.
[20] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns intime series," in *Proceedings of the KDD Workshop*, pp. 359–370, Seattle, WA, USA, 1994.
[21] S. Pavel, Dynamic Time Warping Algorithm Review, vol. 855, Information and Computer Science Department University of Hawaii at Manoa Honolulu, Honolulu, HI, USA.
[22] J. Eamonn and M. J. Pazzani, "Derivative dynamic time warping," in *Proceedings of the 2001 SIAM International Conference On Data Mining*, pp. 1–11, SIAM, Chicago, IL, USA, April 2001.
[23] C. Villani., *Optimal Transport: Old and New*, Vol. 338, Springer Science & BusinessMedia, Berlin, Germany, 2008.
[24] M. Cuturi, "Sinkhorn distances: lightspeed computation of optimal transport," *Advances in Neural Information Processing Systems*, vol. 26, pp. 2292–2300, 2013.
[25] P. Gabriel and M. Cuturi, "Computational optimal transport: with applicationsto data science," *Foundations and Trends R ○ in Machine Learning*, vol. 11, no. 5-6, pp. 355–607, 2019.
[26] L. Ambrosio, "Lecture notes on optimal transport problems," in *Mathematical Aspectsof Evolving Interfaces*, Springer, Berlin, Germany, 2003.
[27] Y. Hao, Y. Zhang, K. Liu et al., "An end-to-end model for question answering over knowledge base with cross-attention combining global knowledge," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, vol. 1, pp. 221–231, Vancouver, Canada, January 2017.
[28] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "Ccnet: criss-cross attention for semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 603–612, Seoul, Korea, 2019.
[29] K.-H. Lee, Xi Chen, G. Hua, H. Hu, and X. He, "Stacked crossattention for image-text matching," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 201–216, Munich, Germany, September 2018.
[30] R. Hou, H. Chang, B. Ma, S. Shan, and X. Chen, "Cross attention network for few-shot classification," *Advances in Neural Information Processing Systems*, vol. 2019, pp. 4003–4014, 2019.
[31] G. Liang, H. Mo, Y. Qiao, C. Wang, and J.-Y. Wang, "Paying deep attention to both neighbors and mt," in *Proceedings of the International Conference onIntelligent Computing*, pp. 140–149, Springer, Bari Italy, October 2020.
[32] A. Arthur and D. Newman, *Uci Machine Learning Repository*, University of California, Irvine, School of Information and Computer Sciences, Irvine, CA, USA, 2007.
[33] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+ d: a largescale dataset for 3d human activity analysis," in *Proceedings of the IEEE Conference Oncomputer Vision And Pattern Recognition*, pp. 1010–1019, Las Vegas, NV, USA, June 2016.
[34] Y. Kawahara, M. De La Bastide, J. P. Hamilton et al., "Improvement of the oryza sativa nipponbare reference genome using next generation sequence and optical map data," *Rice*, vol. 6, no. 1, p. 4, 2013.