


天津大學應用數學中心
Center for Applied Mathematics, Tianjin University

Deep Squared Euclidean Approximation to the Levenshtein Distance for DNA Storage

Alan J. X. Guo (郭嘉祥), Cong Liang (梁聪), Qing-Hu Hou (侯庆虎)

Background

In the DNA storage pipeline, retrieved DNA sequences need to be clustered before they can be decoded.

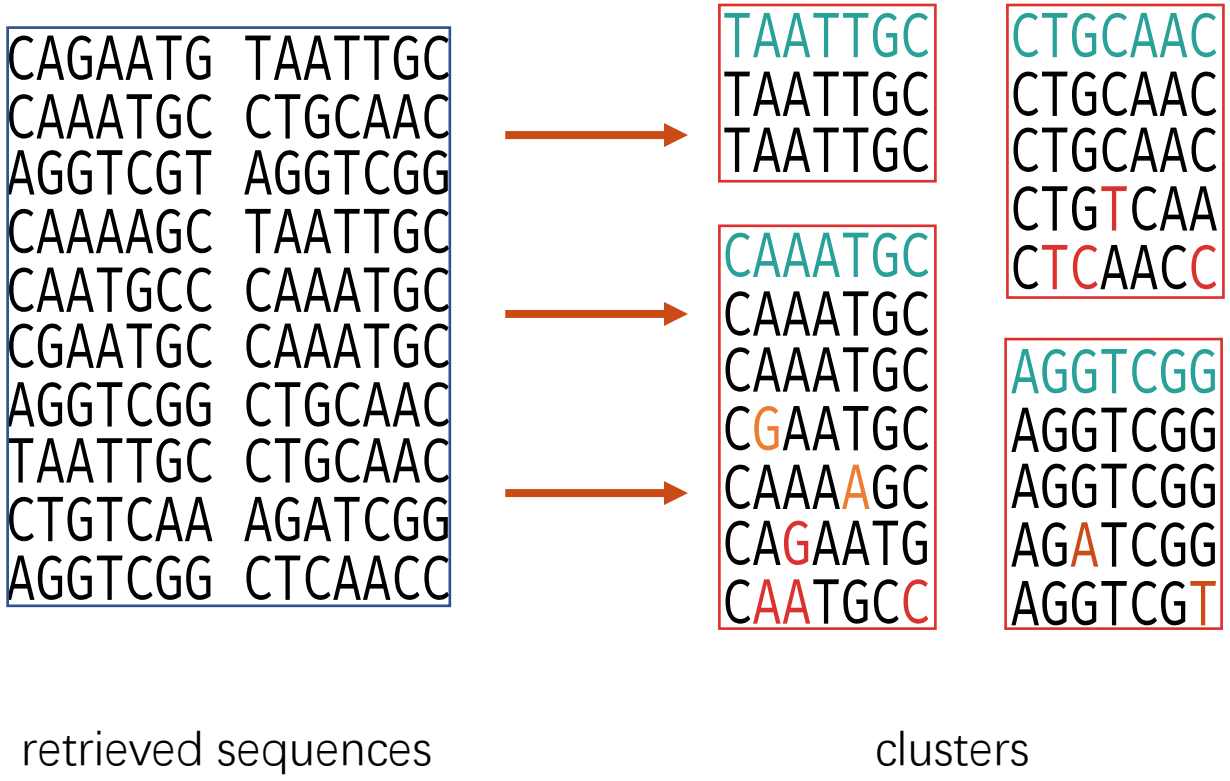
Background

In the DNA storage pipeline, retrieved DNA sequences need to be clustered before they can be decoded.

retrieved sequences

Background

In the DNA storage pipeline, retrieved DNA sequences need to be clustered before they can be decoded.



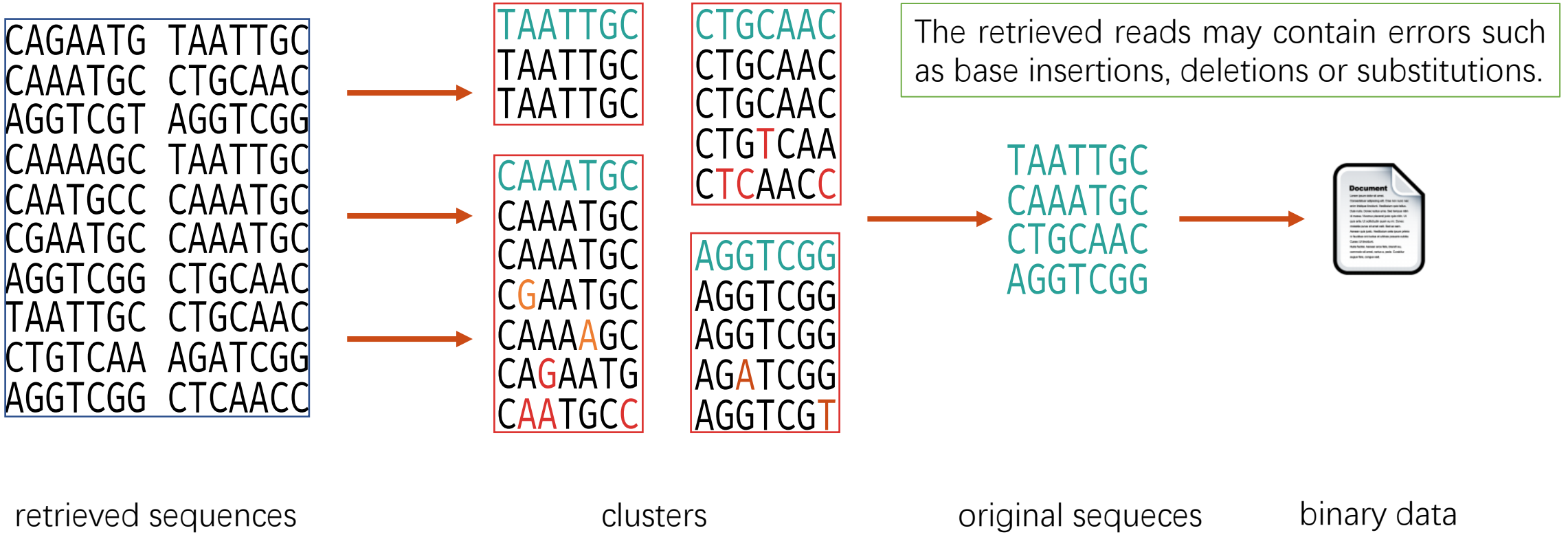
retrieved sequences

clusters

The retrieved reads may contain errors such as base insertions, deletions or substitutions.

Background

In the DNA storage pipeline, retrieved DNA sequences need to be clustered before they can be decoded.



Background

To cluster the retrieved sequences, the Levenshtein distance (Edit distance) is used to evaluate the similarity between two sequences.

Background

To cluster the retrieved sequences, the Levenshtein distance (Edit distance) is used to evaluate the similarity between two sequences.

Levenshtein distance (Edit distance):

The Levenshtein distance between two sequences is the minimum number of insertions, deletions, or substitutions required to modify one string to the other.

Background

To cluster the retrieved sequences, the Levenshtein distance (Edit distance) is used to evaluate the similarity between two sequences.

Levenshtein distance (Edit distance):

The Levenshtein distance between two sequences is the minimum number of insertions, deletions, or substitutions required to modify one string to the other.

- Levenshtein distance cannot be computed in $O(n^{2-\epsilon}), \forall \epsilon > 0$, unless the strong exponential time hypothesis is false.

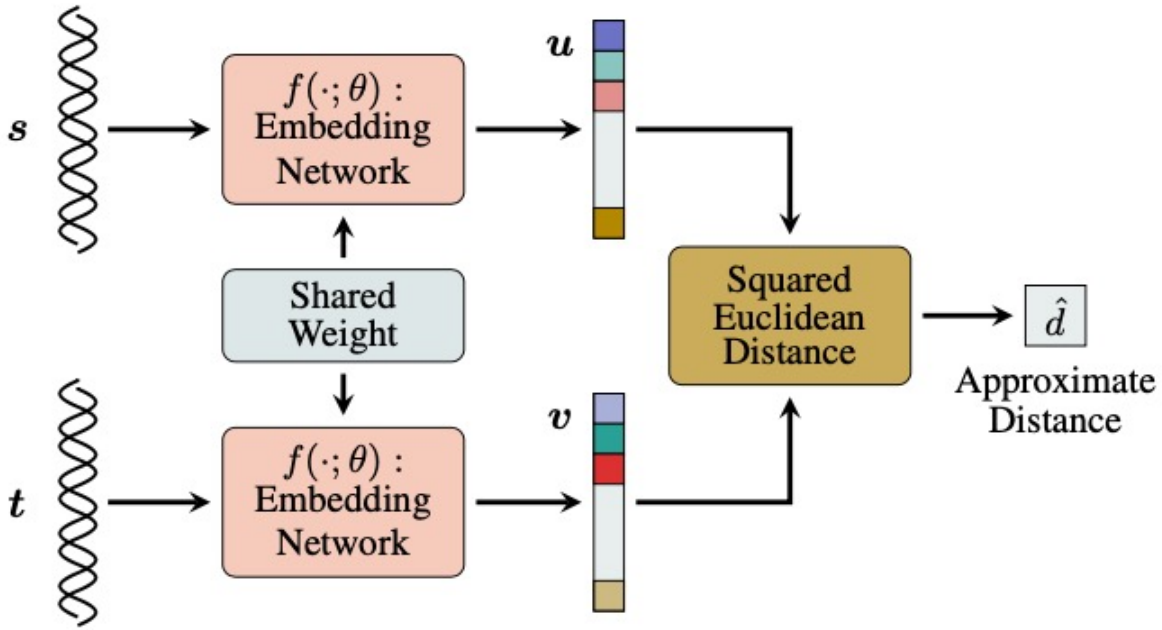
Levenshtein distance embedding

Find an embedding function $f(\cdot)$ that maps the DNA sequences $\boldsymbol{s}, \boldsymbol{t}$ to their embedding vectors $\boldsymbol{u} = f(\boldsymbol{s}), \boldsymbol{v} = f(\boldsymbol{t})$, such that the Levenshtein distance between $\boldsymbol{s}, \boldsymbol{t}$ can be approximated by the commonly used distances between $\boldsymbol{u}, \boldsymbol{v}$, $d_L(\boldsymbol{s}, \boldsymbol{t}) \approx d(\boldsymbol{u}, \boldsymbol{v})$.

Levenshtein distance embedding

Find an embedding function $f(\cdot)$ that maps the DNA sequences $\boldsymbol{s}, \boldsymbol{t}$ to their embedding vectors $\boldsymbol{u} = f(\boldsymbol{s}), \boldsymbol{v} = f(\boldsymbol{t})$, such that the Levenshtein distance between $\boldsymbol{s}, \boldsymbol{t}$ can be approximated by the commonly used distances between $\boldsymbol{u}, \boldsymbol{v}$, $d_L(\boldsymbol{s}, \boldsymbol{t}) \approx d(\boldsymbol{u}, \boldsymbol{v})$.

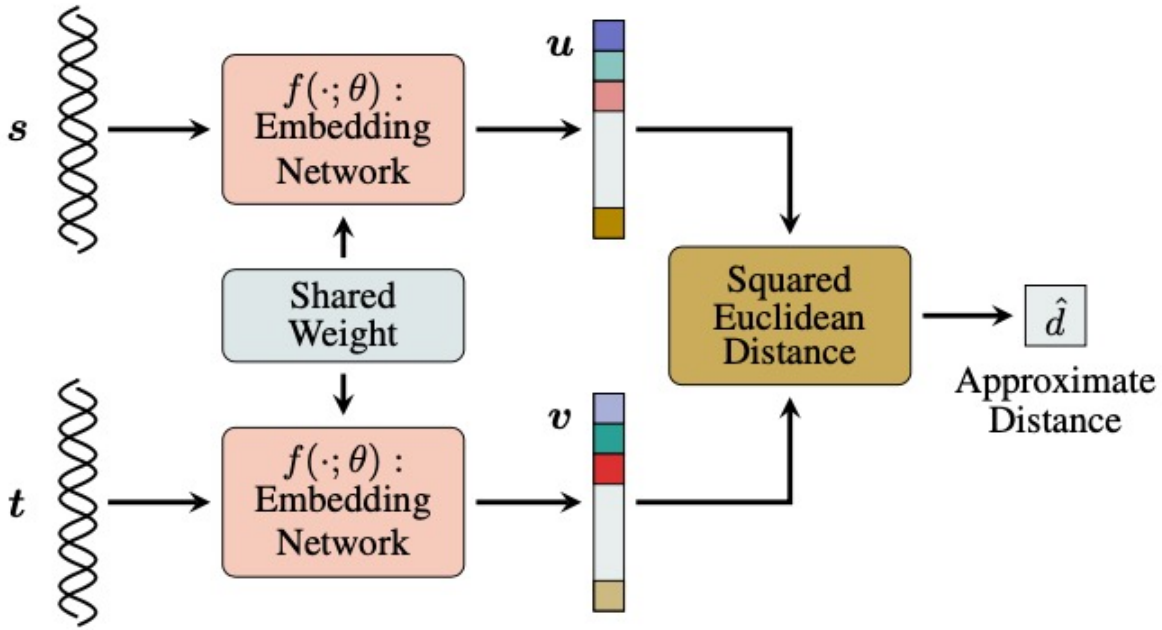
The deep learning-based approach – Siamese Neural Network



Levenshtein distance embedding

Find an embedding function $f(\cdot)$ that maps the DNA sequences $\boldsymbol{s}, \boldsymbol{t}$ to their embedding vectors $\boldsymbol{u} = f(\boldsymbol{s}), \boldsymbol{v} = f(\boldsymbol{t})$, such that the Levenshtein distance between $\boldsymbol{s}, \boldsymbol{t}$ can be approximated by the commonly used distances between $\boldsymbol{u}, \boldsymbol{v}$, $d_L(\boldsymbol{s}, \boldsymbol{t}) \approx d(\boldsymbol{u}, \boldsymbol{v})$.

The deep learning-based approach – Siamese Neural Network



Squared Euclidean Distance:

$$d_{\ell_2^2} = \sum_{i=1}^{n} (u_i - v_i)^2.$$

By optimizing:

$$\hat{\theta} = \arg\min_{\theta} \mathcal{L}(\boldsymbol{d}, \hat{\boldsymbol{d}}; \theta)$$

one will obtain the parameters $\hat{\theta}$ that enables the $f(\cdot;\hat{\theta})$ as the embedding function.

Why do we use the squared Euclidean distance?

- The ground truth discrete distance can be interpreted as the degree of freedom of the difference between the embedding vectors $\boldsymbol{u} - \boldsymbol{v}$.
- The Siamese neural network can be optimized by a better loss.

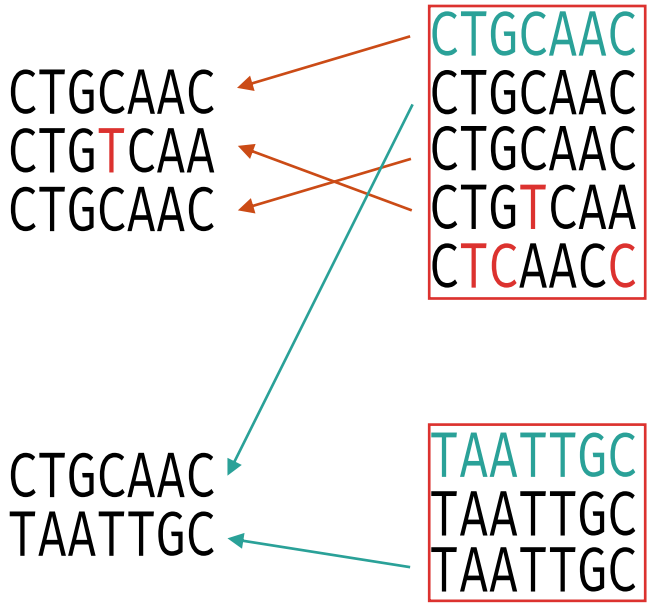
Why do we use the squared Euclidean distance?

- The ground truth discrete distance can be interpreted as the degree of freedom of the difference between the embedding vectors $\boldsymbol{u} - \boldsymbol{v}$.
- The Siamese neural network can be optimized by a better loss.

Notations and Assumptions

Homologous sequences:
Intra cluster sequences, which are similar to each other.

Non-homologous sequences:
Inter cluster sequences, which are none related to each other.



Notations and Assumptions

- **Each element u_i of the embedding vector follows the standard normal distribution $N(0,1)$;**

  By using the Batch Normalization layer, the **mean** and **std** values of the embedded element u_i will be close to 0 and 1 respectively.

Notations and Assumptions

- **Each element u_i of the embedding vector follows the standard normal distribution $N(0,1)$;**

  By using the Batch Normalization layer, the **mean** and **std** values of the embedded element u_i will be close to 0 and 1 respectively.

- **The embedding elements u_i, u_j are independent, iff $i \neq j$;**

  Dependence of embedding elements is a waste of the embedding dimension when the information in the original sequence is not fully expressed.

Notations and Assumptions

- **Each element u_i of the embedding vector follows the standard normal distribution $N(0,1)$;**

  By using the Batch Normalization layer, the **mean** and **std** values of the embedded element u_i will be close to 0 and 1 respectively.

- **The embedding elements u_i, u_j are independent, iff $i \neq j$;**

  Dependence of embedding elements is a waste of the embedding dimension when the information in the original sequence is not fully expressed.

- **If sequences s and t are non-homologous, their embedding elements u_i and v_j are independent.**

From the degree of freedom to the embedding dimension

If s, t are non-homologous sequences:

- By multiplying a rescaling factor $\frac{\sqrt{2}}{2}$, the $u_i - v_i$ follows $N(0,1)$;
- The $u_i - v_i$ and $u_j - v_j$ are independent, iff $i \neq j$.

From the degree of freedom to the embedding dimension

If s, t are non-homologous sequences:

- By multiplying a rescaling factor $\frac{\sqrt{2}}{2}$, the $u_i - v_i$ follows $N(0,1)$;
- The $u_i - v_i$ and $u_j - v_j$ are independent, iff $i \neq j$.

The squared Euclidean distance between u, v follows chi-squared distribution

$$\sum_{i=1}^{d}(u_i - v_i)^2 \sim \chi^2(d),$$

where the d is the dimension of the embedding vector.

From the degree of freedom to the embedding dimension

If s, t are non-homologous sequences:

- By multiplying a rescaling factor $\frac{\sqrt{2}}{2}$, the $u_i - v_i$ follows $N(0,1)$;
- The $u_i - v_i$ and $u_j - v_j$ are independent, iff $i \neq j$.

The squared Euclidean distance between u, v follows chi-squared distribution

$$\sum_{i=1}^{d} (u_i - v_i)^2 \sim \chi^2(d),$$

where the d is the dimension of the embedding vector.

The expectation of chi-squared distribution is the degree of freedom, and should meet the average value of Levenshtein distances between non-homologous sequences.

On the engaged DNA-Fountain dataset, this number is $d = 80$.

From the degree of freedom to the ground truth distance

If $\boldsymbol{s}, \boldsymbol{t}$ are homologous sequences:

We say the degree of freedom with $\boldsymbol{u} - \boldsymbol{v}$ is d, if

$$\boldsymbol{u} - \boldsymbol{v} = \boldsymbol{y}\boldsymbol{P} = (y_1, \dots, y_d, 0, \dots, 0)\boldsymbol{P}$$

where $\boldsymbol{P}$ is an orthogonal matrix, and y_i are i.d.d. and follow $N(0,1)$.

From the degree of freedom to the ground truth distance

If $\boldsymbol{s}, \boldsymbol{t}$ are homologous sequences:

We say the degree of freedom with $\boldsymbol{u} - \boldsymbol{v}$ is d, if

$$\boldsymbol{u} - \boldsymbol{v} = \boldsymbol{y}\boldsymbol{P} = (y_1, \dots, y_d, 0, \dots, 0)\boldsymbol{P}$$

where $\boldsymbol{P}$ is an orthogonal matrix, and y_i are i.d.d. and follow $N(0,1)$.

The smaller the ground truth distance d_L is, the more related the embedding vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ are, and the less free variables y_is are needed to support the $\boldsymbol{u} - \boldsymbol{v}$.

From the degree of freedom to the ground truth distance

If $\boldsymbol{s}, \boldsymbol{t}$ are homologous sequences:

We say the degree of freedom with $\boldsymbol{u} - \boldsymbol{v}$ is d, if

$$\boldsymbol{u} - \boldsymbol{v} = \boldsymbol{y}\boldsymbol{P} = (y_1, \dots, y_d, 0, \dots, 0)\boldsymbol{P}$$

where $\boldsymbol{P}$ is an orthogonal matrix, and y_i are i.d.d. and follow $N(0,1)$.

The smaller the ground truth distance d_L is, the more related the embedding vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ are, and the less free variables y_is are needed to support the $\boldsymbol{u} - \boldsymbol{v}$.

The squared Euclidean distance between $\boldsymbol{u}, \boldsymbol{v}$ is

$$\hat{d} = (\boldsymbol{u} - \boldsymbol{v})(\boldsymbol{u} - \boldsymbol{v})^T = \boldsymbol{y}\boldsymbol{P}\boldsymbol{P}^T\boldsymbol{y}^T = \boldsymbol{y}\boldsymbol{y}^T = \sum_{i=1}^{d} y_i^2$$

and follows the chi-squared distribution with the degree of freedom d

$$\hat{d} \sim \chi^2(d).$$

From the degree of freedom to the ground truth distance

If $\boldsymbol{s}, \boldsymbol{t}$ are homologous sequences:

We say the degree of freedom with $\boldsymbol{u} - \boldsymbol{v}$ is d, if

$$\boldsymbol{u} - \boldsymbol{v} = \boldsymbol{y}\boldsymbol{P} = (y_1, \dots, y_d, 0, \dots, 0)\boldsymbol{P}$$

where $\boldsymbol{P}$ is an orthogonal matrix, and y_i are i.d.d. and follow $N(0,1)$.

The smaller the ground truth distance d_L is, the more related the embedding vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ are, and the less free variables y_is are needed to support the $\boldsymbol{u} - \boldsymbol{v}$.

The squared Euclidean distance between $\boldsymbol{u}, \boldsymbol{v}$ is

$$\hat{d} = (\boldsymbol{u} - \boldsymbol{v})(\boldsymbol{u} - \boldsymbol{v})^T = \boldsymbol{y}\boldsymbol{P}\boldsymbol{P}^T\boldsymbol{y}^T = \boldsymbol{y}\boldsymbol{y}^T = \sum_{i=1}^{d} y_i^2$$

and follows the chi-squared distribution with the degree of freedom d

$$\hat{d} \sim \chi^2(d).$$

By the squared Euclidean embedding, the approximation of ground truth distance d_L can be interpreted as forcing the embedding $\boldsymbol{u} - \boldsymbol{v}$ to have degree of freedom d_L.

The regression

The commonly used approximation errors are usually symmetric on ground truth d.
For example, the MSE is calculated as

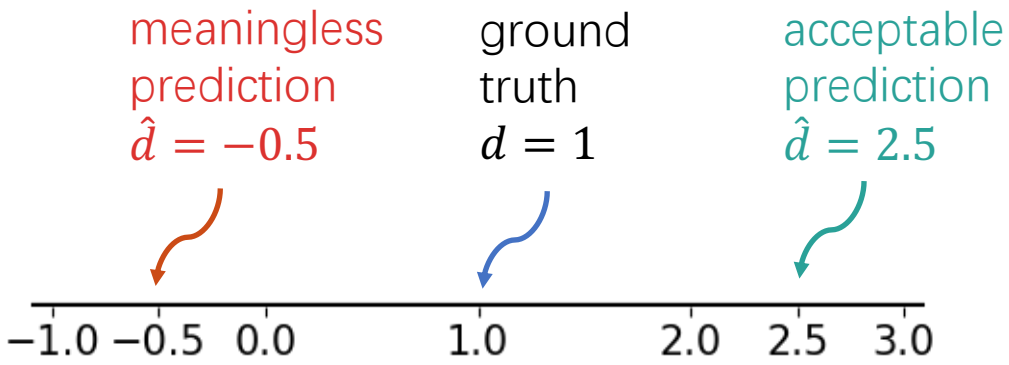
$$\mathrm{MSE}(\hat{d}, d) = (\hat{d} - d)^2.$$

The regression

The commonly used approximation errors are usually symmetric on ground truth d.
For example, the MSE is calculated as

$$\mathrm{MSE}(\hat{d}, d) = (\hat{d} - d)^2.$$

However, when distances are approximated, the approximations are skewed around the ground truth distance.



The regression

By the connection between the ground truth distance and the degree of freedom with the embedding $\boldsymbol{u} - \boldsymbol{v}$, the predicted distance $\hat{d}$ should follow the chi-squared distribution,

$$\hat{d} \sim \chi^2(d).$$

where the d is the ground truth distance.

The regression

By the connection between the ground truth distance and the degree of freedom with the embedding $\boldsymbol{u} - \boldsymbol{v}$, the predicted distance $\hat{d}$ should follow the chi-squared distribution,

$$\hat{d} \sim \chi^2(d).$$

where the d is the ground truth distance.

An entropy style loss function can be defined by

$$\begin{aligned} \mathrm{RE}\chi^2(\hat{d}, d) &= -\log q_d(\hat{d}) \\ &= \frac{d}{2} + \log\Gamma\left(\frac{d}{2}\right) - \left(\frac{d}{2} - 1\right)\log\hat{d} \\ &\quad + \frac{\hat{d}}{2}\log \mathrm{e}, \end{aligned}$$

where the $q_d(x)$ is the pdf of $\chi^2(d)$ distribution.

Experiments

The following three options are available for the proposed approach:

- **The structure of embedding neural network.**

  Plenty structures can be used for the embedding function $f(\cdot)$.

- **Embedding space**

  Instead of the squared Euclidean distance, one can use alternative distances, such as Manhattan distance, Euclidean Distance, etc..

- **Chi-square regressing**

  MAE an MSE can be engaged as the alternative optimization target.

Experiments

Metric	Embed	CNN-ED-5			CNN-ED-10		
		MSE	MAE	REχ^2	MSE	MAE	REχ^2
AE	ℓ_1	4.74 ± 0.03	3.57 ± 0.18	*5.66 ± 0.15*	4.60 ± 0.13	3.70 ± 0.16	*5.26 ± 0.04*
	ℓ_2	6.23 ± 0.01	3.73 ± 0.11	*6.02 ± 0.42*	5.93 ± 0.07	3.56 ± 0.06	*5.00 ± 0.06*
	ℓ_2^2	4.20 ± 0.06	4.12 ± 0.02	4.67 ± 0.09	4.11 ± 0.01	4.12 ± 0.08	4.53 ± 0.03
AE_h	ℓ_1	3.50 ± 0.02	2.50 ± 0.15	*1.89 ± 0.05*	3.44 ± 0.04	2.71 ± 0.19	*1.96 ± 0.01*
	ℓ_2	5.99 ± 0.01	2.69 ± 0.08	*2.59 ± 0.03*	5.88 ± 0.02	2.69 ± 0.14	*2.77 ± 0.05*
	ℓ_2^2	**0.90 ± 0.05**	**0.96 ± 0.09**	**0.90 ± 0.00**	1.11 ± 0.02	1.56 ± 0.15	**0.91 ± 0.01**
OA	ℓ_1	**99.98 ± 0.00**	96.57 ± 0.30	*99.42 ± 0.11*	**99.98 ± 0.01**	96.59 ± 0.27	*99.27 ± 0.04*
	ℓ_2	99.85 ± 0.01	96.40 ± 0.26	*98.34 ± 0.09*	99.66 ± 0.06	96.81 ± 0.09	*98.14 ± 0.02*
	ℓ_2^2	**99.98 ± 0.01**	**99.85 ± 0.08**	**99.98 ± 0.00**	**99.91 ± 0.00**	99.06 ± 0.16	**99.98 ± 0.01**

Metric	Embed	RNN			GRU		
		MSE	MAE	REχ^2	MSE	MAE	REχ^2
AE	ℓ_1	5.25 ± 0.05	4.32 ± 0.43	*5.89 ± 0.18*	4.61 ± 0.14	3.45 ± 0.26	*5.36 ± 0.06*
	ℓ_2	7.15 ± 0.08	5.11 ± 0.44	*6.71 ± 0.33*	7.52 ± 0.15	3.89 ± 0.15	*5.32 ± 0.05*
	ℓ_2^2	4.31 ± 0.01	4.36 ± 0.06	5.41 ± 0.02	3.98 ± 0.02	4.05 ± 0.02	5.51 ± 0.05
AE_h	ℓ_1	4.06 ± 0.05	3.25 ± 0.28	*2.25 ± 0.03*	3.55 ± 0.09	2.48 ± 0.27	*2.09 ± 0.05*
	ℓ_2	6.49 ± 0.15	3.56 ± 0.26	*3.15 ± 0.14*	6.40 ± 0.04	2.75 ± 0.09	*2.73 ± 0.02*
	ℓ_2^2	1.03 ± 0.02	1.14 ± 0.24	**0.91 ± 0.01**	**0.73 ± 0.03**	**0.67 ± 0.02**	**0.88 ± 0.00**
OA	ℓ_1	**99.96 ± 0.01**	96.51 ± 0.42	*99.15 ± 0.13*	**99.98 ± 0.00**	96.72 ± 0.25	*99.40 ± 0.14*
	ℓ_2	99.77 ± 0.02	96.10 ± 0.87	*98.23 ± 0.13*	99.88 ± 0.01	96.25 ± 0.26	*98.21 ± 0.02*
	ℓ_2^2	**99.96 ± 0.00**	99.77 ± 0.18	**99.94 ± 0.00**	**100.00± 0.00**	**99.99 ± 0.00**	**99.97 ± 0.00**

Thank you for listening!

天津大學應用數學中心
Center for Applied Mathematics, Tianjin University