Deep neural networks for climate relation extraction

Jian Zheng, Jianfeng Wang, Shuping Chen, Jiang Li, Yanping Chen, Bingchuan Li

Chongqing Aerospace Polytechnic, Chongqing, China
Corresponding-author :Jian Zheng.
E-mail:zhengjian.002@163.com

Abstract

Climate data composes of time series and space series with unknown. These unknown series contain complex co-variation relations of climate data. The extraction of these relations is essential for further revealing the complex representations between time series and space series in climate data. As an important application, through extracting these co-variation relations, we can further predict the change of climate to provide early warning for natural disasters, e.g., Greenhouse effect. Hence, it is a challenge to explore the relations between climate data. To address this, this work proposes a deep neural network. Based on Brenier theorem, the loss function is derived. Since Brenier theorem rigorously proves that the data distribution in background space is consistent with the data distribution in the feature space with greatest probability, ensuring that the relations extracted from the latent space are as close to that of in background space as possible. Then, the parameters of time series consisting of eight variables are encoded by the first hidden layer in the proposed model. The remaining two hidden layers encode the latitude and longitude in spatial series, respectively. Experimental results show that the proposed method outperforms the state-of-the-art methods with respect to climate relations extracted. Hence, the proposed method is considered a good alternative in capturing relations between climate variables, as well as, between carbon dioxide (CO_2) and surface temperature.

Keywords: Climate, neural networks, relation extraction.

1 Introduction

The climate data contains a rich source of knowledge for relations research of climate. Multiple time series and space series with unknown hide in the large amount of climate data. Usually, these series are high-dimensionality and exist by abstract complex forms. Moreover, some ineffaceable redundant information that brings a great deal of trouble for climate patterns discovered, such as, noise, etc, also hides in climate data. Facing to high-dimensional and complex climate data, because manually curating these inner relations is time consuming and expensive, there has been growing interest in developing computational approaches for automatically extracting relations from climate data. Relations extraction for climate aims to automatically extract by taking advantage of machine learning and contributes to various fields of climate research. Since the curse of dimensionality and ineffaceable redundant information of disturbing, it is a challenge for relations extraction from climate data.

Some study field, such as images, ecological and medical etc, great efforts have been made for relations extraction. In these field, some methods have been also successfully employed for automatic relations extraction, including (i) Pattern-based method, such as, pattern structures of syntactic trees in [1], similarly, method in [2], such method needs the crafting or defining of some patterns and rules according to features of tasks. Unfortunately, it is hard to defining of some patterns and rules in high-dimensional data. Moreover, Due to the diversity and complex data forms, pattern-based method is easy to suffer from low recall rates.

(ii) Feature-based method, e.g., in [3] and in [4], this method relies on variation of features, and is very skill dependent tasks. For instance, in [5], the patterns are extracted based on domain knowledge features. As well as the feature representations in [6] are also obtained based on the feature coupling generalization.

(iii) Kernel-based method, e.g., the neighborhood hash graph kernel method in [7], it can effectively capture syntactic features from the structural data, for example, a hybrid kernel-based method is used for relations extraction [8], beyond that, also using the multiple kernel methods, such as methods in [9] and in [10]. Kernel-based method needs to select suitable kernel functions, while this is a difficult task to design suitable kernel functions.

(iv) deep architecture-based method, e.g., con-volutional neural networks (CNNs) in [11], this method shows exciting potential in automatic feature learning [12] and in capturing the correct low-dimensional represents as accurately as possible [13] [14] [15]. In actual application, CNNs are widely applied in relations extraction work. For instance, CNNs in [16] and in [17] is used for medical relation extraction. In addition, the multichannel CNN [18] is also used for biological relation extraction. Certainly, sparse auto-encoders (SAEs) can also capture low-dimensional representations from high-dimensional data [19]. For instance, the multi-modal deep neural networks are used to explore informative and heterogeneous features from different feature groups [20]. Although the model in [20] learns better variables representations, the model needs to predefine loss function in different task features, which is very hard to predefine loss function in application. To learn multi-channel features representations in recognition tasks, a deep framework is designed in [21].

This indicates that deep architecture-based method has outstanding advantages to capture variable relations. This is because deep architectures are not simply to learn an identity function, more specifically, they squeeze redundant information out of data by learning under strong constraints [22]. From the perspective of the internal architecture, an encoder in deep architecture achieves the mapping of background space to latent space (feature space), while a decoder reconstructs the original input according to latent space [23]. Obviously, obtaining the data distribution in latent space becomes very critical, because this influences the final output reconstructed by a decoder. The Brenier theorem can effectively solve the issue that data distribution in latent space is close to original data distribution in background space, since the Brenier theorem can calculate the optimal distance between background space and latent space from the view of geometry.

To extract relations from climate, we designed deep neural networks possessing three hidden layers. Our primary goal is to capture the relations from climate. However, our final goal is to explore the ability of complex relations extracted using this architecture possessing deep paradigm. To achieve our studied goals, we developed the proposed model in the following steps: 1) the data in background space is sampled by the sampling theorem, to ensure that the discrete surface reconstructed quickly converge to an original surface. 2) the loss function in proposed model is derived according to Brenier theorem. Finally, we validate our thought-on climate datasets.

We summarize the main contributions of this work as follows.

1) The proposed model successfully captures these co-variation representations between climate data, implying that this architecture possessing deep paradigm is better than that of possessing non-deep paradigm to obtain relations between complex variables.

2) The decline in the correlation between seasons reduces the possibility of climate change maintaining the same pattern in different seasons.

3) The trajectory of climate change shows continuity when CO_2 concentration is low. In the areas with high surface temperature, CO_2 concentration shows a local peak, meanwhile, this local peak phenomenon stays for a while.

2 Notation and Preliminary

Lemma 1. [24]. Then for any $\upsilon = \sum_{i=1}^{k} v_i \delta_{yi}$, with $\sum_{i=1}^{k} v_i = \mu(\Omega)$, there exists $h = (h_1, ...h_k) \in \mathrm{R}^k$, unique up to adding a constant $(c,...,c)$, so that $w_i(h) = v_i$, for all i. The vectors h are exactly maximum points of the concave function. Then for any $\upsilon_1, ..., \upsilon_k > 0$ with $\sum_{i=1}^{k} v_i = \mu(\Omega)$, there exists $h = (h_1, ...h_k) \in \mathrm{R}^k$, unique up to adding a constant $(c,...,c)$, so that $w_i(h) = v_i$, for all i. The vectors h are exactly maximum points of the concave function

$$E(h) = \sum_{i=1}^{k} h_i v_i - \int_0^h \sum_{i=1}^{k} w_i(\eta) d\eta_i \quad (1)$$

on the open convex set $H = \{h \in \mathrm{R}^k \mid w_i(h) > 0, \forall i\}$. Furthermore, ∇u_h minimizes the quadratic cost $\int_\Omega |x - T(x)|^2 \, d\mu(x)$ among all transport maps $T \# u = \nu$, where the Dirac measure $\nu = \sum_{i=1}^{k} v_i \delta(y - y_i)$.

Lemma 2. Brenier theorem [25]. Suppose x and y are the Euclidean space R^n, and the transportation cost is the quadratic Euclidean distance $c(x, y) = |x-y|^2$. If u is absolutely continuous and u and v have finite second order moments, then there exists a convex function $\mu : X \to \mathrm{R}$, such that the gradient map ∇u gives the unique solution to the Monge's problem, where u is called Brenier's potential, ∇u is called Brenier mapping or the optimal transmission mass mapping. In general, u is not unique.

Lemma 1 and Lemma 2 provide a theory for deriving loss function, in section 3.1 we use them to derive the loss function in proposed model.Symbols of appearing and their meaning are given in Table 1.

Table 1 Symbol description

Symbols	Implication
h	vectors
c	a constant
H	open convex set
v	Dirac measure
T	transport map
$c(x, y)$	Euclidean distance
u	convex function
∇u	Brenier mapping
δ	a constant
ε	a constant
Ω	convex area
p	a point
φ	Kantorovich's potential
α , $\alpha \in [0,1]$	a confidence interval.
n	the number of neurons

3 Methodology

In this section, we explore how to select the loss function for the proposed model. Then, the sampling condition is given. To further analyze the trajectory of climate change, the correlation diagram is described. Finally, in subsection 3.4, the proposed model is implemented.

3.1 Loss function

Lemma 1 interprets the optimal transportation mapping from a geometric point of view. The optimal transportation mapping is exactly what we expect to find, because it helps that the distribution reconstructed is close to the original distribution. In fact, Brenier theorem (i.e., lemma 2) indicates that the optimal transmission mapping is a gradient mapping of a convex function. The convex function is also called Brenier potential energy, i.e., $\mu : \Omega \to \mathrm{R}$, where, Ω is itself convex R area. Hence, we only need calculate a gradient mapping of a convex function (or Brenier potential energy).

Reference [25] indicate that if the source probability measure u satisfies some very broad conditions, such as absolute continuity, or the finite of second moment, the optimal transmission mapping exists and is unique. The gradient mapping of Brenier's potential function is $\nabla u : \Omega \to \Omega$, which maps point $p \in \Omega$ to $\nabla u(p) \in \Omega$. Therefore, the gradient mapping should satisfy the Monge-Ampere equation, having that

$$\det(D^2 u) = \frac{u}{v \circ \nabla u} \tag{2}$$

The existence of solution of Monge-Ampere equation has been proved by Gu et al [24].

Above analyzing shows that selection loss functions is equivalent to calculate Brenier's potential of convex functions. References [26] and [27] prove that the Brenier's potential u and the Kantorovich's potential φ is related by following equation

$$u(x) = \frac{1}{2} | x |^2 - \varphi(x) \tag{3}$$

E.q (3) shows that calculation Brenier's potential can be converted to calculate Kantorovich's potential. Kantorovich's potential can be calculated by calculation Wasserstein distance [28]. Consequently, the selection of loss function is converted to calculate Wasserstein distance. In regard to the calculation of Wasserstein distance, please see in [29] [30][31].

As discussed in the section, we get a proper loss function through calculating the Wasserstein distance. Noting that there are many methods for the selection of loss function, e.g., Classification Loss, Regression Loss, etc. However, the above involved method provides a reference for the selection of loss functions.

3.2 Data sampling

To accurately discover the relations from climate, we need to consider sampling density. Since sampling density is related to the accuracy of surface reconstructed, sampling conditions should be strictly considered. Sampling conditions should be that at least there is one sampling point inside any geodesic disk with radius δ . In addition, the distance between any two sampling points is no less than the threshold ε . To ensure that the reconstructed discrete surface converges to the original smooth surface, an appropriate ε and δ should be considered. Many methods can be used to measure the distance between any two sampling points, such as Hausdorff distance, geodesic distance, curvature measure and Laplace-Beltrami operator etc.

3.3 Correlation diagram

Climate data belongs to a typical multivariate distributed data, therefore, Markov random field model is very suitable to analyze the kind of correlation to multivariate distributed data [32]. To describe the correlation between climate variables, according to the principle of Markov random field model [32] [33], the inverse co-variance matrix is used. Due to the non-zero element in the fitted inverse co-variance matrix is used to represent the conditional correlation between related variables, there should resolve the statistical confidence level of the non-zero element. If α ($\alpha \in [0,1]$) is a confidence interval, the fitted inverse co-variance matrix should mistakenly obtain a non-zero element with the probability of α [33].

The fitted inverse co-variance matrix not only reduces the number of parameters estimated, but also ensures the reliability of non-zero elements [33]. This is because it effectively controls the unnecessary correlation degree. Moreover, it also ensures the credibility between each variable with correlation. Obviously, the authenticity of method is not lost due to the over-simplification of correlation. Hence, to guarantee a sufficient statistical confidence degree of the finally constructed correlation diagram, based on above discussed, we use $\alpha = 0.05$.

4 Model

4.1 Model implementation

The proposed deep model with three hidden-layers is as shown in Fig.1. The first layer in proposed model encodes time series parameters consisted of the 8 variables (KM, KMLS, KH, KHLS, KHSFC, RI, 25(100 hPa), 27(100 hPa)). The remaining two layers encodes spatial series parameters, i.e., encoding latitude and encoding longitude. This doing is that the same layer is helpful for the relations extraction between variables of the same type. Importantly, due to each layer corresponds to a series, three layer architectures are beneficial to better maintain the consistency between time series and

spatial series, thereby encoding more compact.

Model training. For inputting dataset (see Section 5.1 for detail), 80% of the data are used as training sample for training our model. During training, we dynamically adjust the iteration epoch according the observed training precision. The training stops until our model converges.

Model testing. The left 20% of data are then used for testing sample.

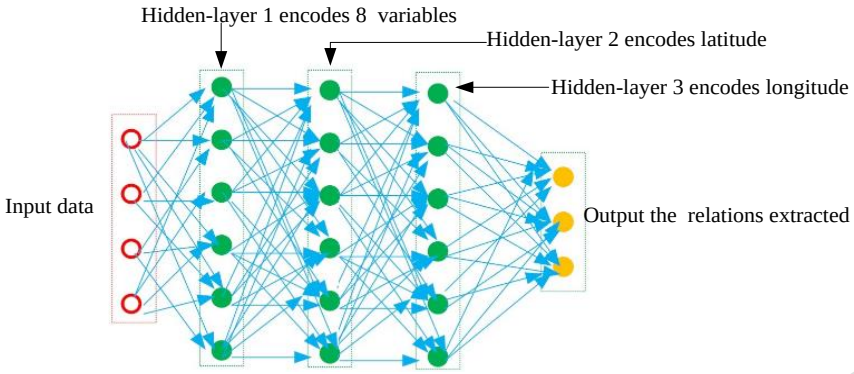


Fig.1 Model architecture. The first hidden-layer encodes eight variables. The second and third hidden-layer encode the longitude and the latitude, respectively. The eight variables are described at https://atmosphere.copernicus.eu/

From the view of inner architectures, compared with complicated deep architectures, e.g., multi-layer CNNs, Generative Adversarial Networks, our model has relatively shallow layers. In spite of this, but our model possesses a deep architecture basic paradigm, meanwhile, the loss function derived by Brenier theorem compensates for the weakness of shallow layers to a certain extent. Overall, our model is capability of achieving relations extraction from the complicated climate data.

4.2 Hyper-parameters

The proposed model has some hyper-parameters, such as neurons volume, activation functions, and learning rate, etc, therefore, we carefully studied part of them in the range of values. For other hyper-parameters, since they have no substantial impact on experimental results, their default values are adopted.

(i) Optimizer. Adam not only inherits the ability from AdaGrad to effectively deal with sparse gradient, but also has the same ability as RMSProp to handle with non-stationary targets [34]. Compared with other optimizer, such as RMSprop, SGD, Momentum and Nesterov, etc, Adam show better performance in high-dimensional data. Hence, Adam is used as an optimizer for our model.

(ii) Activation function. Common activation functions are ReLu, Sigmoid, tanh, eLU, etc. We verify the four activation functions. In addition, the suitable number of neurons is also explored in range of from $\sqrt[2]{n}$ to $\sqrt{n}$ (n is data volume). Through analyzing AUC value of hyper-parameters, we adjusting them accordingly. The tuned results are that the number of neurons is $\sqrt{n}$, and the ReLu is used as activation function in the first and second layer, and the tanh is used in the third layer.

(iii) Learning rate. Adam can automatically provide an adaptive learning rate for different learning tasks, so there is no need to manually configure the learning rate.

5 Experimental setting

5.1 Dataset

We explore the relations of climate in an annual cycle, which provides a reference for climate change in the next an annual cycle. So climate data and CO_2 data in an annual cycle are used as the studied objects.

The ECMWF climate data is used in this work (https://atmosphere. copernicus.eu/), of which each sample has 512 components, consisting of 8 variables (KM, KMLS, KH, KHLS, KHSFC, RI, 25(100 hPa), 27(100 hPa)) at different locations, i.e., 8 latitudes and 8 longitudes, in Table 2.

Table 2 Datasets description.

Dataset	Description	Data volume	Dimensionality
Climate	low spatial, 8 latitudes, 8 longitudes, 8 variables	40000	512
CO_2	low spatial, 8 latitudes, 8 longitudes, 8 variables	65000	512

5.2 Comparison methods and assessment metrics

We opt for the three state-of-the-art of typical relation extraction methods, i.e., (I) Feature-based method, method in [5]. (II) Kernel-based method, method in [8]. (III) Deep architecture-based method, method in [17]. For the three competitors, their optimal parameters observed in the corresponding literature are used. Unless otherwise stated, all experiments are run on the same experimental settings.

Receiver operating characteristic curve (ROC) and corresponding area under the curve (AUC) are used to assess the precision of relations extraction.

6 Results and discussion

In this section, we address entire experimental results. The low-dimensional representations from these high-dimensional climate data are presented, and these relations extraction are visualized .

6.1 Extraction precision

The AUCs of methods are addressed in Fig.2. It can be seen that our method outperforms competing methods in extracted precision on climate dataset and CO_2 dataset. As for the two datasets, the extracted accuracy of our method reaches above 85%. While for the three competitors are below 75%. In addition, deep method, e.g., method in [17], is superior to traditional method, e.g., method in [5] and [8]. This implies that this architecture possessing deep paradigm is better than that of possessing non-deep paradigm to obtain relations between complex variables.

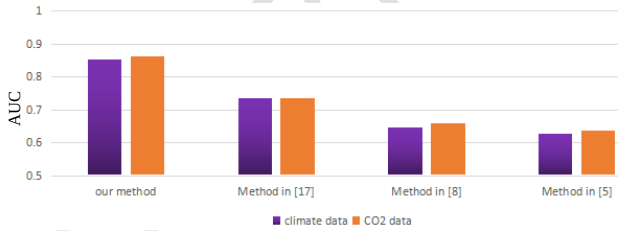


Fig.2 Comparison of accuracy extracted

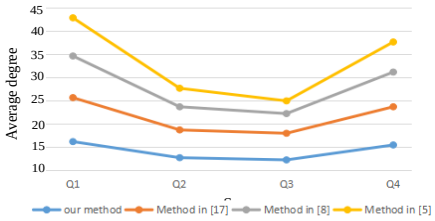


Fig.3 Correlation diagram

6.2 Change trajectory of climate

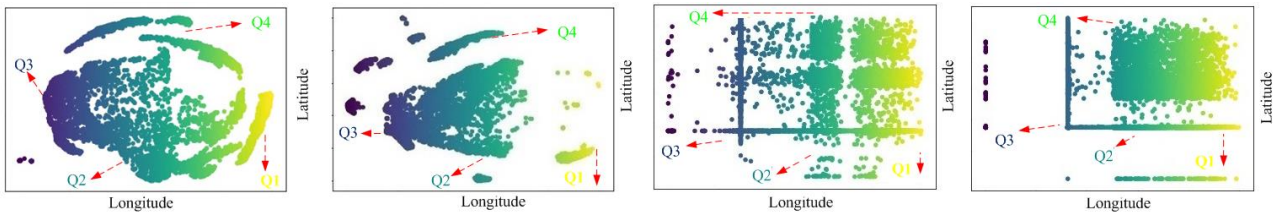
Fig.3 addresses the correlation diagram. In Section 3.3, we determine the statistical confidence degree of the fitted inverse co-variance matrix, so this is feasible to directly study the number of edges in correlation diagram. Fig.4 displays the trajectory of climate change in an annual cycle, and we visualize the trajectory using 2-dimension.

Results indicate that in the first three quarters, the number of edges in correlation diagram drops, but in the fourth quarter show a rising trend, in Fig.3. Through analyzing Fig.3 and Fig.4, several observations can be obtained.

(i) Our method and competing methods find the trajectory of climate change. However, the trajectory discovered by our method outperforms the three competitors.

(ii) The trajectory climate change shows obviously partial continuity in some seasons as shown in Fig.4, e.g., Q1 and Q4, meaning that climate presents periodic change (partial continuity) and a-periodic change.

(iii) The decreasing number of edges demonstrates that the correlation of between seasons is decreasing, implying that maintaining the same regularity in different seasons is getting lower and lower.

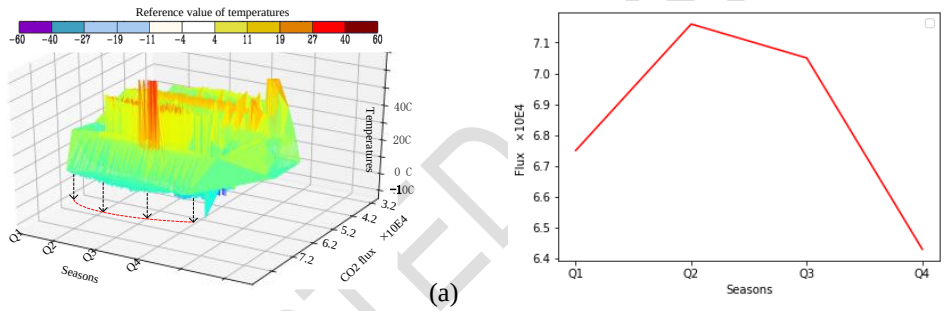


(a) our method (b) method in [17] (c) method in [8] (d) method in [5]

Fig.4. Climate trajectory. Colors represent quarters of a year, i.e., first quarter (Q1), second quarter (Q2), third quarter (Q3), fourth quarter(Q4).

6.3 Co-variation relations extracted

The relations of CO_2, surface temperatures and seasons are addressed in Fig.5 (a). Results show this encoding method, i.e., our method, can accurately capture the co-variation relations of atmospheric carbon dioxide flux, time series and surface temperatures. Obviously, CO_2 flux reaches the maximum amount in second quarter (Q2), as shown in Fig.5 (b).



(a)

CO_2 flux and temperatures (b) CO_2 flux and seasons

Fig.5. Relations of CO_2 flux, surface temperatures and seasons.

Several observations can be obtained from Fig.5, having that

(i) CO_2 concentration presents a local peak in the areas of presenting higher surface temperature, e.g., 40^o *C*. This phenomenon of local peak occurs concentrated within a certain period of time, i.e., CO_2 concentration concentrates with a high probability in Q2.

(ii) CO_2 concentration is related to the trajectory of climate change. In Q2, CO_2 concentration is higher than the other 3 quarters in Fig.5 (b). Correspondingly, the trajectory of climate change in Q2 shows discontinuity, as shown in Fig.4 (a). However, the continuity of trajectory in Q1 and Q4 is more obvious than that of in Q2 and Q3, meanwhile, CO_2 concentration in Q1 and Q4 is lower than that of in Q2 and Q3.

Above results indicate that the proposed model can successfully learn relations between variables through filtering this non-eigenvalue information from complicated data. This indicates that models possessing deep architectures paradigm has more advantages than that of owning non-deep architectures in capturing relation between complex variables.

7 Conclusion

In this work, we investigated the issue of the relations extraction from complex climate variables. To address this, a deep neural network is designed to explore this interesting relations in this climate variables. We successfully capture the co-variation relations CO_2 flux, surface temperatures and seasons, thereby further revealing the complex representations between time series and space series in meteorological data.

Funding

The research funding is Supported by the Science and Technology Research Program of Chongqing Municipal Education Commission of China under Grant KJQN201903003. And the Science and Technology Research Program of Chongqing Municipal Education Commission of China under Grant KJQN202003001.

Declarations

All authors have no conflicts of interest to declare that are relevant to the content of this article. All authors agree with availability of data and material in this work.

References

1. Zhang ming, Zhou Guodong, AW, Aiti. Exploring syntactic structured features over parse trees for relation extraction using kernel methods. Information Processing & Management. 2008; 44(2):687-701.
2. K. Raja, S. Subramani, J. Natarajan. PPInterFinder—a mining tool for extracting causal relations on human proteins from literature. Database. 2013.
3. S. Kim, H. Liu, L. Yeganova, W.J. Wilbur. Extracting drug-drug interactions from literature using a rich feature-based linear kernel approach. J. Biomed. Informat. 2015; 55:23-30.
4. A. Raihani, N. Laachfoubi. Extracting drug-drug interactions from biomedical text using a feature-based kernel approach. J. Theor. Appl. Informat. Technol. 2016; **92** (1):109.
5. J. Björne, S. Kaewphan, T. Salakoski. UTurku: drug named entity recognition and drug-drug interaction extraction using SVM classification and domain knowledge. International Workshop on Semantic Evaluation. 2013. p.651-659.
6. Y. Li, X. Hu, H. Lin, Z. Yang. Learning an enriched representation from unlabeled data for protein-protein interaction extraction. BMC Bioinformatics. 2010; 11(2):S7.
7. Y. Zhang, H. Lin, Z. Yang, Y. Li. Neighborhood hash graph kernel for protein-protein interaction extraction. J. Biomed. Informat. 2011; 44 (6):1086-1092.
8. M.F.M. Chowdhury, A. Lavelli. FBK-irst: a multi-phase kernel based approach for drug-drug interaction detection and classification that exploits linguistic information. Atlanta, Georgia, USA. 2013; 351:53.
9. P. Thomas, M. Neves, T. Rocktäschel, U. Leser. WBI-DDI:drug-drug interaction extraction using majority voting. DDI Challenge at Semeval. 2013. p.628-635.
10. W. Zheng, H. Lin, Z. Zhao, B. Xu, Y. Zhang, Z. Yang, J. Wang. A graph kernel based on context vectors for extracting drug-drug interactions. J. Biomed. Informat. 2016; 61:34-43.
11. S.Zheng, Y. Hao, D. Lu, H. Bao, J. Xu, H. Hao, B. Xu. Joint entity and relation extraction based on a hybrid neural network. Neurocomputing. 2017; 257:59–66.
12. Z. Zhao, Z. Yang, L. Luo, H. Lin, J. Wang. Drug drug interaction extraction from biomedical literature using syntax convolutional neural network. Bioinformatics. 2016; 32(22):3444-3453.
13. Jianzhu Ma, Michael Ku Yu , Samson Fong. Using deep learning to model the hierarchical structure and function of a cell. Nature Methods. 2018; 15:290–298.
14. Y. Goldberg, A. Zakai, D. Kushnir, Y. Ritov. Manifold learning: The price of normalization. J. Mach. Learn. Res. 2018; 9:1909–1939.
15. Le Cun, Y., Bengio, Y. & Hinton, G. Deep learning. NATURE. 2015; 521:436–444.

16. Z. Zhao, Z. Yang, H. Lin, J. Wang, S. Gao. A protein-protein interaction extraction approach based on deep neural network . Int. J. Data Min. Bioinformatics. 2016; 15(2):145-164.
17. S. Liu, B. Tang, Q. Chen, X. Wang. Drug-drug interaction extraction via con-volutional neural networks. Comput. Math. Methods Med. 2016.
18. C. Quan, L. Hua, X. Sun, W. Bai. Multichannel convolutional neural network for biological relation extraction. BioMed Res. Int. 2016.
19. Xinyu Guo, Ali A. Minai, Long J. Lu. Feature Selection using Multiple Auto-Encoders. IEEE 2017 International Joint Conference on Neural Networks (IJCNN). 2017. p.4602-4609.
20. L. Zhao, Q. Hu, W. Wang. Heterogeneous Feature Selection with Multi-Modal Deep Neural Networks and Sparse Group Lasso. IEEE Transactions on Multimedia. 2015; 17:1936-1948.
21. Z. Zhu, P. Luo, X. Wang, X. Tang. Deep learning multi-view representation for face recognition. arXiv preprint arXiv. 2014; 1406.6947.
22. P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. The Journal of Machine Learning Research. 2010; 11:3371-3408.
23. Y. Bengio, L. Yao, G. Alain, P. Vincent. Generalized denoising auto-encoders as generative models, in Advances in Neural Information Processing Systems. 2013. p.899-907.
24. Gu, Feng Luo, Jian Sun, S.-T. Yau. Variational principles for minkowski type problems, discrete optimal transportation", and discrete monge-ampere equations. Asian Journal of Mathematics (AJM). 2016; 20(2):383-398.
25. Brenier, Yann. Polar factorization and monotone rearrangement of vector-valued functions. Commun. Pure Appl. Math. 1991; 44(4):375–417.
26. Kehua Su, Wei Chen, Na Lei, Junwei Zhang, Kun Qian, Xianfeng Gu. Volume preserving mesh parameterization based on optimal mass transportationation. Computer-Aided Design. 2017; 82: 42-56.
27. Chen, Haodi, Huang, Genggeng, Wang, Xu-Jia. Convergence rate estimates for aleksandrov's solution to the monge-ampere equation. Siam Journal On Numerical Analysis. 2019;57(1):173-191.
28. Na Lei, Kehua Su, Li Cui, Shing-Tung Yau, Xianfeng David Gu. A geometric view of optimal transportation and generative model. Computer Aided Geometric Design. 2019; 68:1-28.
29. Kantorovich, L.V. On a problem of Monge. *Usp.* Mat. Nauk. 1948; 3:225–226.
30. Villani Cédric. Topics in optimal transportation. graduate studies in mathematics. American Mathematical Society, Providence, RI. 2003; 58.
31. Villani Cédric. Optimal Transport: Old and New. Springer Science & Business Media. 2008; 338.
32. Rue H, Held L. Gaussian Markov Random Fields: Theory and Applications Monographs on Statistics and Applied Probability. London:Chapman & Hall. 2005; 104.
33. Meinshausen N, Buhlmann P. High-dimensional graphs and variable selection with the Lasso. Annals of statistics. 2006; 34(3):1436.
34. Diederik P. Kingma, Jimmy Lei Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv. 2015;1412.6980v8.