Representation Learning and Translation between the Mouse and Human Brain using a Deep Transformer Architecture

Minxing Pang [1] **Jesper Tegnér** [1 2]

Abstract

Recent progress in single-cell genomics has produced large single-cell data-sets of cell-types and organs from mouse and human samples. Since it is still difficult to transfer data across species we conceptualize this problem as language translation task between mouse and human, requiring a latent space in which we can translate accordingly. To this end, we developed a deep learning module using a transformer architecture and using the Mouse and Human Brain Atlas to embed the data as a piece of a manifold amenable for cross-species translation. Neither PCA and SAUCIE could align the data across species in this manner, and with our representation we also demonstrate better clustering accuracy (86.7%) compared with PCA and SAUCIE. Computing Wasserstein distances between all cell-types and tissue types demonstrated overall a shorter distance in the latent space between similar cell types across human and mouse. This suggests that the intrinsic geometry of the organization of the nervous system between species share sufficient structure to enable translation across species.

1. Introduction

Rapid progress in the development of single-cell RNA sequencing (scRNA-seq) technologies in recent years has provided many valuable insights into complex biological systems (Stuart & Satija, 2019; Shalek et al., 2014). For example, scRNA-seq profiling of the adult mouse nervous system has provided a deep mapping of cell-types in different organs and during development, and these studies are in effect producing a reference atlas for studying the mammalian nervous system (Zeisel et al., 2018). The processing of such data presents new challenges, and machine learning algorithms have proved to be useful to address topics such as low-level bioinformatics analysis, clustering (Butler et al., 2018), visualization (Moon et al., 2019), missing datapoints, i.e. imputation (Wang et al., 2019), data integration (Hie et al., 2019) and prediction of perturbation responses (Lotfollahi et al., 2019). A common theme across these methods is the representation of the single-cell gene expression count matrix as a low-dimensional object for either efficient data pre-processing or downstream analysis including data fusion within or between different data-modalities.

Interestingly, these large amounts of data are currently being generated in a number of different species, including the mouse model, which is an important model organism for development, diseases, and drug development. It would therefore be advantageous to be able to transfer information, data, and insights between different model systems. Yet, this is still an open challenge since current tools as a rule targets a single data-modality or organ system within the confinement of a given species. Such cross-species comparisons are complicated by many biological and technical factors (Shafer, 2019). Some early work in this direction attempting to align the cross-species data sets includes (Ding et al., 2019; Stuart et al., 2019). Nevertheless, they are either based on a biological process rather than data or project the data points(cells) to several clusters that contains the same cell type in each cluster. Here we address this problem by asking to what extent we can learn a mapping or a translation between the human and mouse neural systems. We rephrase the cross-species alignment as a language problem in that we would like to translate the language within one system (mouse brain) into the other language (human brain). Using the language analogy we search for a latent space representation of the problem that enables such a translation thus avoiding the high-dimensionality of the original single-cell gene expression matrix. In natural language processing, the essence of neural machine translation (Bahdanau et al., 2015) is an embedding-transformation process, which can be classified as generalized data fusion. Here we explore and adapt a successful Transformer language neural network architecture to the problem of cross-species alignment, targeting a translation between the human and mouse nervous system.

[*]Equal contribution [1]King Abdullah University of Science and Technology, Saudi Arabia [2]Karolinska Institutet, Sweden. Correspondence to: Jesper Tegnér <jesper.tegner@kaust.edu.sa>.

Presented at the ICML 2020 Workshop on Computational Biology (WCB).

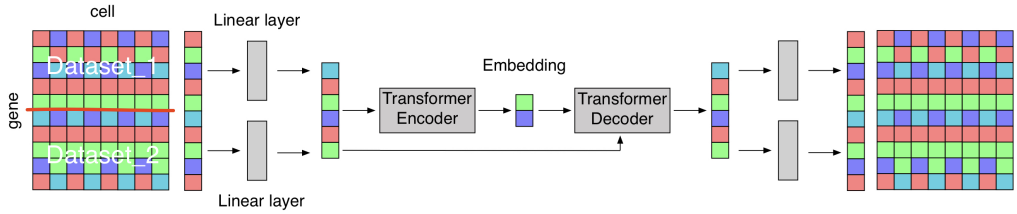

Figure 1. Overview of the model architecture

2. Model

2.1. Neural Network Architecture

Our model contains three main parts: linear transformation, transformer encoder and transformer decoder. It should be noted that in natural language processing, the input data is in order, which mainly includes three parts of embedding, that is, word vector embedding, position embedding, and language type embedding. Because the data of the single-cell expression matrix is order-independent, our model discards the position code and only contains the gene code, that is, the source code, and can use the information capture capabilities of the encoder and decoder in the language translation model.

As shown in Figure 1, suppose we have dataset $X_A \in \mathbb{R}^{a\times n}$ and dataset $X_B \in \mathbb{R}^{b\times m}$, where a the amount of cells in the dataset X_A, n is the amount of genes in the dataset X_A, b the amount of cells in the dataset X_B, m is the amount of genes in the dataset X_B, the values of the matrix represent gene expression values for each cells. Notice that the expression matrix is sparse(typically over 80% of the values are zero).

Through the linear layer W_a and W_b, we can obtain the lossless intense representation of the expression matrix(since we can reconstruct the raw expression matrix by simple inverse transformation). e.g, the lossless intense representation of X_A and X_B are $X_A^* = W_a X_A \in \mathbb{R}^{a\times n}$ and $X_B^* = W_b X_B \in \mathbb{R}^{b\times m}$. (Devlin et al., 2018) shows that randomly mask some of the input and impute back is an efficient way to train a large model in an unsupervised approach. Then we randomly mask 10% of the data, which is a method to make the model more robust.

After linear layer, let $C_A = S_A W_A \in \mathbb{R}^{1\times n}$ and $C_B = S_B W_B \in \mathbb{R}^{1\times m}$ be the trainable source embedding that denote the source of the data, where $S_A \in \mathbb{R}^{1\times k}$ and $S_B \in \mathbb{R}^{1\times k}$ are one-hot vectors that denote the dataset, and k is the amount of data sets. Then we extend the C_A and C_B to the matrix whose shapes are (a, n) and (b, m) The input of the Transformer Encoder can be calculate as $I_A = X_A^* + C_A$ and $I_B = X_B^* + C_B$

As a result we obtain the embeddings E_A and E_B from Transformer Encoder

$$E_A = MultiHead(I_A, I_A, I_A) \tag{1}$$

$$E_B = MultiHead(I_B, I_B, I_B) \tag{2}$$

Inspired by the BERT (Devlin et al., 2018), we utilize Attention mechanism to merge the information after Transformer Encoder

$$I'_A = Attention(I_A, E_A, E_A) \tag{3}$$

$$I'_B = Attention(I_B, E_B, E_B) \tag{4}$$

And the output of Transformer Decoder can be calculated as

$$X'_A = MultiHead(I'_A, I'_A, I'_A) \tag{5}$$

$$X'_B = MultiHead(I'_B, I'_B, I'_B) \tag{6}$$

Finally, we use Mean Square Error $L_A = \|X_A - X'_A\|_2$ and $L_B = \|X_B - X'_B\|_2$ as loss function for reconstruction. Since the intense representation is lossless transformation of the raw data, it is equivalent to reconstruct data in the raw space.

2.2. Alignment in the latent manifold space as derived from scRNA-seq data

Here we make the explicit assumption that the cells in different biological systems can have similar functions and such functions are encoded in expression of different genes. Using the analogy from language translation problem, we benefit from that the embedding of English and German have a similar structure. Since we live in the same world, different languages can be seen as the different representations of the same things. The difference is that in scRNA-seq, the function of cells might be much different, and some datasets might contain the cells that are not included in another data set. Yet, based on the corresponding assumption, that the "same" system in two different species is performing similar tasks based on a similar architecture, the shape of the manifold of the different datasets in the embedded space is similar.

Our model therefore have the objective of learning an alignment between two sets such that the manifold shape is similar in the shared embedding space. In this paper, we use the adversarial criterion to measure the similarity, which is denoted as $\|\cdot\|_D$. We denote the source dataset as $\mathscr{X}$ and the target matrix as $\mathscr{Y}$, what we want to do is to find a transformation W^* such that

$$W^* = \underset{W \in M_d(\mathbb{R})}{\arg\min} \|WX - Y\|_D \tag{7}$$

where d is the dimensionality of the embeddings, $M_d(\mathbb{R})$ is the space of $d \times d$ matrices of real numbers, X is matrix of size $d \times n$ sampled from $\mathscr{X}$, Y is matrix of size $d \times n$ sampled from $\mathscr{Y}$

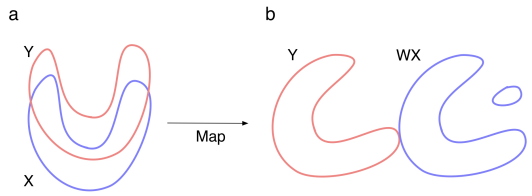

Figure 2. Overview of mapping. (**a**) Dataset X and dataset Y in the embedding space. (**b**) After mapping, assume that dataset X contains more cell types, dataset X would separate into two parts, one of them is similar to dataset Y. And the other part contains the cells that very different from all the cells in dataset Y.

The rationale for employing a linear transformation to align the data and to use GANs for training, originates from recent results on word translation. Here a (Mikolov et al., 2013) achieved better results on the word translation task compared to more advanced strategies like multi-layer neural networks. Finally, (Lample et al., 2018) construct linear language mapping through adversarial training, which demonstrates the effectiveness of learning linear mapping by GANs.

2.3. Data-sets, preprocessing, and parameters

Data-sets. We performed computational experiments on Human Brain Atlas and Mouse Brain Atlas, where we demonstrate the visualization of the result and compare the predictive performance of our proposed method against SAUCIE (Amodio et al., 2019) and PCA (Lever et al., 2017). Human Brain Atlas (HBA) (https://portal.brain-map.org/atlases-and-data/rnaseq#Human_Cortex) include 50,000 genes of 25,000 cells for more than 10 cell types. And Mouse Brain Atlas (MBA) (http://mousebrain.org/) includes 25,000 genes of 160,000 cells for more than 16 cell types, which provides a clearer picture of cell diversity by region and a reference atlas for studying the mammalian nervous system.

Preprocessing. Cells and genes were filtered using the python package scprep (https://github.com/KrishnaswamyLab/scprep). Secondly, we log-transformed gene expression profiles. For each data set, we then scaled it to $[-50, 50]$. Finally, for each data set, we project it to 1200 dimensions as intense representation using linear Principle Component Analysis algorithm.

Parameters. In (i) the embedding stage, the batch size is 1024, the learning rate is 0.0003, the embedding size for Transformers is 1024, the total amount of parameters in the embedding model is 300 million. (ii) In the alignment stage, we choose one layer linear networks without activation function as the generator and Multi-Layer Perceptron as the discriminator, the learning rate is 0.00003, the batch size is 1024. In the embedding task, we compared several machine learning algorithms, namely, multitask SAUCIE (Amodio et al., 2019), PCA and our Transformers based model on two data sets. We randomly selected 1000 cells in ever cell type for visualization.

3. Results

Here we first show that the embedding of our model aligns the data for human and mouse at the level of cell-type/tissues, sufficient to reconstruct each data-sets, and detecting parts of the data which does not match across species.

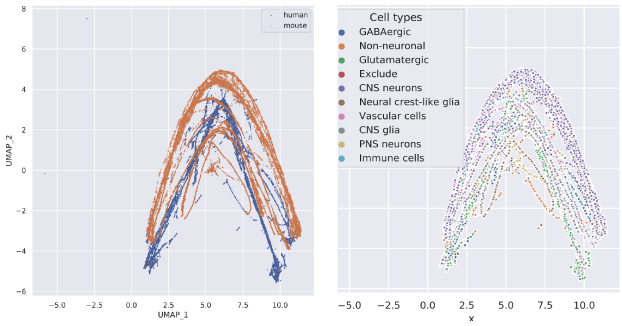

Figure 3. The embedding from our models, the top four labels in the right subfigure belong to Human Brain Atlas, others are from Mouse Brain Atlas

Figure 3 illustrates that the embeddings using our model preserved biological knowledge better than the other two models. For example, for each cell type, the distribution of cells in the embedding space is continuous and separated. Here the 1024-dimension embeddings are visualized in two-dimension space. Note that since the two datasets (HBA and MBA) are operated through the same Transformer encoder and decoder, the embeddings of them are similar with respect to the shape of the manifolds. Moreover, since a

Table 1. Average Wasserstein distance from Human to Mouse for Figure 5

	CNS	PNS	NON-NEURAL
CNS	7.91 ± 0.012	8.85 ± 0.011	8.3 ± 0.003

source embedding is added to both data set prior being processed by the Transformer Encoder, the HBA data set and MBA data set can readily be separated.

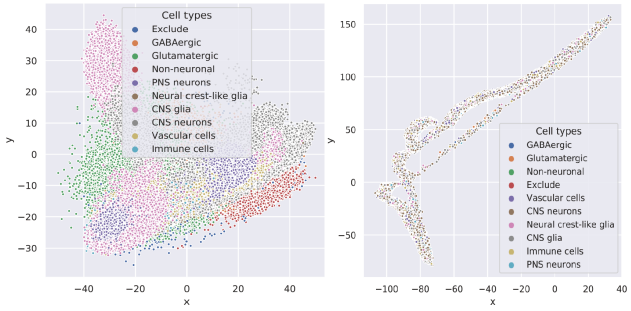

Figure 4. The left subfigure is 2-D embedding from PCA, and the right figure is The 2-D embedding from SAUCIE

Such an embedding using the transformer, contrasts with both a PCA or SAUCIE embedding. Performing a dimensionality reduction using PCA, into the 2-dimensional space (Figure 4), we find that the PCA embedding model cannot distinguish between different species or cell types. Using the SAUCIE model, with its default parameters, to first embed high-dimensional single-cell data with PCA into 1024 dimensions, and use the multi-layer perceptron model to embed in the 2-dimensional space. As evident when visualizing the embedding (Figure 4) and compared with Figure 3, the SAUCIE representation does not capture the intrinsic geometry of the data-sets.

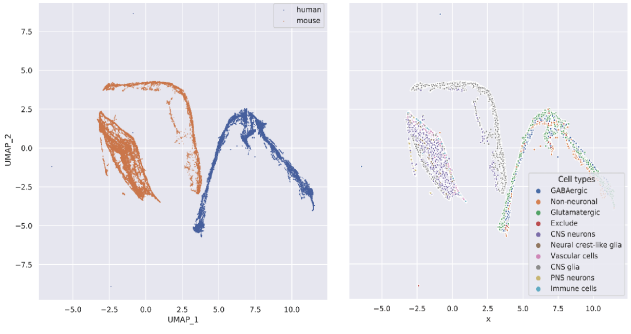

Figure 5. Data integration results. the top four labels in the right subfigure belong to Human Brain Atlas, others are from Mouse Brain Atlas

Next, we asked to what extent would the species alignment (Figure 5) using a transformer be informative from a biological standpoint. As illustrated in the Method section, it is to be expected that different parts of the data would be separated from the main manifold after an alignment. Interestingly, this was confirmed using the data. The left part of data points is from MBA data set, which contains the cells from *CNS*, *PNS* and *ENS* while the right part of the data only from *CNS*. And the right part of the MBA data set, which is from *CNS*, is similar to the HBA data set. To quantify such a similarity we computed the wasserstein distances between all cell-types and tissue types. A proper alignment would position the human *CNS* closer to the mouse *CNS* as compared with other mouse tissues. This was indeed confirmed (Table 1).

Table 2. Hierarchical cluster accuracy

MODELS	TAXONOMY-1	TAXONOMY-2
SAUCIE	66.7%	53.3%
PCA	80.0%	60.0%
OURS	**86.7%**	**66.7%**

Next we asked whether our embeddings, evidently useful for cross-species mapping, would be beneficial for clustering. To this end we performed hierarchical cluster algorithm (Rokach & Maimon, 2005) on the embeddings of Mouse Brain Atlas and compared the hierarchical cluster results with the ground truth. Our model performs well in this unsupervised classification problem, the accuracy rate at the first taxonomy reaches 86.7%. The only two wrong classification results are Sypathetic and Enteric. However, the size of these two categories is relatively small, so this error may be caused by the imbalance of the data set. However, for a relatively large set of cell types that have undergone extensive training, the accuracy is higher.

In conclusion, the results suggest that the intrinsic geometry of the organization of the nervous system between species share sufficient structure to enable translation across species.

We compare the result with SAUCIE and PCA in Table 2. In SAUCIE, they directly compress the data from high-dimension space to 2-dimension space. It is probable that the embedding is highly compressive, so its hierarchical information can not be discovered by hierarchical cluster directly. In conclusion, our models show that Transformers architecture preserves the biological information well when it was used to embed the scRNA-seq expression matrix from high dimensional space to relatively low-dimension space.

References

Amodio, M., van Dijk, D., Srinivasan, K., Chen, W. S., Mohsen, H., Moon, K. R., Campbell, A., Zhao, Y., Wang, X., Venkataswamy, M., Desai, A., Ravi, V., Kumar, P.,

Montgomery, R., Wolf, G., and Krishnaswamy, S. Exploring single-cell data with deep multitasking neural networks. *Nature Methods*, 16(11):1139–1145, 2019.

Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, 36(5):411–420, 2018.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

Ding, H., Blair, A., Yang, Y., and Stuart, J. M. Biological process activity transformation of single cell gene expression for cross-species alignment. *Nature Communications*, 10(1):4899, 2019.

Hie, B., Bryson, B., and Berger, B. Efficient integration of heterogeneous single-cell transcriptomes using scanorama. *Nature Biotechnology*, 37(6):685–691, 2019.

Lample, G., Conneau, A., Ranzato, M., Denoyer, L., and Jégou, H. Word translation without parallel data. In *International Conference on Learning Representations*, 2018.

Lever, J., Krzywinski, M., and Altman, N. Principal component analysis. *Nature Methods*, 14(7):641–642, 2017.

Lotfollahi, M., Wolf, F. A., and Theis, F. J. scgen predicts single-cell perturbation responses. *Nature Methods*, 16 (8):715–721, 2019.

Mikolov, T., Le, Q. V., and Sutskever, I. Exploiting similarities among languages for machine translation, 2013.

Moon, K. R., van Dijk, D., Wang, Z., Gigante, S., Burkhardt, D. B., Chen, W. S., Yim, K., Elzen, A. V. D., Hirn, M. J., Coifman, R. R., Ivanova, N. B., Wolf, G., and Krishnaswamy, S. Visualizing structure and transitions in high-dimensional biological data. *Nature Biotechnology*, 37 (12):1482–1492, 2019.

Rokach, L. and Maimon, O. Clustering methods. In Maimon, O. and Rokach, L. (eds.), *Data Mining and Knowledge Discovery Handbook*, Data Mining and Knowledge Discovery Handbook, pp. 321–352. Springer US, Boston, MA, 2005.

Shafer, M. E. R. Cross-species analysis of single-cell transcriptomic data. *Frontiers in Cell and Developmental Biology*, 7(175), 2019. ISSN 2296-634X. doi: 10.3389/fcell.2019.00175.

Shalek, A. K., Satija, R., Shuga, J., Trombetta, J. J., Gennert, D., Lu, D., Chen, P., Gertner, R. S., Gaublomme, J. T., Yosef, N., Schwartz, S., Fowler, B., Weaver, S., Wang, J., Wang, X., Ding, R., Raychowdhury, R., Friedman, N., Hacohen, N., Park, H., May, A. P., and Regev, A. Single-cell rna-seq reveals dynamic paracrine control of cellular variation. *Nature*, 510(7505):363–369, 2014.

Stuart, T. and Satija, R. Integrative single-cell analysis. *Nature Reviews Genetics*, 20(5):257–272, 2019.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, William M., I., Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902.e21, 2019.

Wang, J., Agarwal, D., Huang, M., Hu, G., Zhou, Z., Ye, C., and Zhang, N. R. Data denoising with transfer learning in single-cell transcriptomics. *Nature Methods*, 16(9): 875–878, 2019.

Zeisel, A., Hochgerner, H., Lönnerberg, P., Johnsson, A., Memic, F., van der Zwan, J., Häring, M., Braun, E., Borm, L. E., La Manno, G., Codeluppi, S., Furlan, A., Lee, K., Skene, N., Harris, K. D., Hjerling-Leffler, J., Arenas, E., Ernfors, P., Marklund, U., and Linnarsson, S. Molecular architecture of the mouse nervous system. *Cell*, 174(4): 999–1014.e22, 2018.