Lecture: Introduction to Interactive Visual Data Analysis (IVDA) (2021)
J. Bernard (Professor)

Visualizing Neural Networks with the Grand Tour

Agarwal Nidhi[1] and Kumar Abhinav[2]

[1]Zurich University, Switzerland 21-717-673
[2]Zurich University, Switzerland 21-744-115

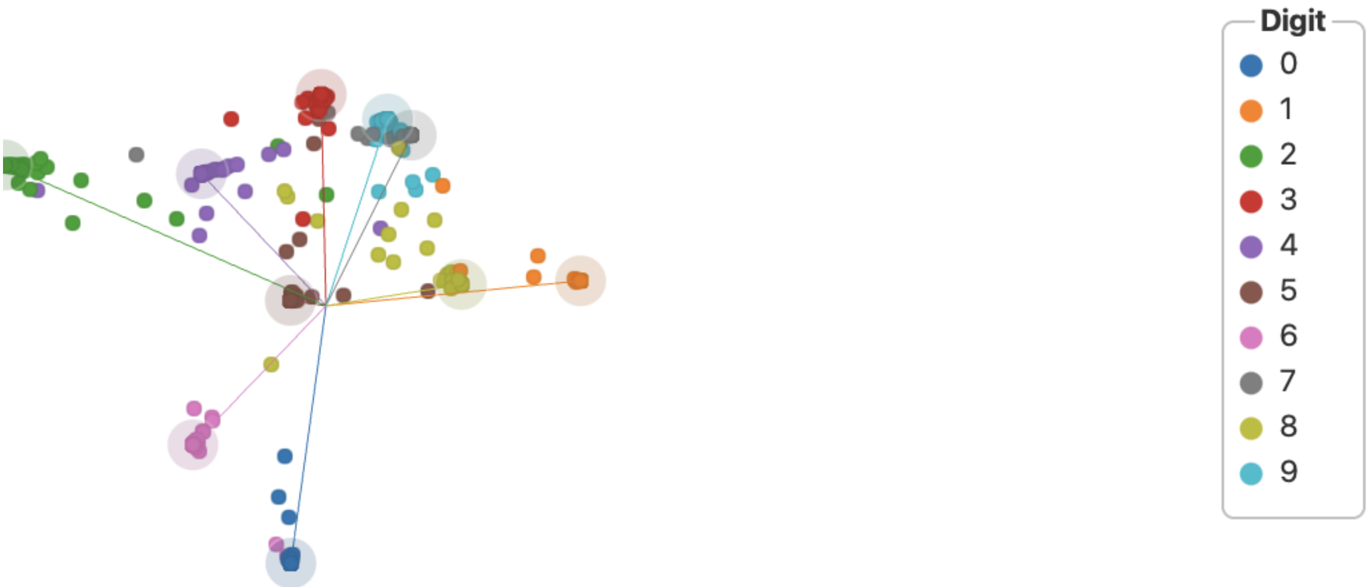

Figure 1: The Grand Tour visualization[1]

Abstract

Grand Tour is a visualization technique created to project a high-dimensional dataset into 2 dimensions of screen. It uses animation to smoothly project all faces of dataset in a gradual way for better visualisation to the viewer. Compared to other non-linear projection methods, Grand Tour is based on linear method. This report shows how the linearity of Grand Tour aids in understanding the process of deep learning and in comparing various training epoch during training phase.

1. Introduction

Neural network is a machine learning model, modeled after the human brain, that is designed for classification and regression tasks. This model is good for identifying simple patterns but shows limitations while applied to complex real objects like handwriting and image categorization. To overcome this, neural network with a large number of hidden layers known as deep learning also known as stacked neural network is introduced. It's hard to understand the deep learning process as they are difficult to understand and interpret. We need to look into different states of the network on training data and how weight gets updated through different layers.

Primitive way for understanding Neural Network is through observing its output based on input data and visualisation is created based on it. This can give a picture of activation patterns of neural networks, however it doesn't explain the relationship between different datasets for a given model and how data flow through the neural networks. Grand Tour is developed to visualize the response of a neural network and shows relationships between weights when data is changed. It tells us through visualisation the difference between training epochs and how models converge or diverge on a given dataset. It does that by creating a random rotation of the dataset and then projecting this on a two-dimensional screen. It is developed with keeping consistency in design so that if data had been changed visualisation would change accordingly.

2. Related Work

t-SNE (t-distributed stochastic neighbor embedding)[5] is a technique to reduce nonlinear dimensionality and visualize high-dimensional data in a low-dimensional space of two or three dimensions. Alike data objects are modeled by neighboring points and unalike data objects are modeled by points which are distant and have high probability. t-SNE can be further modified to allow a trade-off between temporal coherence and spatial coherence. The adaptation can be done in a controlled fashion (structure is preserved at a specific time step). This nonlinear projection method is called Dynamic t-SNE[2]. UMAP (Uniform Manifold Approximation and Projection)[4] is similar to t-SNE and can be applied also for general nonlinear dimension reduction. It looks for closest possible fuzzy topological structure in lower dimension that project higher dimensional data. Principal Component Analysis (PCA) is the classic technique of linear dimensionality reduction, that chooses to project data in order to preserve the most variance possible. This visualization technique makes it possible to observe patterns in lower dimensions that are relevant in complete parameter space, where it would otherwise be unclear by the high dimensionality of the parameter space.

t-SNE, dynamic t-SNE and UMAP are better at providing two-dimensional images where similar points tend to be clustered together very effectively but not so particularly good to understand the behavior of neuron activations at a finer scale. For both t-SNE and UMAP, the position of each single data point depends essentially on the whole data distribution in embedding algorithms. This makes it harder to judge the underlying change in data from the change in visualization as the model fails to establish a direct correspondence between data and visuals. Visualization of classifier behavior is better with Grand Tour which also has the capability of direct manipulation. Through the training epochs, PCA projections are decipherable and consistent. However, the decision to select the principal component of softmax activation becomes difficult if for example the first two principal components are not considerably better than the third. Therefore, instead of PCA, this data can be visualized by smoothly animating random projections, using the Grand Tour.

3. Relation to IVDA

The Grand Tour method is a direct example of Model Explainer which was taught to us in Week 11 of Interactive Visual Design. Grand Tour is a linear visualisation tour that assists in capturing the essence of models and helps in understanding the progress of models. Lack of transparency in deep learning models provides a challenge for its usage in real world scenarios. Tools like Grand Tour provide insight into its functioning and how the models reach its output can be easily understood by this tool. Interactive view which is part of Grand Tour which helps in viewing model states at each epoch can be recorded for further analysis.

As it was discussed in lecture, there is demand for visualisation tools that provide reliable insight about the model to amplify the trust. As deep learning has gained significant traction it is important to develop and verify such tools. These tools give better understanding about models like how a model works correctly and why a model fails and what can be done to improve model performance. There are many stages in machine learning where visualisation can enhance the experience of workflow. Grand Tour is aiding in feature/parameter analysis and also to understand Deep Neural Network. This aligns with the VIS4ML goals stated in [3]. In the deep learning process, visualisation is used during and after the training. This helps in comparing and selecting the correct model in the given domain and also teaching the deep learning concepts to non-expert users.

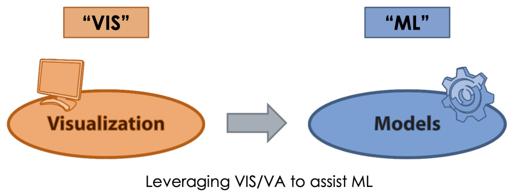

Figure 2: Leveraging Vis 4 ML

4. Main Part

A classic linear projection method, the Grand Tour visualizes high-dimensional data in two dimensions and projects every possible view of the dataset by streamlining animations over multiple iterations. Arbitrarily smooth changing rotation of the dataset is produced by the Grand Tour, and the data is then projected in two-dimensions linearly. Beginning with a random velocity, the data points are rotated smoothly around the origin in high dimensional space, and then projected on 2D for visualization. To show how grand tour function to reduce high dimension to 2D, we are showing rotating 3D cube whose facets are all visible in 2D.

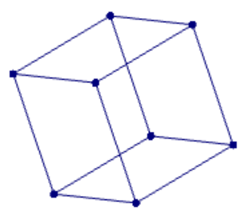

Figure 3: 3D cube whose facets are all visible in 2D[1]

The Grand Tour of the Softmax Layer: Since the axes of the softmax layer have strong semantics, it is easier to understand this layer. The network's certainty about predicting that the given input belongs to the i-th class corresponds to the i-th axis. The Grand

Tour allows us to understand the relative difficulty of classifying different datasets and qualitatively assess the performance of our model. For instance, we can clearly see that for the MNIST dataset (contains grayscale images of 10 handwritten digits), data points are classified most confidently, where the digits are close to one of the ten corners of the softmax space. For Fashion-MNIST (contains grayscale images of 10 types of fashion items), the segregation is not as clean, and more points appear within the bulk of space.

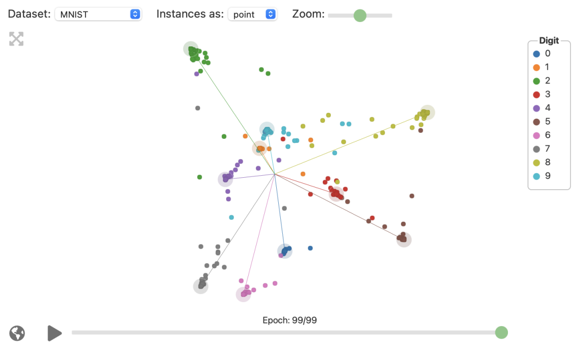

Figure 4: The Grand Tour of softmax layer in the last (99th) epoch, with MNIST dataset[1]

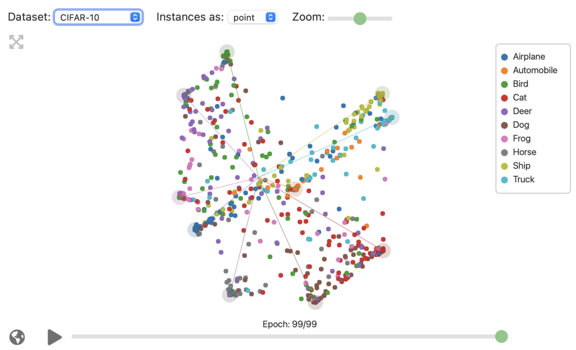

Figure 5: The Grand Tour of softmax layer in the last (99th) epoch, with CIFAR-10 dataset[1]

The Grand Tour of Training Dynamics: The Grand Tour also gives a qualitative assessment of over-fitting by comparing the visualizations of training and testing data. For example, in the MNIST dataset in epoch 99, the testing image is consistent with the training set. Conversely, in CIFAR-10 the training and testing set images are inconsistent. There is a clear contrast in the two distributions. This implies that the model overfits the training data of CIFAR-10 and does not generalize the testing data well.

The Grand Tour of Layer Dynamics: Clearly defining parts of the input (or process) we seek to capture, discard and distill from the representation is critical in a successful design of visualization. The Grand Tour can be made invariant to rotations in data. Therefore, it is an attractive choice when scalings and nonlinearities are more important than the rotational factors in linear transformations of neural networks.

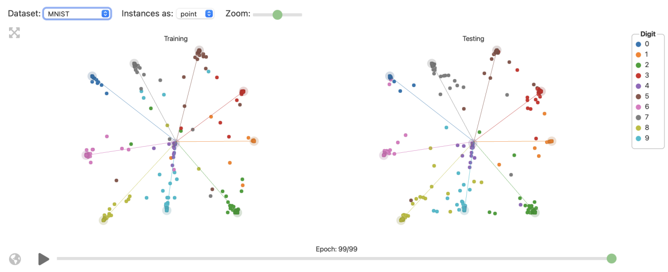

Figure 6: There is less difference between training and testing sets of MNIST dataset[1]

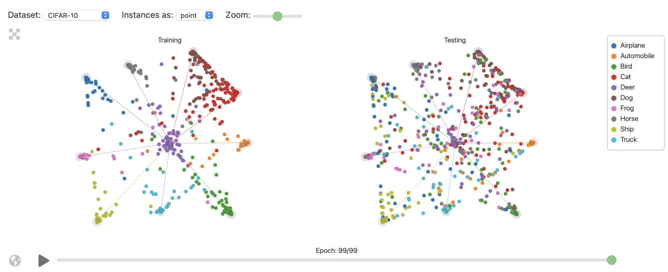

Figure 7: The color of points are more mixed in testing (right) than training (left) set, showing an over-fitting in the training process of CIFAR-10 dataset[1]

In summary, the Grand tour aids in grasping the difficulty of classifying various datasets and assesses the model's performance in the Softmax layer. It also helps in comparison of how the training and testing datasets are fit by the model. These methods show the user what happens in intermediate layers and how changes in data impact the deep learning model. Since it can be made invariant to rotations of data, it also helps in better visualization when rotational factors are not relevant in linear transformation of the network.

5. Discussion

Grand Tour[1] provides a linear model to project n-dimensional features on a two-dimensional screen which is it's strength. This approach has a few shortcomings too which we want to discuss in this section. When topology of data is concerned, like clustering the data or dimensional reduction of downstream models, non-linear models like UMAP or t-SNE are good with fewer dimensions and have more freedom in projection of data. This is one of the strengths of non-linear methods. Grand Tour uses a single animated view of the model to contrast training epochs and dimensionality reduction methods. This decision is made as no such views are available which are highly mentioned in literature or conferences so this can create a bias for given visualisation. Nonetheless, the use of current animation and vis provide a natural way of direct manipulation and continuum of rotations. With CNN, the Grand Tour can run into scalability issues with large sparse matrices present in layers. For simplification of approach, brute force is used to compute the alignment of convolution layers by directly writing matrix representation. To mitigate that, SVD of multi-channel 2D convolutions can be calculated efficiently and can be used for alignment. To improve computation and see various activation layers, it is becoming more common to have non-sequential links in neural networks such as highway branches or dedicated branches for different tasks. Sometimes some neural units are also turned off in a few computations to get better results through fewer units in the model. These non-determinant improvements are not incorporated in Grand Tour and additional research is needed to add such features in Grand Tour for better visualisation.

6. Opinion

After reading, our opinion is that there are still some important cases that are missing in this tool. In this paper, only classification problems have been studied and no regression use cases have been examined. It is also not mentioned how this visualization technique can be extended to regression models. Also important features in dataset are not shown and their impact on result. We suggest, adding this visualisation can help in identifying important features and generate more derived features around it. In neural network, for each layer weights are calculated and applied to input data. How these weights get updated and their change process in each epoch should be shown in visualisation. At last, we want to state that it can conceal internal features of the network. To view them, we need slices through the high dimensional space which can be helpful to understand network better.

7. Conclusion

While the t-SNE and UMAP techniques are quite powerful, they often fail to offer the needed data-visual correspondences that a linear method like Grand Tour can offer. When direct manipulation from the user is accessible and preferable, the Grand Tour method is quite useful. Methods that highlight the best of both worlds can be designed such that non-linear dimensionality technique is used to visualize the intermediate activation layers and the concluding projection can be visualized by the Grand tour. Grand Tour method helps researchers and practitioners in visual analytics of deep learning to understand the model performance. This shows that visual analytics as an integral component in addressing pressing issues in modern AI, helping to discover and communicate insight, discerning model bias and understanding models.

References

[1] Daniel Asimov. “The Grand Tour: A Tool for Viewing Multidimensional Data”. In: *SIAM Journal on Scientific and Statistical Computing* 6.1 (1985), pp. 128–143. DOI: 10.1137/0906011. URL: https://doi.org/10.1137/0906011 (cit. on pp. 1–3).

[2] Paulo E. Rauber, Alexandre X. Falcão, and Alexandru C. Telea. “Visualizing Time-Dependent Data Using Dynamic t-SNE”. In: (2016). Ed. by Enrico Bertini, Niklas Elmqvist, and Thomas Wischgoll. ISSN: -. DOI: 10.2312/eurovisshort.20161164 (cit. on p. 2).

[3] Dominik Sacha et al. “VIS4ML: An Ontology for Visual Analytics Assisted Machine Learning”. In: *IEEE Transactions on Visualization and Computer Graphics* 25.1 (2019), pp. 385–395. DOI: 10.1109/TVCG.2018.2864838 (cit. on p. 2).

[4] “Uniform manifold approximation and projection for dimension reduction. UMAP”. In: *UMAP* (). URL: https://umap-learn.readthedocs.io/en/latest/ (cit. on p. 2).

[5] Wikipedia contributors. “T-distributed stochastic neighbor embedding — Wikipedia, The Free Encyclopedia”. In: (2021). [Online; accessed 11-December-2021]. URL: https://en.wikipedia.org/w/index.php?title=T-distributed_stochastic_neighbor_embedding&oldid=1040619466 (cit. on p. 2).