Deep Learning Report

End Sem Major Exam

Research Paper Implementation

Interpretable part-whole hierarchies and conceptual-semantic relationships in neural networks (CVPR-2022)

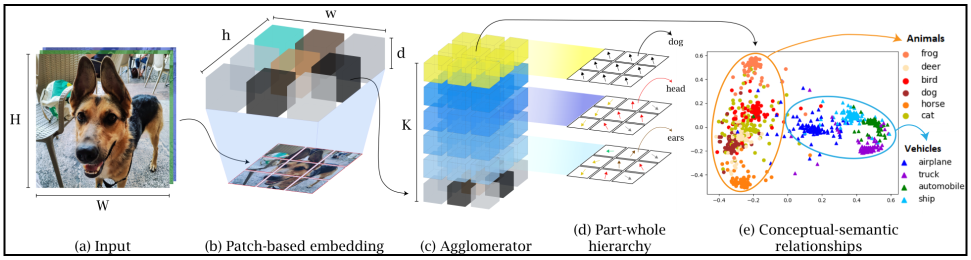
Authors of University of Trento - Department of Information Engineering and Computer Science - DISIVia Sommarive, 9, 38123 Povo, Trento TN:

- Nicola Garau
- Niccol´o Bisagno
- Zeno Sambugaro
- Nicola Conci

Ayush Abrol (B20AI052)

Overview

- The paper discusses the interpretability and understanding of the network's response to a given input limitations of current neural network topologies.
- The paper introduces ***Agglomerator***, a framework that organizes input distribution to meet the conceptual-semantic hierarchical structure across classes and may offer a representation of part-whole hierarchies using visual cues. The paper compares this strategy to other cutting-edge ones by evaluating it on widely used datasets and demonstrating that it produces models that are easier to understand.



(a) Input (b) Patch-based embedding (c) Agglomerator (d) Part-whole hierarchy (e) Conceptual-semantic relationships

[Paper Link]

Datasets

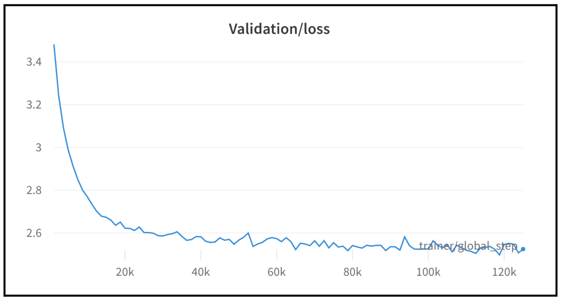
- Datasets such as SmallNORB, MNIST, FashionMNIST, CIFAR-10 and CIFAR-100 have been used in the paper.
- I have implemented the paper and reproduced the results on the CIFAR-10 dataset using the code given in the official github repository.

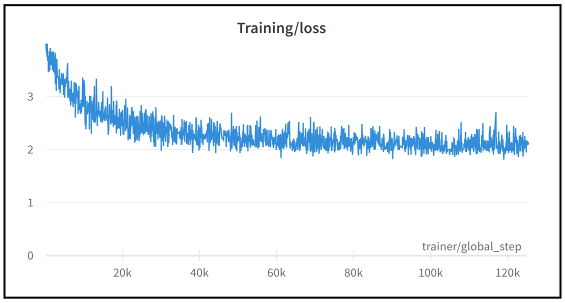
Approach

- Islands of agreement are produced during the propagation phase of the Agglomerator architecture. During this phase, neighbor levels on the same layer are encouraged to reach a consensus by forming islands of agreement. The islands of agreement represent the part-whole hierarchies at different levels. In other words, they are clusters of neurons that agree on a particular representation of an object or concept. These islands help in creating interpretable part-whole hierarchies and hierarchical organization of the feature space contained in data, which is one of the main goals of this architecture.

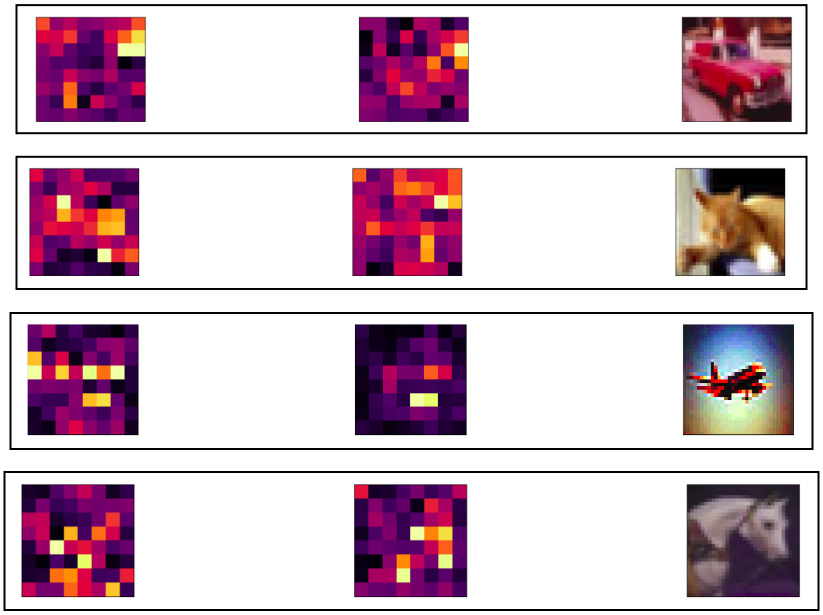
PART 1: Implementation

- We loaded the pre-trained model given in the official repository and fine-tuned it for the CIFAR10 dataset. Logs of training are logged in the wandb server.





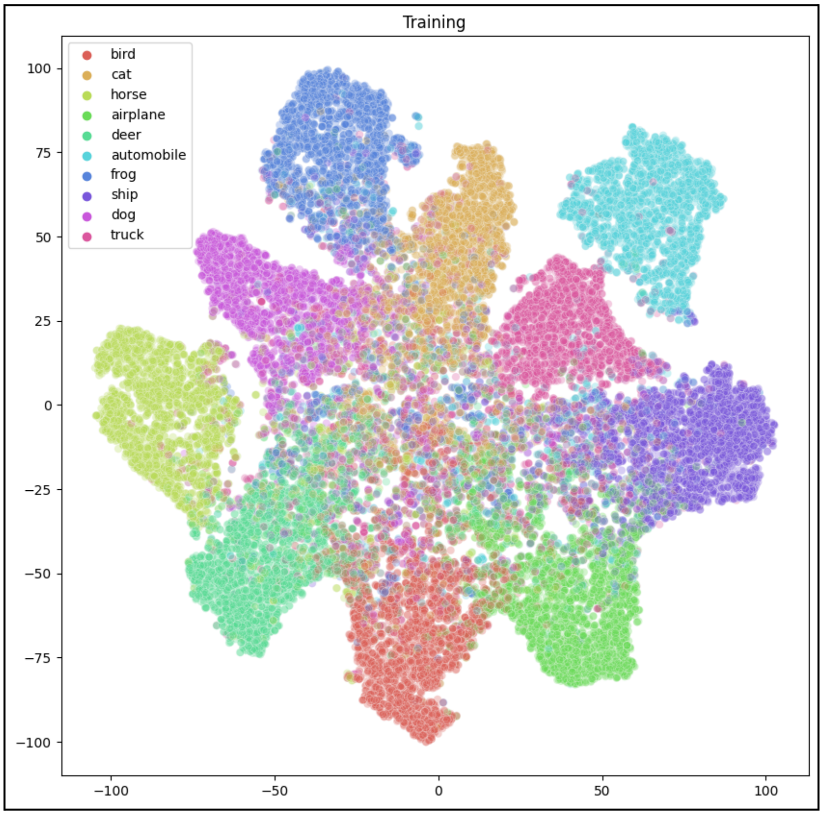
- Generated the plot islands when testing the model.



These are the plot islands obtained for levels(k) set to 2 for CIFAR10 dataset

We can see the areas where the object is present in the image is highlighted a little brighter in the plot (thus, the name island)

- Latent space organization as the representation of conceptual-semantic relationship in data. (t-SNE Plot to visualize the clustering)

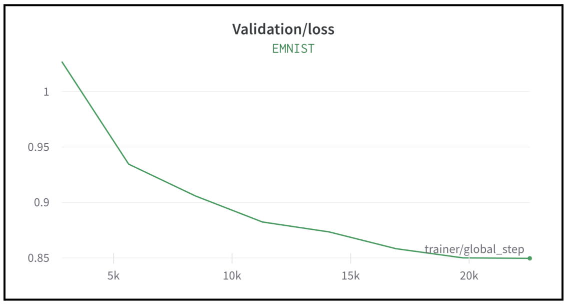


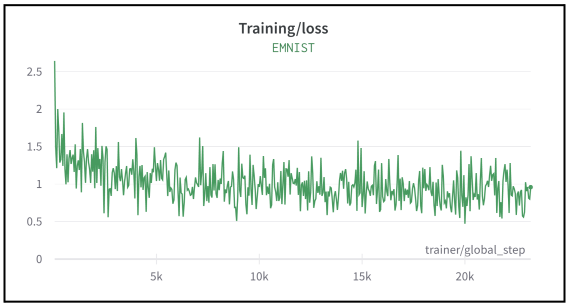
Through this t-SNE plot for CIFAR10, observation can be made that the rightmost 4 colors are the classes (ship, airplane, automobile and truck) and the other classes are ((cat, frog, dog, horse, deer and bird)

PART 2: Application/ Implementation on a new dataset (EMNIST)

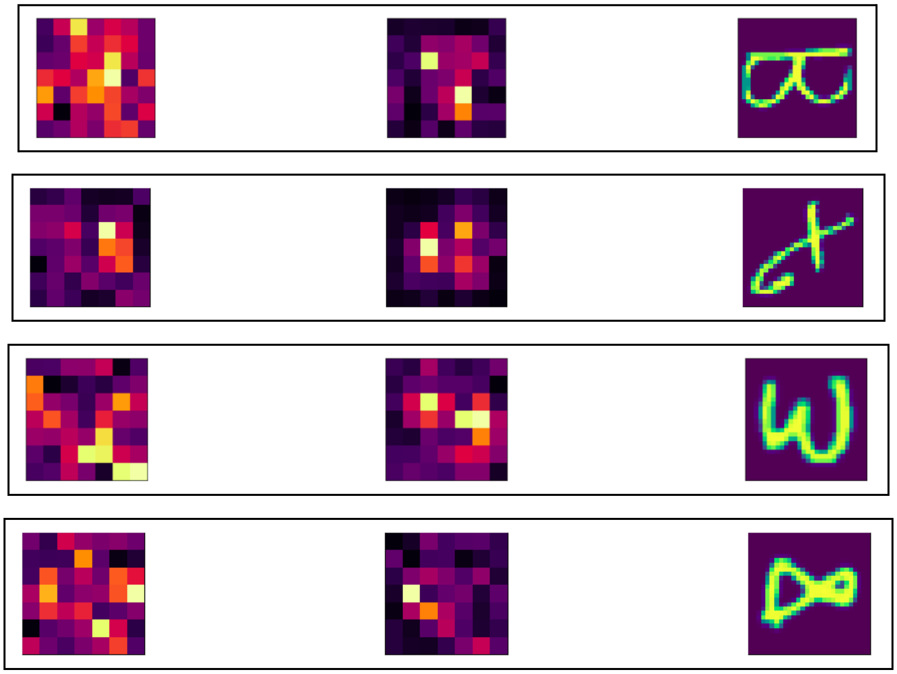
- For the application part ,we will be using the EMNIST dataset for the application part and we will observe that the model is able to cluster the letters of both (A-z) and (a-z) together and digits (0-9) together in the t-SNE plot and the interpretable part-whole hierarchies and conceptual-semantic relationships can be seen.

- Loss Plots for training the model on EMNIST



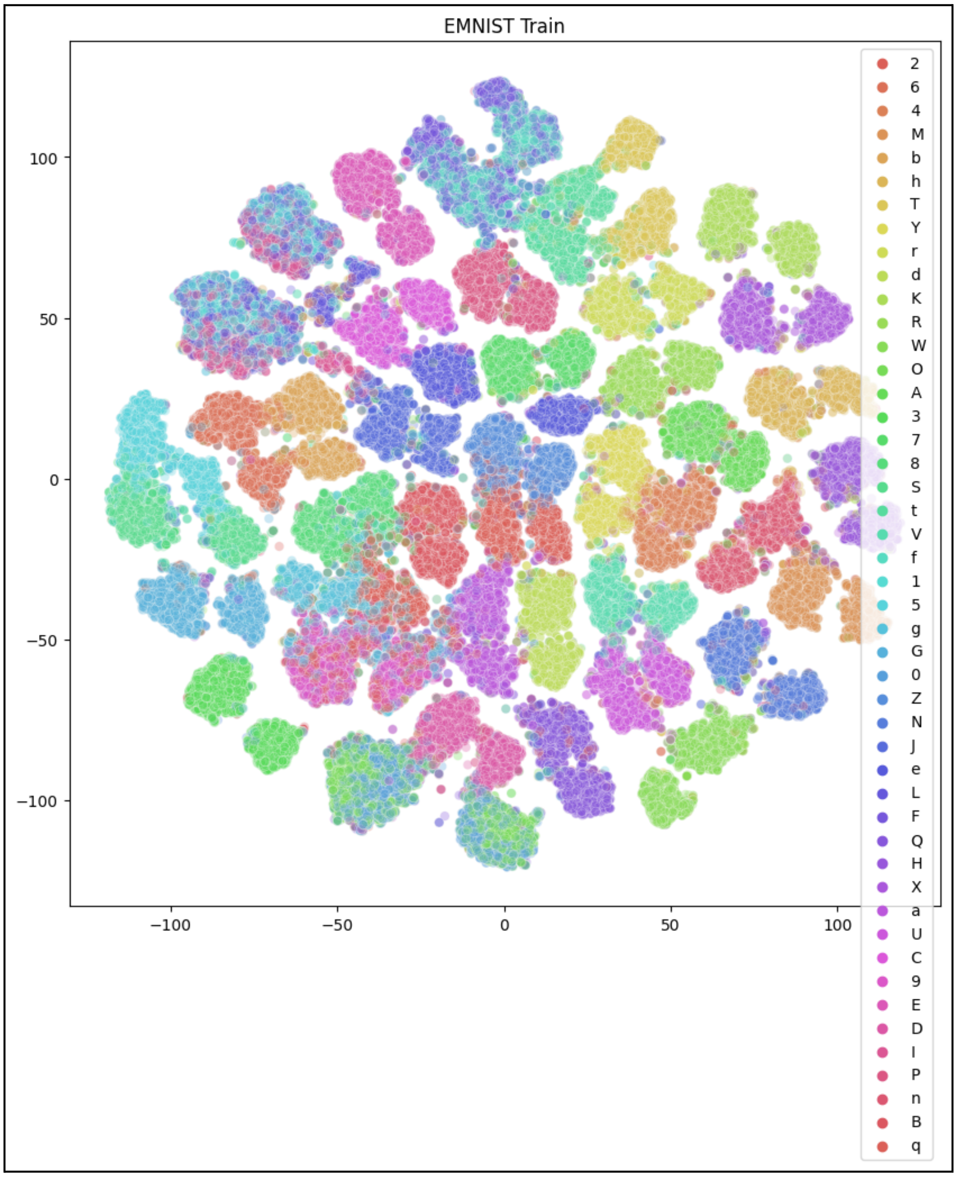


- Plot islands created for EMNIST model



These are the plot islands obtained for levels(k) set to 2 for EMNIST dataset

- Latent space organization as the representation of conceptual-semantic relationship in data. (t-SNE Plot to visualize the clustering)



We can observe here that the classes are not overlapping with each other and the model is able to cluster the classes properly. We can see some clusters of classes like (A-Z) and (a-z) are overlapping with each other and digits (0-9) are overlapping with each other. But overall, the model is able to cluster the classes properly and the interpretable part-whole hierarchies and conceptual-semantic relationships can be seen as digits are on one side and letters are on the other side of the t-SNE plot.

PART 3: Asking ChatGPT to perform the same task

Answer by ChatGPT for Interpretability of part-whole hierarchies on EMNIST

- The Agglomerator framework aims to organize the input distribution to align with the conceptual-semantic hierarchical structure of the dataset. In the case of the EMNIST balanced dataset, this could mean that the Agglomerator model will create islands of agreement that represent different levels of the part-whole hierarchies of handwritten characters. For example, the islands of agreement could represent features such as strokes, loops, or angles that are common to certain subsets of characters.
- The t-SNE plots will provide a visual representation of the feature space created by the Agglomerator model. The plots will allow researchers and practitioners to visualize the islands of agreement and identify any areas of overlap or separation between different clusters of neurons. This can help in understanding the patterns in the data and the relationships between different characters.
- Since the EMNIST balanced dataset is a well-studied benchmark dataset, we can expect that the Agglomerator framework will achieve competitive performance on the dataset. The t-SNE plots will provide a visual representation of the interpretability of the model and can be used to compare it with other cutting-edge approaches.
- In summary, we can predict that the Agglomerator framework applied to the EMNIST balanced dataset will create interpretable neural network models that better capture the hierarchical structure of the dataset. The t-SNE plots will provide a visual representation of the feature space created by the model and can help in understanding the patterns in the data and the relationships between different characters.

Is the answer given by ChatGPT correct?

- The answer given is based on general predictions and assumptions based on the characteristics of the Agglomerator framework and the EMNIST balanced dataset. However, when implementing it I faced some problems like class imbalance in EMNIST, complexity of relationships among classes like 8 and ‘B’; 0 and ‘O’ etc.