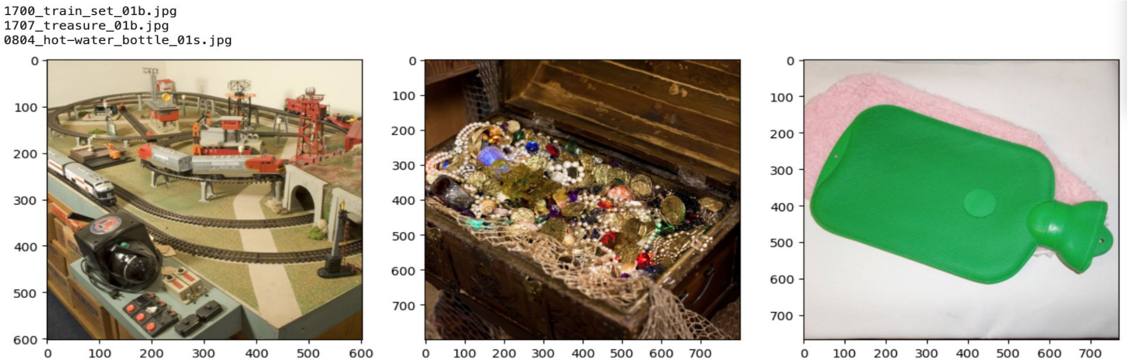
# Mimicking Human Image Similarity Judgment Using Siamese Neural Network

By: Bonnie, You, Bereket, Melika

neuromatch academy

# Introduction

- Humans have the ability to make similarity judgments about objects. These judgments are formed through hierarchical categorization of core criterias of the objects such as color, shape, purpose, etc.




- Our idea was inspired by a NIH paper that demonstrated how human object similarity judgment can be modeled using an embedding vector obtained through an optimization technique.
- In our implementation we aim to achieve the same objective using Siamese neural network to directly extract embedding vectors from the input images.
- Our project trains a neural network model that can capture the core dimensions used by humans to judge similarity between objects and apply it to recognizing the "odd one out" in a set of triplet natural images.

# Objective

**Research Question:** Can we mimic human visual similarity judgment using a neural network that can learn representations from images used in odd-one-out tasks?

## Evolution of our hypothesis:

**Initial Hypothesis:** The neural network can be trained to learn representation that can be model or capture the core criteria used by humans in similarity judgement.

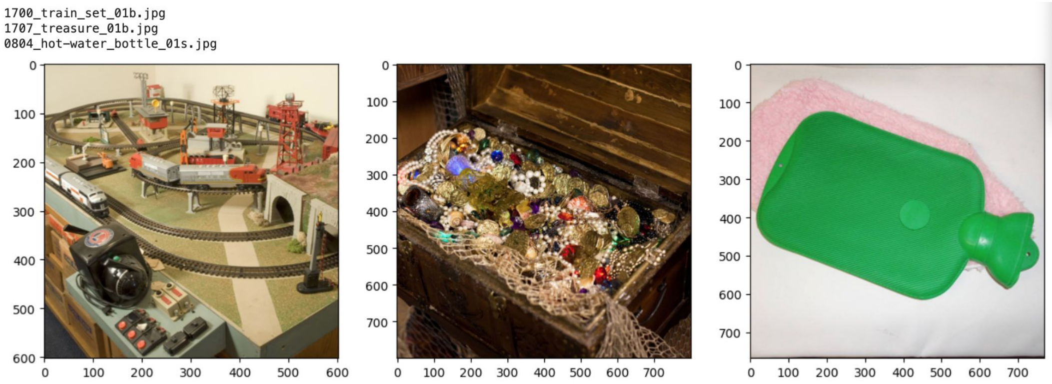
After reading the paper **Mahner, F.P et.al 2024 *rXiv preprint arXiv:2406.19087*** we learned that

- The embedding vector learned from the neural network can not be interpreted directly as the dimension of the embedding vectors are not disentangled.
- The embedding vector learned by neural network is focused on the visual representation of the images, while human rely on semantic representation of images.

**Final Hypothesis:** We hypothesize that the neural network will perform to a level close to humans in identifying odd-out images and the learned representations will not be affected by noise and geometric variations introduced to the input images.

# Dataset

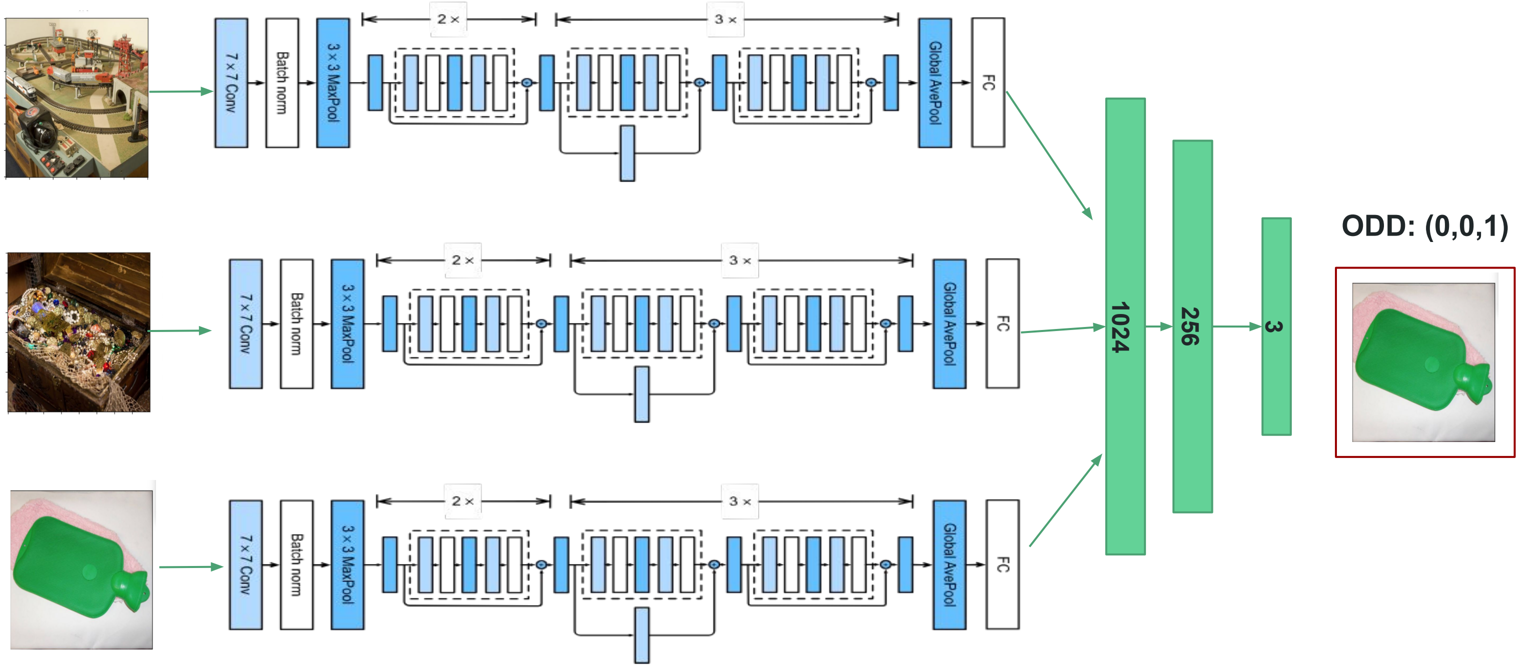
- We chose to use the THINGS Similarity dataset from the THINGS database, that has training, testing, and validation images of natural objects.
- Our dataset consists of the *image dataset* and *triplet dataset*
  - Triplet Dataset – txt files with sets of triplets of image indices; the left-most index represents the labelled odd-one-out image
  - Image Dataset – 1854 reference images sorted in alphabetical order.
- The training dataset was truncated to less than 1% of the original data size due to the limited  computational resource available




# Model

**Pre-trained ResNet18 for feature extraction (Fine-tuned)**

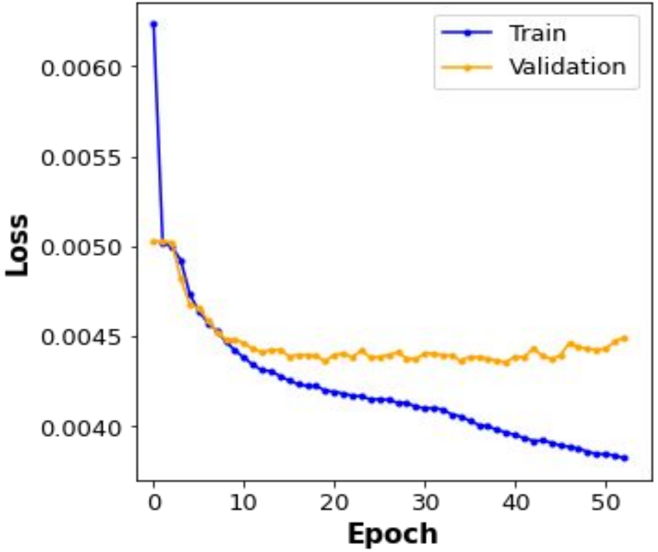
**Fully connected layers for classification (ReLU + 50%Dropout)**

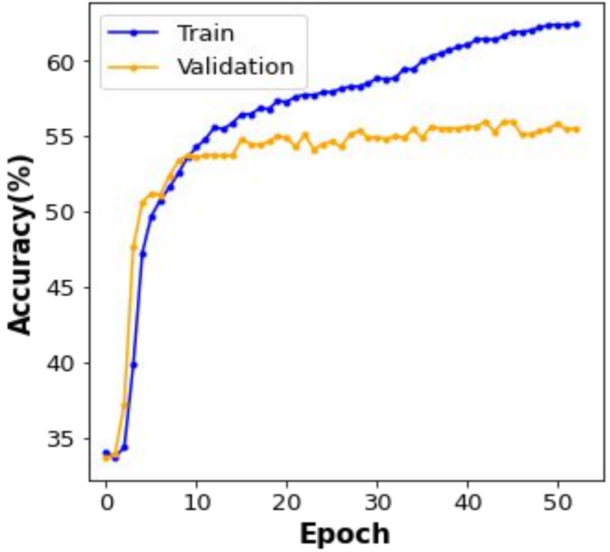


# Model Training

**Loss function:** Binary cross entropy + L2 regularization
**Learning rate**: 0.0001
**Early stopping**

**Dataset size:**
Training-50000 Validation-10000 Test-10000

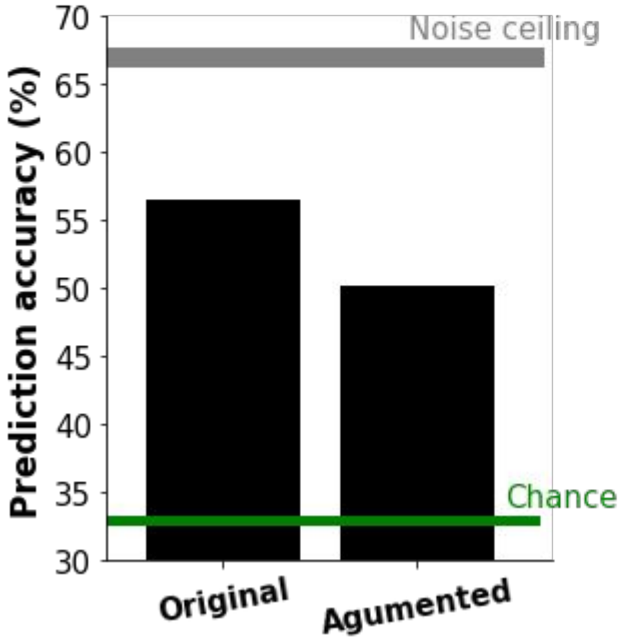


# Model Evaluation

**Original dataset**

| | |
|---|---|
| Accuracy | 56.48% |
| Precision | 56.60% |
| Recall | 56.53% |
| F1 | 56.10% |




**Augmented dataset**

| | |
|---|---|
| Accuracy | 50.24% |
| Precision | 50.40% |
| Recall | 50.37% |
| F1 | 49.97% |

# Limitations

- The model was trained on a small subset of the triplet dataset due to the limited available GPU .
- The embedding vectors are of high dimensions and lack interpretability.
- Embedding vectors obtained from the neural network are more biased towards visual representation instead of semantic representation.

# Future Direction

- Improving the performance of the model by training with more triplet datasets to improve the accuracy of the model (our current training only takes a very small subset of the provided training triplets).
- Convert the representations into interpretable embedding vectors.
- Incorporating semantic representation into the features learned by the neural network.