Project report

## Introduction

First of all, we tried to farmiliarize ourselves with what hard samples are.

In machine learning, "hard samples" and "rare samples" refer to specific types of data instances that can pose challenges for models during training or testing. Let's break down each term:

1. **Hard Samples**:

- **Definition**: Hard samples are data instances that are particularly challenging for a machine learning model to correctly predict or classify.

- **Characteristics**:

- They may be on the boundary between different classes, making the decision boundary ambiguous.

- They could have complex patterns or relationships that are difficult for the model to capture.

- Hard samples might be outliers or instances with unusual characteristics.

2. **Rare Samples**:

- **Definition**: Rare samples are instances that occur infrequently or are underrepresented in the dataset.

- **Characteristics**:

- They belong to a minority class in a classification problem, making them rare compared to the majority class.

- Rare samples may have limited representation in the training data, leading to challenges in learning their characteristics.

- In some cases, rare samples might be outliers or examples that deviate significantly from the majority of instances.

**Significance**:

- **Challenges**: Hard and rare samples can pose challenges during the training of machine learning models. Models may struggle to generalize well to these samples, leading to lower performance on them.

- **Importance**: Understanding and correctly handling hard and rare samples are crucial for improving the robustness and generalization of machine learning models, especially in tasks where correctly identifying such samples is essential (e.g., fraud detection, medical diagnosis of rare diseases).

**Handling Strategies**:

1. **Data Augmentation**: Generate additional training examples by applying transformations to existing samples.

2. **Resampling Techniques**: Over-sample rare classes or under-sample majority classes to balance class distribution.

3. **Ensemble Methods**: Combine predictions from multiple models to improve overall performance, especially on hard samples.

4. **Focus on Model's Attention**: Design models that can focus more on hard or rare samples during training.

Identifying and addressing the challenges posed by hard and rare samples contribute to the overall improvement of a machine learning model's performance and reliability.

Some methods for estimating the hardness or rareness of a sample include: Confidence score, Prediction entropy, Margin sampling, Uncertainty quantification, etc

## Related work

For our project, we had recommended papers which contain contain related work. We considered the foremost 2.

1. Characterizing Datapoints via Second-Split Forgetting (SSFT)
2. Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics.

**Charaterizing Datapoints via Second-Split Forgetting (SSFT)**

In this article, the authors propose a new forgetting metric - a complementary metric that tracks the epoch after which an original training example is forgotten during fine tuning on a randomly held out partition of the data.

The metric also aims to classify what type of hardness the example might have - mislabeled examples or rare examples. The advantage here is that mislabeled examples can be removed from the training set.

One major result of this paper is showing that mislabeled examples are learnt quite slowly but forgotten quickly. Also, it further distinguishes complex samples from rare samples by the fact that the former are never forgotten while the latter are.

**Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics**

The idea of this metric leverages on the behaviour of the model on individual instances during training by measuring the confidence of the model in the true class and how it changes across epochs in one training. Here, confidence is defined as the mean of the model probability of the true label across epochs. Variability is the standard deviation of the same random variable (model probability) across epochs.

The result of the paper is that, hard samples had low variability and low confidence, while ambiguous samples had high variability.

Other recommended papers which we looked at are

1. **Prioritized Training on Points that are Learnable, Worth Learning, and not yet Learnt**: It talks about RHO-LOSS which is a technique designed to accelerate training on large-scale datasets. It addresses issues with existing data selection methods by choosing points that significantly reduce the model's generalization loss. Unlike other methods, RHO-LOSS avoids "hard" (high loss) points that are often noisy or less relevant to the task. It outperforms optimization-based and curriculum learning approaches, selecting points that are both learnable and worth learning. RHO-LOSS achieves faster training and higher accuracy across various datasets and model architectures, such as MLPs, CNNs, and BERT. In a case study on the Clothing-1M dataset, RHO-LOSS trains 18 times faster and achieves a 2% higher final accuracy compared to uniform data shuffling.
2. **An Empirical Study of Example Forgetting during Deep Neural Network Learning**: The study investigates neural network learning dynamics during single classification tasks, exploring the occurrence of forgetting events where examples shift from correct to incorrect classification. Key findings include high-frequency forgetting for specific examples, architecture-independent forgettable examples, and the possibility of omitting a significant fraction of examples from the training set while maintaining strong generalization performance. The research sheds light on forgetting phenomena in the absence of clear distributional shifts in the data.

## Our work

We conducted experiments with metrics based on latent space dynamics during model training, compared the results with label-based methods described above, and examined the hard samples these 3 approaches found.

### Metrics

Consider - latent space vector for i-th example at e-th training epoch.

**Variance shift:** average distance, latent space moved from its average coordinates during the training epochs

**Mean shift:** average distance between latent spaces during epochs

**Velocity: ,** where. The idea is to track the “second derivative”. Hypothesis is: hard examples(as it shown in SSF paper) train longer, so they should not “slow down” as easy examples.

**Distance:** Avg distance between centre of latent space and embedding over epochs

### Uncertainty

Another label based approach([Swayamdipta](https://arxiv.org/search/cs?searchtype=author&query=Swayamdipta,+S) et al, Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics) used model uncertainty of true labels to identify hard examples.

Besides that, we studied 2 approaches to estimate uncertainty, which we used to compare with our metrics.

**Max probability:** - inverse probability of most confident class, according to model. This approach is easy to calculate but requires model calibration.

**Entropy based:** Entropy of the distribution over classes probability for

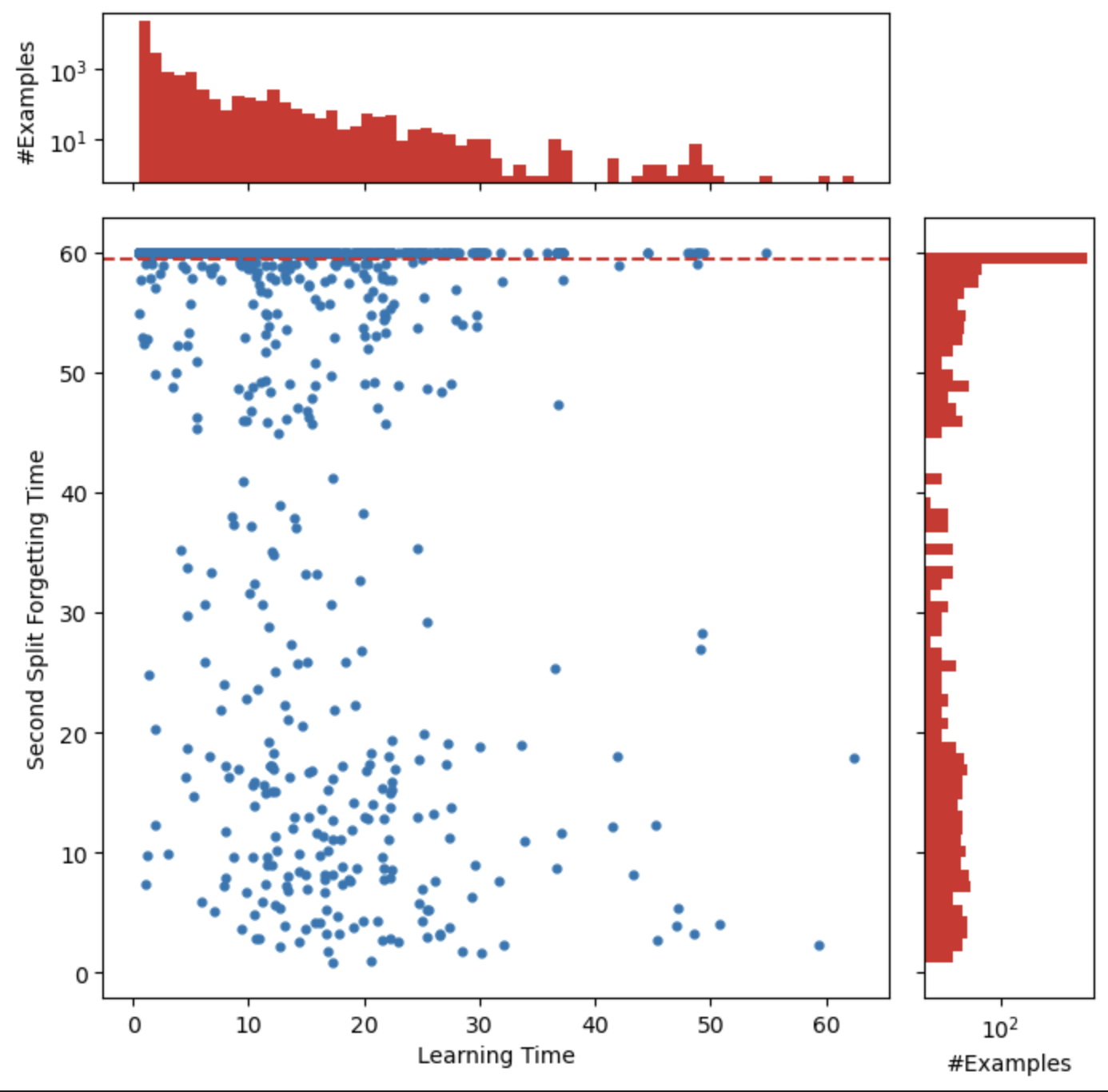
### Experiments design

All experiments were conducted on MNIST dataset. First, we reproduced Second split forgetting([Pratyush Maini](https://arxiv.org/search/cs?searchtype=author&query=Maini,+P) et al) code to identify hard samples in the dataset. By design of experiment noise and perturbation were introduced to the train dataset to generate such samples. We used exactly that dataset to evaluate our approach. Then we trained the same model during 30 epochs and saved latent spaces and output probabilities for train and test split. After label and latent-space based metrics were calculated for each example in the dataset and visualised latent space embeddings compressed by PCA. Finally we compared our methods results with other methods numerically and visually and found hard samples.

### Experiments results

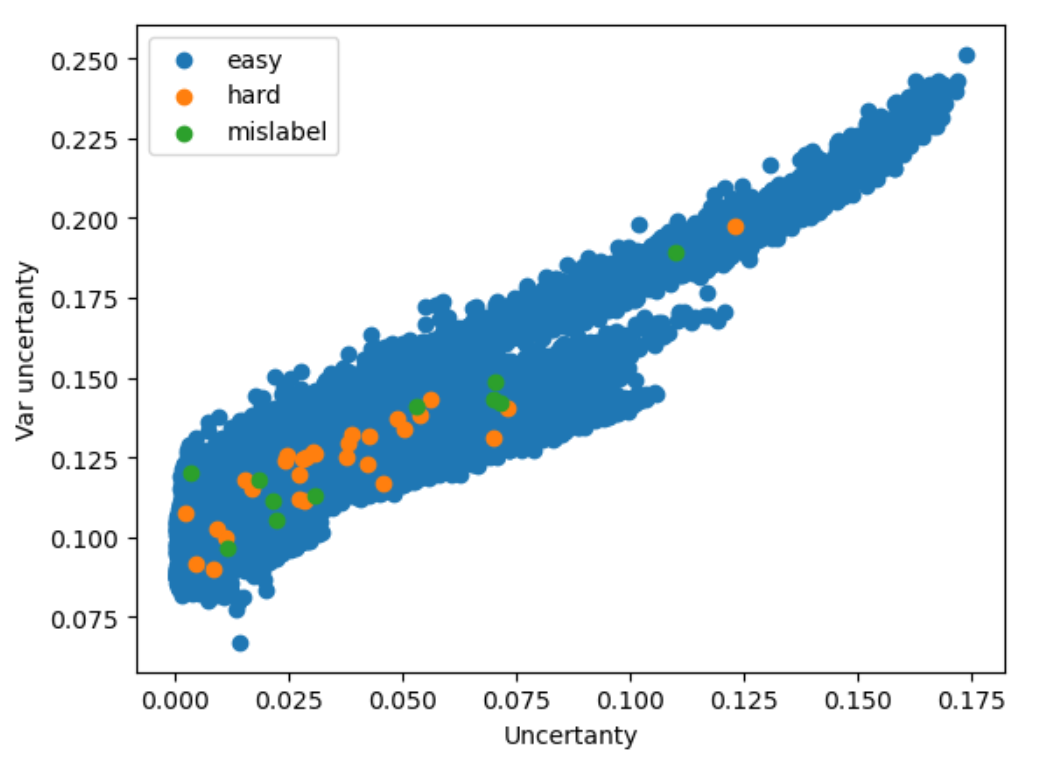
**Second split forgetting**

We obtained similar results as authors provide in papers.



Number of hard samples is very small, so it is hard to study it.

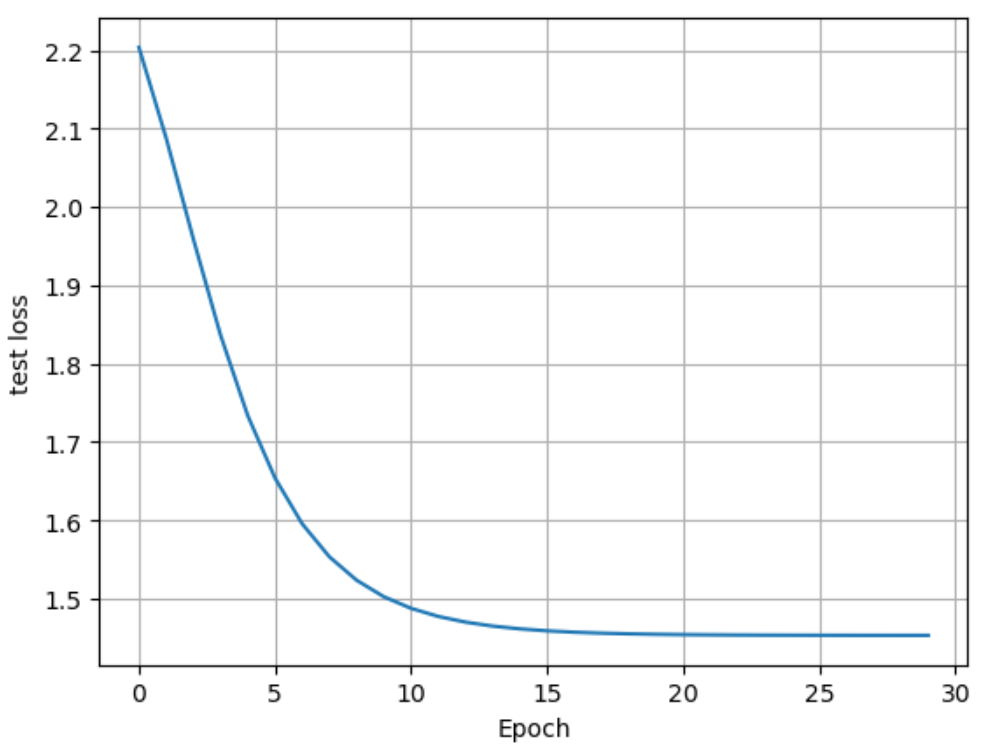
**Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics**

****

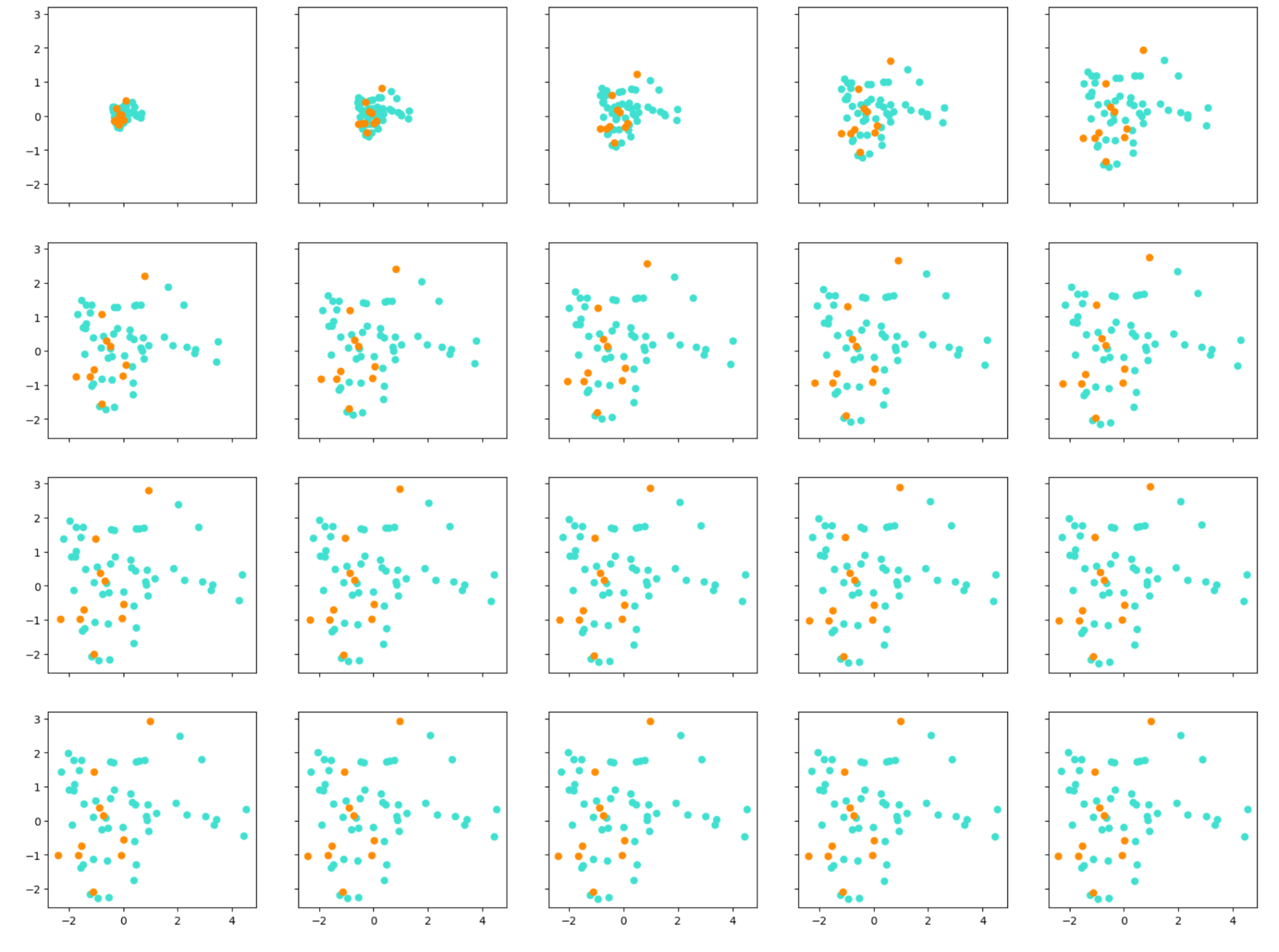
In our experiment distribution of uncertainty and variability metrics become degenerate, but hard samples are also located at left-bottom.

**Model training**

We trained LeNet model for 30 epochs with SGD.

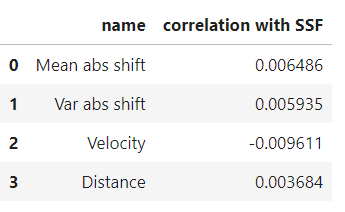


Test loss was the following. Model achieved 99.99% accuracy. During the training latent space embeddings were saved. Here visualised PCA of their dynamic:

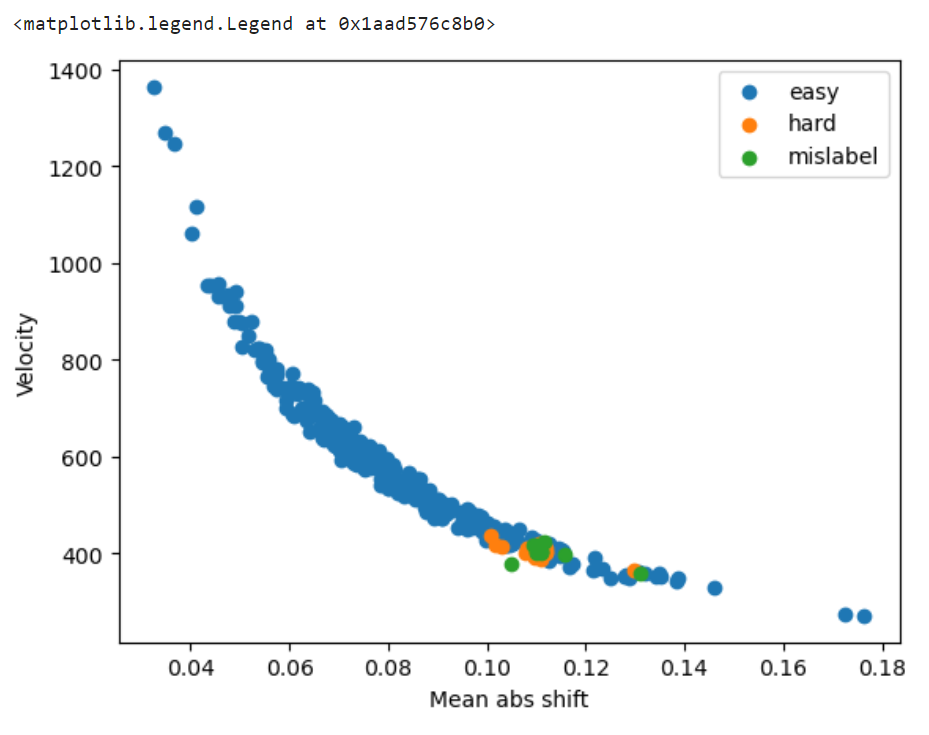


**Second split forgetting comparison**

Correlation was calculated between corresponding metric of latent space dynamic and from the paper’s code. The results don’t seem to be promising but correlation is not the good way to compare these 2 approaches.

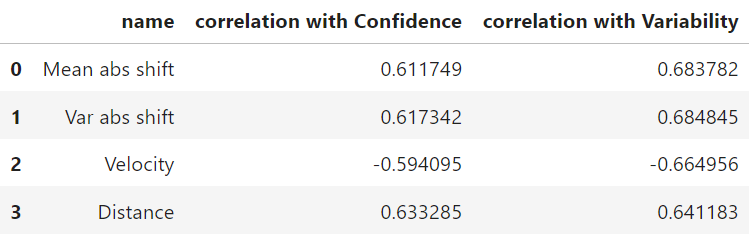
****

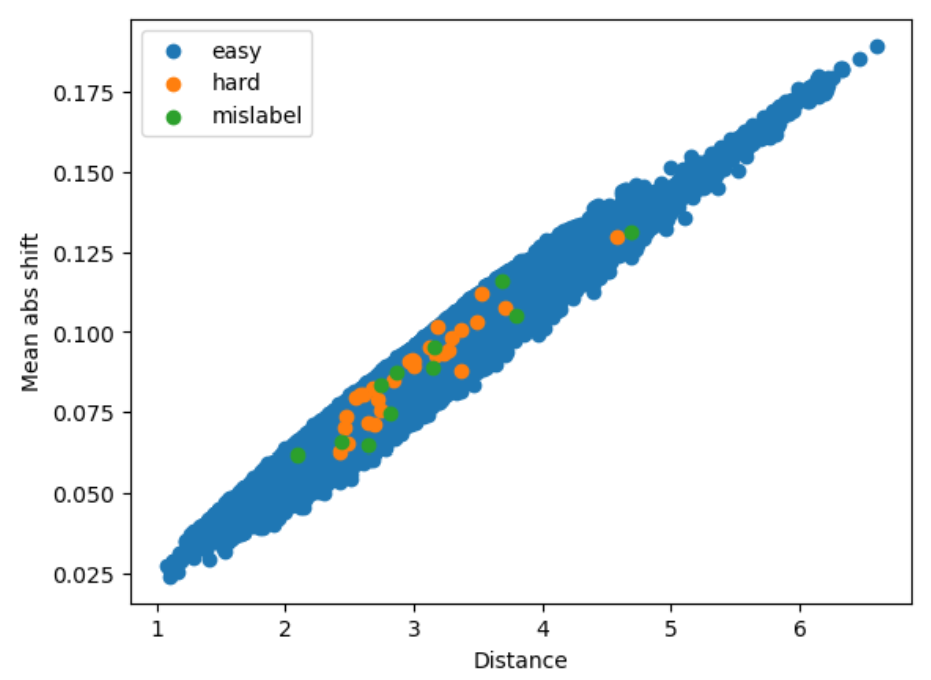
The plot below demonstrates how complex(yellow) and mislabelled(green) points, selected from the right-top and right-bottom from the learning-forgetting plane are concentrated in one region on Velocity/Mean abs shift plane.

****

**Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics comparison**

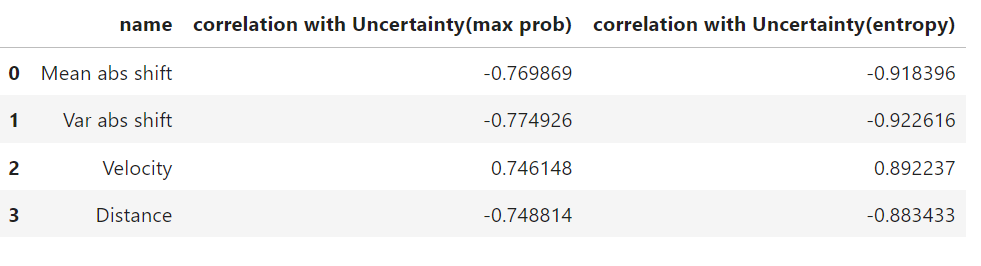
The correlation here is significant. The region of hard samples is also in left-bottom, but biased.

****

****

**Uncertainty correspondance with proposed metrics**

Results are the following:

****

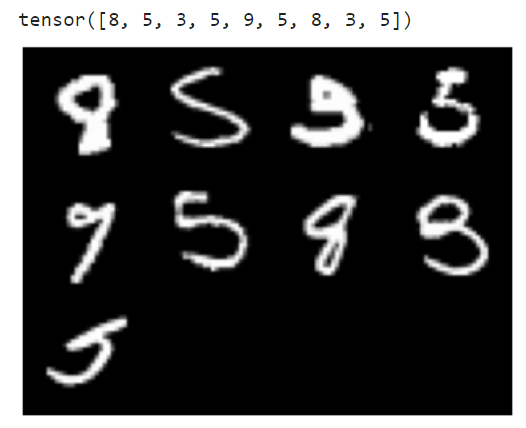
Visually, certain and uncertain examples looks similar in different metrics:



**Unsupervised latent space complex examples**

To estimate how good proposed metrics perform in absence of labels we computed these metrics on test split and studied regions where hard samples should be, visually.

Complex examples according ssf:



Complex examples according to combination of MeanAbsShift and Velocity:

