SUMMER
INTERNSHIP 2023

REPORT WEEK 2

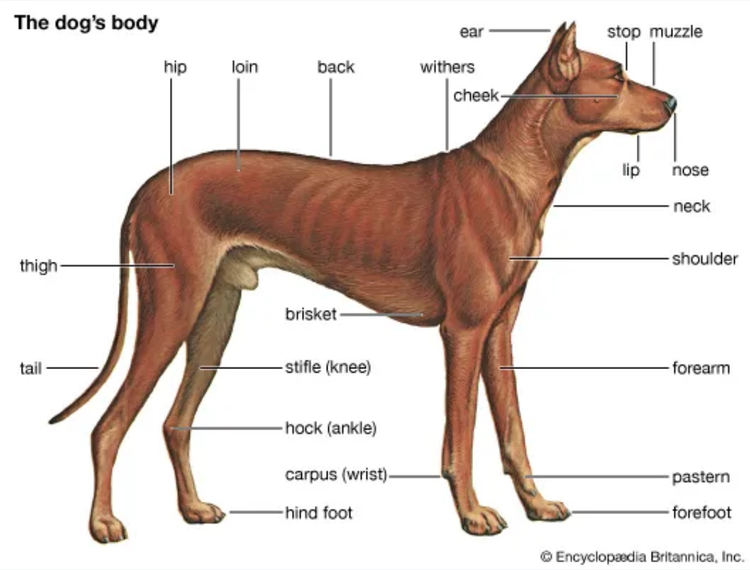
LE VIET HUNG - AI INTERN

CONTENT

[1] Insufficient Quantity of Training Data

Insufficient quantity of training data is defined as the situation in machine learning that training dataset is too small and insufficient to effectively capture underlying patterns of target problem, so it can lead to several problems, such as overfitting, poor generalization, and limited model performance.

Example: In dog and cat classification , if we get the lack of training data, the model will not learn underlying feature of the dog, such eye, nose, tail,...



[2] Nonrepresentative Training Data

Non-representative training data is defined as the situation in machine learning that the training dataset is not generalized enough , the training data don't cover all cases of real problem that are already occurred.

[2] Nonrepresentative Training Data

Example: Dog and cat Classification

In large dataset of dog-and-cat , we can capture many views of image's dog or cat . In contrast, in smaller dataset , we get just a little of image's style

Small Dog-and-cat dataset

Large Dog-and-cat dataset

[3] Poor-Quality Training Data

Poor-quality training data is a general definition that the training data is inaccurate, incomplete, inconsistent, outdated, or otherwise unreliable for the intended use or analysis

Poor-quality training data can be include some types:

- Mislabeled data
- Biased data
- Outlier data

[3] Poor-Quality Training Data

- **Mislabeled data:** Data points with mislabeled labels or can mislead the model during training and lead to inaccurate predictions.
- **Example: We** We get a dataset for cat and dog classification, but some of the data of cat is mislabeled as dog

Dog

[3] Poor-Quality Training Data

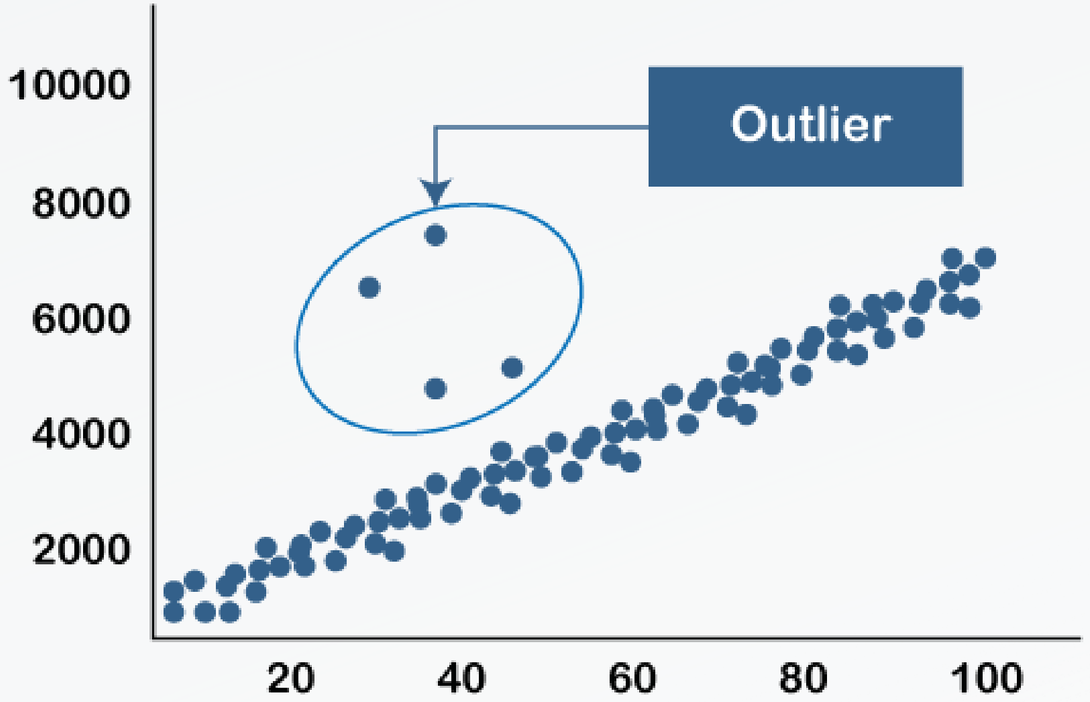
- **Biased data:** Training data that reflects biased views can lead to biased predictions by the model.
- **Example:** Suppose you collect data for predicting the salary. you collected the rich people , but in fact the actual data include both rich and poor people.

collected data
(rich people)

actual data
(both rich and poor people)

[3] Poor-Quality Training Data

- **Outlier data:** Outliers and noisy data can distract the model's understanding of the underlying patterns, making it less effective in making accurate predictions.
- **Example:** We get a dataset to predict something, with some of outliers here



[4] Irrelevant Features

I**rrelevant Feature** is a term used to describe attributes or features of the data that do not significantly contribute to predicting the output or provide any useful information to the machine learning model

[4] Irrelevant Features

Example: In this dataset, the "Size," "Bedrooms," "Location," and "Has Pool" are the features, and "Price" is the target variable we want to predict.

Size (sq. ft.)	Bedrooms	Location	Has Pool	Price (USD)
1500	3	Suburb	No	200,000
2000	4	City Center	Yes	350,000
1800	3	Suburb	No	220,000
1200	2	Rural Area	No	150,000
2400	4	City Center	Yes	400,000

In this small dataset, we observe that when "Has Pool" is "Yes," the price tends to be higher. However, as we analyze more data, we might find that the presence of a pool has little to no impact on the housing price. In this case, the "Has Pool" feature is likely an irrelevant feature because it doesn't provide meaningful information for predicting the housing price.

[5] Overfitting and Underfitting the training data

	Overfitting the training data	Underfitting the training data
Definition	Overfitting occurs when a model learns the training data too well, capturing both the underlying patterns and the noise in the data. The model becomes overly complex, leading to poor performance on new, unseen data.	Underfitting, on the other hand, happens when a model is too simple to capture the underlying patterns in the training data. It fails to learn the relationships and exhibits poor performance on both the training data and unseen data.
Performance	The model performs poorly on new, unseen data, as it has memorized the training data's noise and cannot generalize to different samples.	An underfitting model will have low accuracy or high error on the training data since it fails to learn the underlying patterns effectively.
Cause	Overfitting can occur due to a small training dataset, too many irrelevant features, high model complexity	Underfitting often results from using an overly simplistic model, insufficient training, or inadequate feature representation

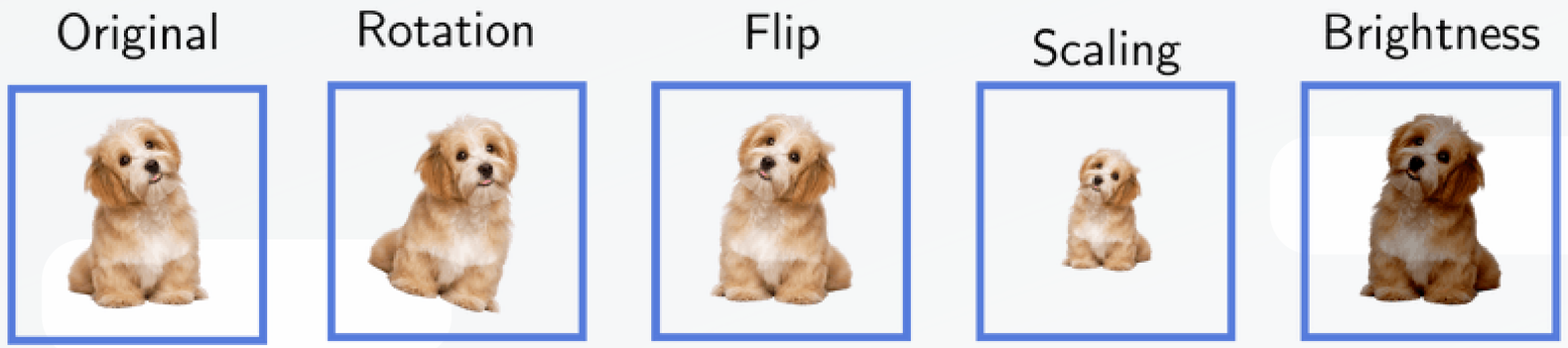
[6] Methods

To deal with these problem, we get some of the strategies below:
1) Data augmentation
2) Improve model complexity

[6] Methods

1) Data augmentation:

You can artificially enrich your dataset through data augmentation techniques. For example, you can apply transformations, rotations, flips, or other modifications to your existing data to create new samples.



[6] Methods

2) Improve model complexity:

You also can improve the complexity of model to effectively learn feature of dataset.
For example:
- In machine learning, for linear regression you can add more parameters to improve the model complexity
- In deep learning model, you can add more layers, use the techniques to reduce the overfitting and underfitting or apply the the-state-of-art model to boost your model.

Thank you!