WE WILL START AT 1:05 PM

ML SYSTEMS DESIGN MEETUP GROUP

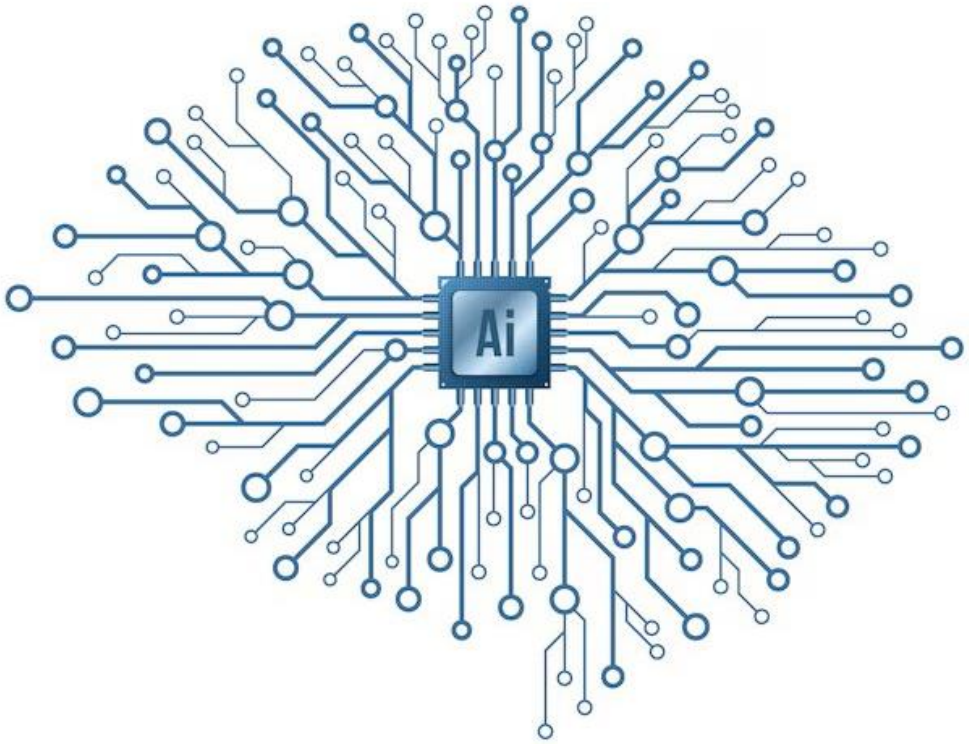
HETAV PANDYA

IMPORTANT: THIS WORKSHOP WILL BE RECORDED

HETAV PANDYA

AGENDA

- Introduction
- Data labelling
- Labelling functions
- Natural labelling
- Weak supervision models
- Semi-supervised learning
- Transfer learning
- Class imbalance problems
- Resampling
- Cost sensitive learning
- Data Augmentation
- Open Q&A



INTRODUCTION

- Welcome to ML Systems Design Meetup Group
- Designing Machine Learning Systems – Chip Huyen
- Free Access – City Library
- Frequency – Biweekly - Monthly
- Questions: Meetup Event Chat

FREE ACCESS

Via Burnaby Public Library

WHAT ARE LABELS?

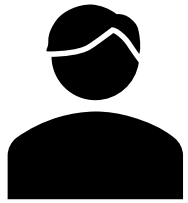
A method to teach your models the patterns in your dataset

Bad

Good

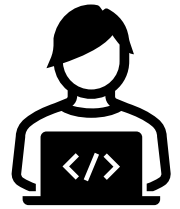
Logs, debug files, system metrics

TYPES OF LABELS



Hand labelling

Natural labelling

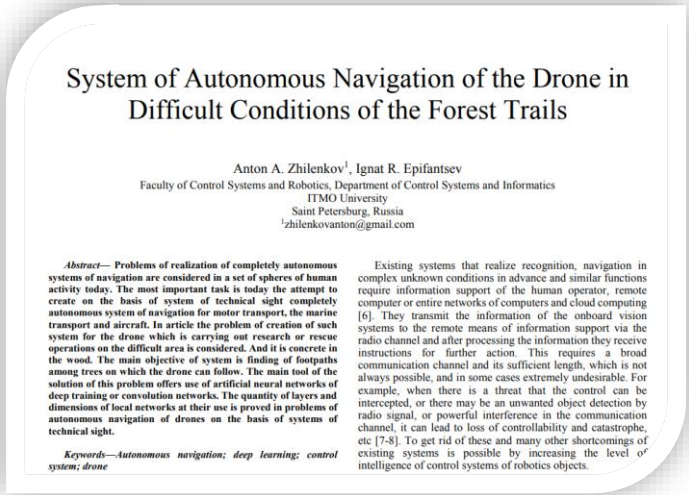


Programmatic labelling

NATURAL LABELLING

Tasks with natural labels are tasks where the model's predictions can be automatically evaluated or partially evaluated by the system.

Fig. 3. Demonstration of the input classification result



System of Autonomous Navigation of the Drone in Difficult Conditions of the Forest Trails

Anton A. Zhilenkov[1], Ignat R. Epifantsev
Faculty of Control Systems and Robotics, Department of Control Systems and Informatics
ITMO University
Saint Petersburg, Russia
[1]zhilenkovanton@gmail.com

***Abstract*— Problems of realization of completely autonomous systems of navigation are considered in a set of spheres of human activity today. The most important task is today the attempt to create on the basis of system of technical sight completely autonomous system of navigation for motor transport, the marine transport and aircraft. In article the problem of creation of such system for the drone which is carrying out research or rescue operations on the difficult area is considered. And it is concrete in the wood. The main objective of system is finding of footpaths among trees on which the drone can follow. The main tool of the solution of this problem offers use of artificial neural networks of deep training or convolution networks. The quantity of layers and dimensions of local networks at their use is proved in problems of autonomous navigation of drones on the basis of systems of technical sight.**

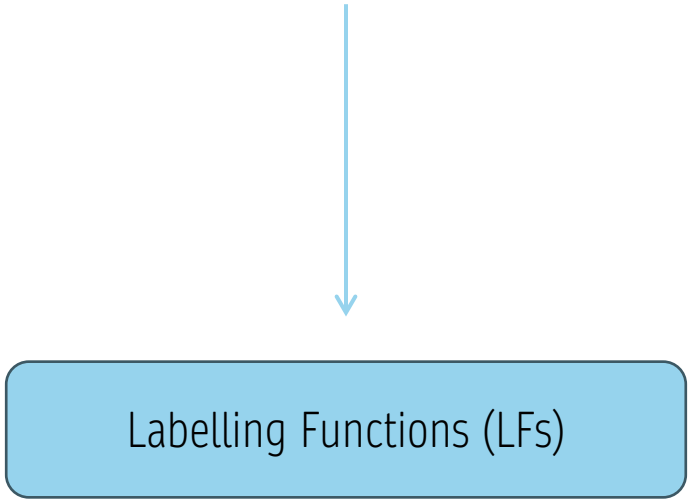
Keywords—Autonomous navigation; deep learning; control system; drone

Existing systems that realize recognition, navigation in complex unknown conditions in advance and similar functions require information support of the human operator, remote computer or entire networks of computers and cloud computing [6]. They transmit the information of the onboard vision systems to the remote means of information support via the radio channel and after processing the information they receive instructions for further action. This requires a broad communication channel and its sufficient length, which is not always possible, and in some cases extremely undesirable. For example, when there is a threat that the control can be intercepted, or there may be an unwanted object detection by radio signal, or powerful interference in the communication channel, it can lead to loss of controllability and catastrophe, etc [7-8]. To get rid of these and many other shortcomings of existing systems is possible by increasing the level of intelligence of control systems of robotics objects.

[1] A. A. Zhilenkov and I. R. Epifantsev, "System of autonomous navigation of the drone in difficult conditions of the forest trails," *2018 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering*

PROGRAMMATIC LABELLING

The process of automatically generating labels for data using algorithms, rules, or models rather than manual annotation



WHAT ARE LFS?

Labelling functions (LFs) are heuristics that replace the need for hand labelling!

LFS VS HAND LABELLING

Hand labeling	Programmatic labeling
Expensive: Especially when subject matter expertise required	**Cost saving**: Expertise can be versioned, shared, and reused across an organization
Lack of privacy: Need to ship data to human annotators	**Privacy**: Create LFs using a cleared data subsample and then apply LFs to other data without looking at individual samples
Slow: Time required scales linearly with number of labels needed	**Fast**: Easily scale from 1K to 1M samples
Nonadaptive: Every change requires relabeling the data	**Adaptive**: When changes happen, just reapply LFs!

IF LFS NEEDED, THEN TROUBLE LIKELY...

THE PAIN OF HANDLING LACK OF LABELLED DATA!

Method	How	Ground truths required?
Weak supervision	Leverages (often noisy) heuristics to generate labels	No, but a small number of labels are recommended to guide the development of heuristics
Semi-supervision	Leverages structural assumptions to generate labels	Yes, a small number of initial labels as seeds to generate more labels
Transfer learning	Leverages models pretrained on another task for your new task	No for zero-shot learning Yes for fine-tuning, though the number of ground truths required is often much smaller than what would be needed if you train the model from scratch
Active learning	Labels data samples that are most useful to your model	Yes

"Semi-Supervised Learning Literature Survey"

BALANCE IS KEY

...ESPECIALLY WHEN IT COMES TO TRAINING DATASET!

Think about anomaly detection models, rare disease detection etc.

REEL VS REAL

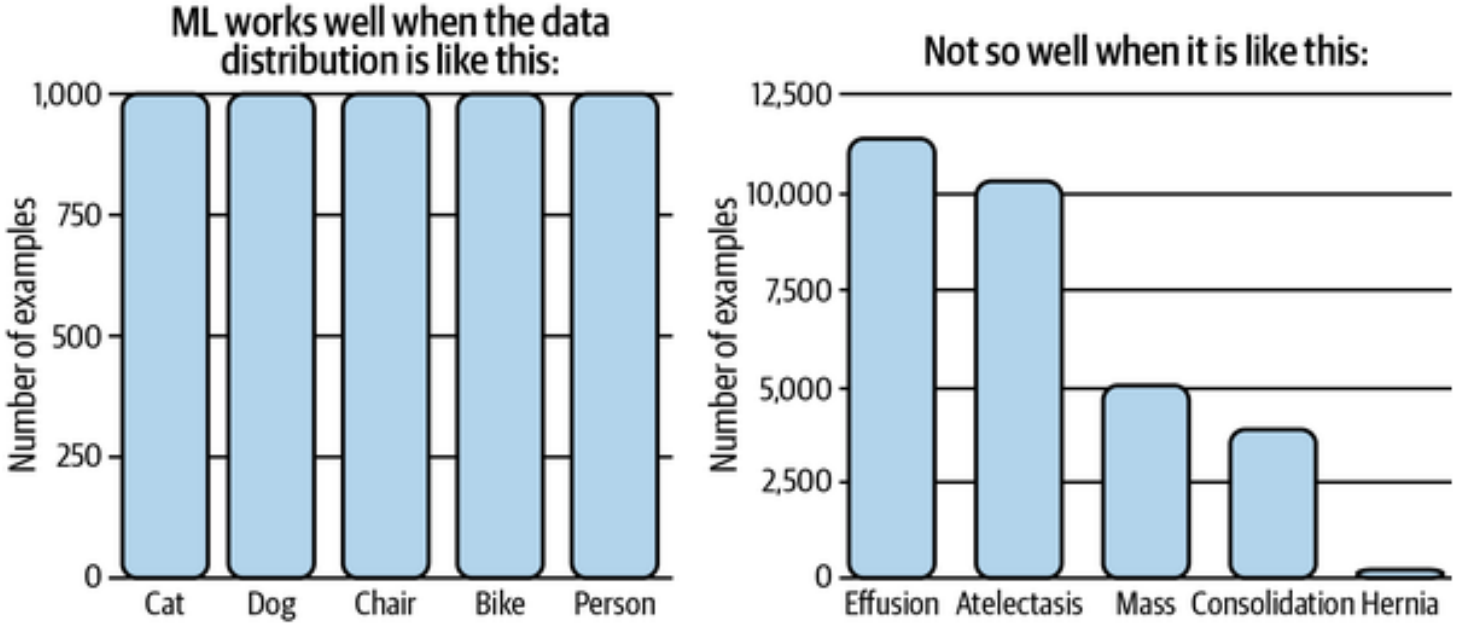
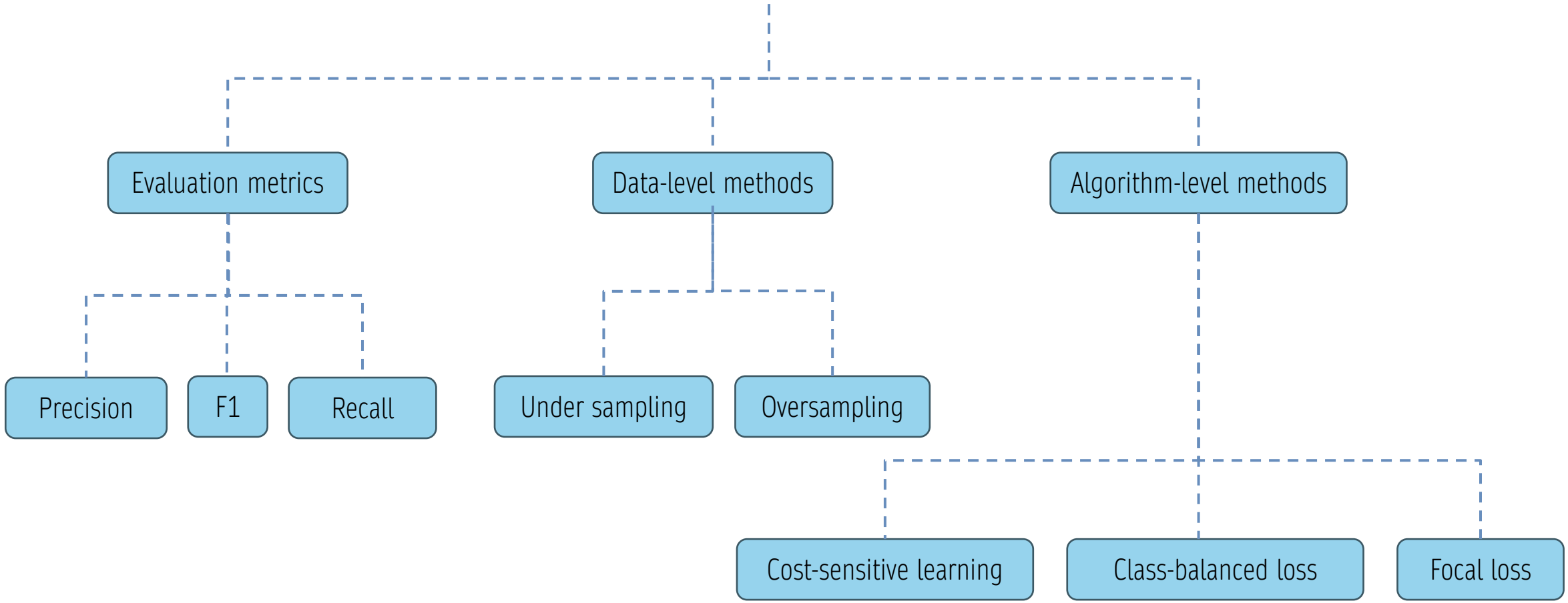


Figure 4-8. ML works well in situations where the classes are balanced. Source: Adapted from an image by Andrew Ng[26]

LET'S HANDLE *CLASS IMBALANCE*



EVALUATION METRICS

Model A	Actual CANCER	Actual NORMAL
Predicted CANCER	10	10
Predicted NORMAL	90	890

Model B	Actual CANCER	Actual NORMAL
Predicted CANCER	90	90
Predicted NORMAL	10	810

	Predicted Positive	Predicted Negative
Positive label	True Positive (hit)	False Negative (type II error, miss)
Negative label	False Positive (type I error, false alarm)	True Negative (correct rejection)

	CANCER (1)	NORMAL (0)	Accuracy	Precision	Recall	F1
A	10/100	890/900	0.9	0.5	0.1	0.17
B	90/100	810/900	0.9	0.5	0.9	0.64

Precision = True Positive / (True Positive + False Positive)

Recall = True Positive / (True Positive + False Negative)

F1 = 2 × Precision × Recall / (Precision + Recall)

DATA LEVEL METHODS

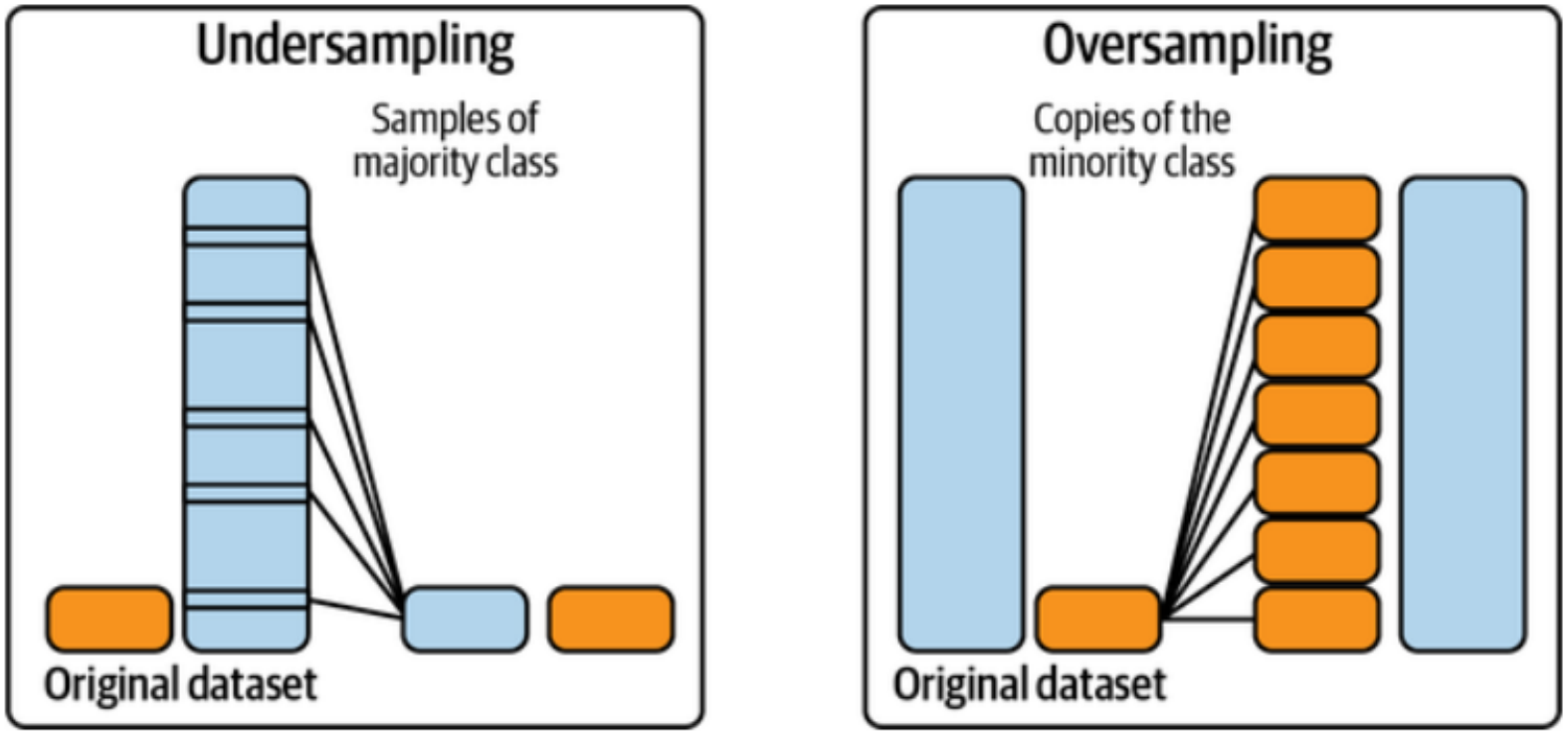


Figure 4-10. Illustrations of how undersampling and oversampling work. Source: Adapted from an image by Rafael Alencar[37]

Tomek links

SMOTE (synthetic minority oversampling technique)

ALGO LEVEL METHODS

Class-balanced loss

$$W_i = \frac{N}{\text{number of samples of class i}}$$

Modified Loss function

$$L(x;\theta) = W_i \sum_j P(j|x;\theta)\,\text{Loss}(x,j)$$

Cost-sensitive learning

	Actual NEGATIVE	Actual POSITIVE
Predicted NEGATIVE	$C(0, 0) = C_{00}$	$C(1, 0) = C_{10}$
Predicted POSITIVE	$C(0, 1) = C_{01}$	$C(1, 1) = C_{11}$

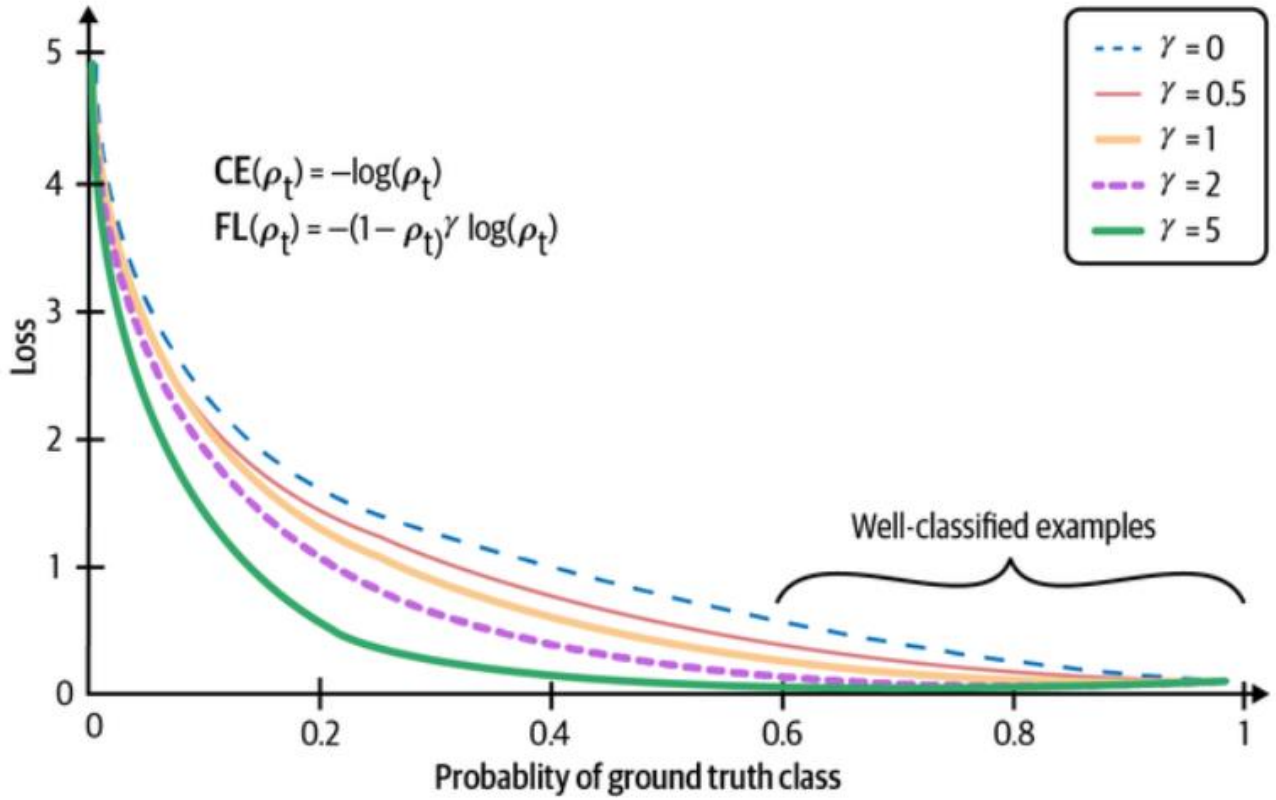
Modified Loss function

$$L(x;\theta) = \sum_j C_{ij} P(j|x;\theta)$$

ALGO LEVEL METHODS

Focal Loss Techniques

Focal Loss modifies the standard cross-entropy loss to address this issue by down-weighting the loss contribution from easy examples and focusing more on hard, misclassified examples.



NOW LET'S AUGMENT

"make (something) greater by adding to it; increase" ~ Oxford dictionary

Image Augmentation:

- Geometric Transformations: Rotation, translation, scaling, and flipping.
- Color Adjustments: Brightness, contrast, saturation, and hue changes.
- Noise Addition: Adding random noise to images.
- Cropping: Random cropping of image sections.
- Normalization: Standardizing pixel values.
- Affine Transformations: Shearing and elastic distortions.

Python toolkits:

TensorFlow (tf.image)

Keras ImageDataGenerator

Albumentations

NOW LET'S AUGMENT

"make (something) greater by adding to it; increase" ~ Oxford dictionary

Language Augmentation:

- Synonym Replacement: Replacing words with their synonyms.
- Back-Translation: Translating text to another language and then back to the original language.
- Random Insertion: Inserting random words into the text.
- Random Deletion: Removing words from the text.
- Text Generation: Using models like GPT to generate variations of the text.

Python toolkits:

NLTK (Natural Language Toolkit)

TextAttack

Transformers

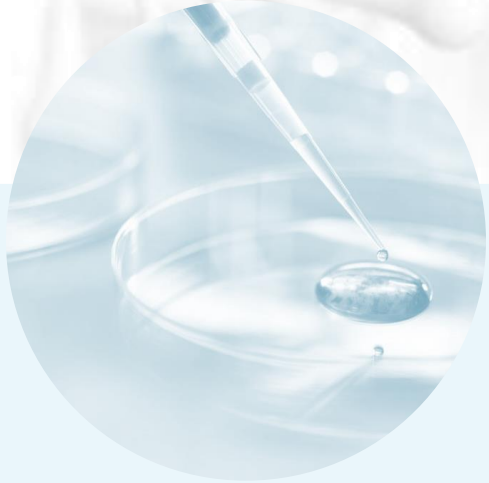
HOW TO IMPLEMENT BACK-TRANSLATION?

YOUR VOICE MATTERS!

Please take some time to fill up our very very short feedback form ☺

If you would like to connect with me, feel free to scan this!

THANK YOU!



Q&A TIME