



2025 Autumn Introduction to Machine Learning Homework 4 Announcement

Release Date: 2024/11/25 15:00

Homework 4

- Deadline: **11:55 a.m., Dec. 14th (Sun), 2025**
- **Kaggle and Report (70%)**: Participat in the **Kaggle competition** and introduce your work.
 - Competition (50%): Get good score.
 - Competition summary slides (20%): **A 3-5 pages slides** (e.g., PowerPoint/Google slides/canva ... No template) to describe your work and highlight how you achieve the performance.
- **Handwritten Questions (30%)**: Answer questions about deep learning.
 - Answer the questions in the report.
 - You **must use the template** and in **digital-typed** (no handwritten scan)
 - In English

Links

- [Questions and Report template](#)
- [Link to Kaggle](#)

Image classification task

Identification of Human (real) VS. AI-generated images

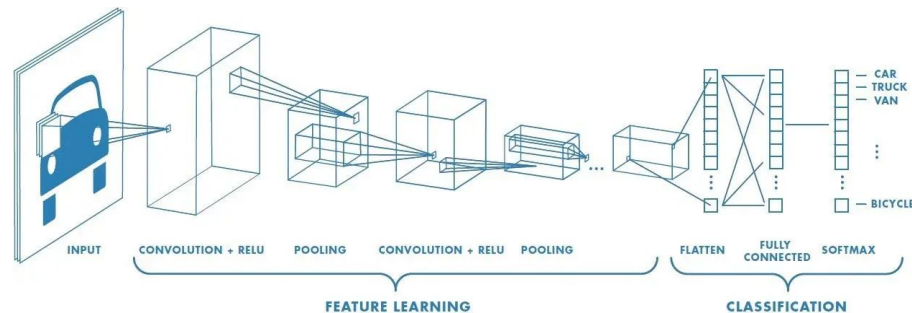
Training data

- 2 classes
 - Real (0) or AI-Generated (1)
- Unlabeled data



Hint

- Convolution Neural Network
- Transfer Learning
- Semi-Supervised learning



Join the Competition



SEAN · COMMUNITY PREDICTION COMPETITION · PRIVATE · 25 DAYS TO GO

Join Competition



Overview

Homework 4 - Identify whether the image is human vs. AI-generated images.

In this task, you're going to train a classification model to classify images into "Real" or "AI-Generated"

It is a binary classification problem, which

- Real (Human-generated) is labeled as 0
- AI is labeled as 1

Don't forget to rename the team name to your [STUDENT_ID]

Watch out the DEADLINE, No late submission in Homework 4

Download the Dataset

Settings Overview **Data** Code Models Discussion Leaderboard Rules Team

Dataset Description

 Edit

The dataset contains 3 parts

1. Train
 - Labeled (Real / Generated)
 - Unlabeled
2. Test

For the training set, the labels are as their folder name. The mapping rule is

```
real: 0
generated: 1
```

Files

80713 files

Size

4.9 GB

Type

jpg

License

CC BY-NC-SA 4.0

data (3 directories)

About this directory

This directory does not have a description yet.



test
16.8k files



train
2 directories



unlabeled
22.4k files

Data Explorer

4.9 GB

- ▼ data
 - ▶ test
 - ▶ train
 - ▶ unlabeled

Summary

▶ 80.7k files

 Download All

+ New Version

Set the Team Name

If TAs cannot find your student ID on the leaderboard, you may get 0.
Please make sure you set the team name to your student ID correctly

- You must set your team name as your **student ID**.



Your Team

Everyone that competes in a Competition does so as a team - even if you're competing by yourself. [Learn more](#).

General

TEAM NAME

412345678

This name will appear on your team's leaderboard position.

Competition Rules

- You are allowed to use any open-source resources/libraries but you must specify them in the report.
 - model architecture (ex: ResNet, ViT, etc.)
 - **But, only ImageNet pretrained weight is allowed**
- Your model size (parameters) have to **less than 100M**.
- You **CAN NOT** use any external data.
- You have to train (finetune) your model **by yourself**.

Any violation of the above rules will result in 0 pts for this part.

ResNet50_Weights.IMAGENET1K_V1:	
These weights reproduce closely the results of the paper using a simple training recipe.	
acc@1 (on ImageNet-1K)	76.13
acc@5 (on ImageNet-1K)	92.862
min_size	height=1, width=1
categories	tench, goldfish, great white shark, ... (997 omitted)
num_params	25557032 ~25.6M
recipe	1000k
GFLOPS	4.09
File size	97.8 MB

Grading Criteria – Performance (50%)

- Public leaderboard
 - As a reference
- Private leaderboard. Your competition score is X and the final score on this part is S
 - Less than weak-baseline ($X < \text{w.baseline}$): $S = 0$
 - Between weak-baseline and the 3rd place: $S = 60 + (X - \text{w.baseline}) / (f1sc.\text{rank3} - \text{w.baseline}) * (100 - 60)) * 0.5$
 - Rank1, 2, and 3: $X=100$; $S= 100 * 0.5$

Leaderboard

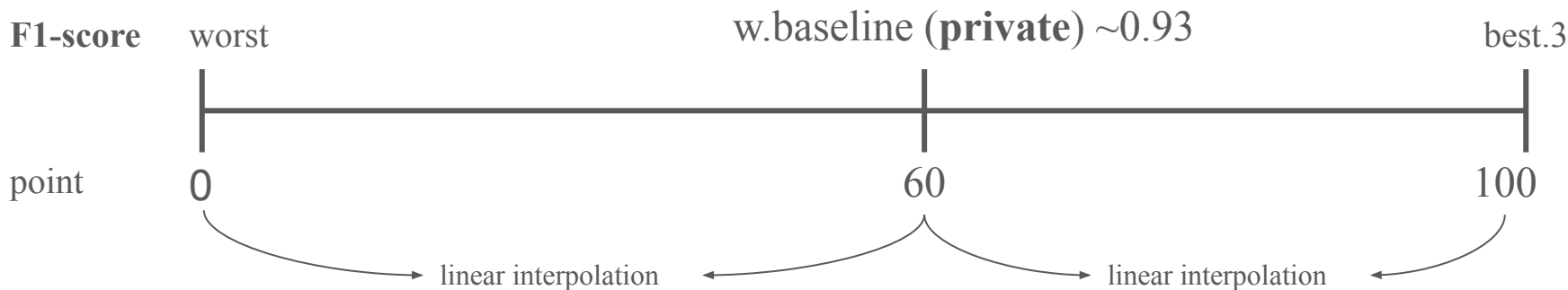
Raw Data Refresh

Search leaderboard

Public Private

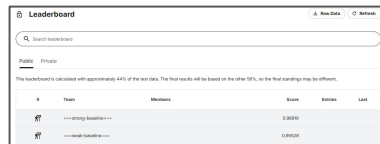
This leaderboard is calculated with approximately 44% of the test data. The final results will be based on the other 56%, so the final standings may be different.

#	Team	Members	Score	Entries	Last
1	==strong-baseline==		0.98919		
2	==weak-baseline==		0.95528		



Report (Competition report slide presentation - 20%)

- **Make slides** (e.g., MS PowerPoint, Google Slides, etc.) to summarize your work to TAs within **3-5 pages** (no title page and thank-you page needed).
- Things you should cover, include and not limit to
 - How do you process the data? Have you done any special processing that significantly boosted performance?
 - What is your model architecture? Do you make any changes/modifications to the model? Does these modifications improve performance?
 - How do you train the model? Do you use any special techniques (e.g., ensembles or other methods) to improve performance?
 - Other details you want to mention that improve the performance
- Paste the snapshot of your Kaggle public leaderboard in the appendix page (Not count to the page limit, -5 pts if not provided)
- Hint
 - Make your slide presentation clear and informative, and TAs will evaluate its completeness and content.
 - Read some literature to see how they report and highlight their achievement.



The image shows a screenshot of the Kaggle public leaderboard. At the top, there's a search bar and tabs for 'Public' and 'Private'. Below the tabs, a note states: 'This leaderboard is calculated with approximately 44% of the test data. The final results will be based on the other 56%, so the final standings may be different.' The main part of the image is a table with columns: '#', 'Team', 'Mean', 'Score', 'Status', and 'Last'. There are two rows of data visible, both with placeholder text '---getting leaderboard---' in the 'Team' column and '0.0000' in the 'Score' column.

#	Team	Mean	Score	Status	Last
1	---getting leaderboard---		0.0000		
2	---getting leaderboard---		0.0000		

Numpy & PyTorch

- Numpy Tutorial: [Link](#)
- PyTorch Tutorial: [Link](#)
 - Free to use any modules and functions

Environment

- Python version: 3.9 or higher
- If you have a GPU
 - [Conda](#)
 - [Miniconda](#)
 - [virtualenv](#)
 - ...
- If you don't have a GPU
 - [Google Colab](#)

Handwritten Questions (30%)

2-1 (10%) Based on the “SVMs vs. Logistic regression” slide, explain which kinds of training data points mainly determine the classifier learned by SVM and which kinds of points influence Logistic Regression, and briefly justify your answer by referring to the shapes of hinge loss and logistic loss.

2-2 (15%) For an SVM with a Linear Kernel, determine whether to use the Primal or Dual Form for the datasets below. Justify your choice based on: (a) Optimization variables & computational complexity, (b) Memory requirements (specifically the size of the Gram Matrix), (c) Prediction cost.

- Dataset A: $N=100$, $M=20,000$
- Dataset B: $N=1,000,000$, $M=20$

Handwritten Questions (30%)

2-3 (5%) To train a neural network, what do we need to optimize it? (How do we know the network is good or not?) Also, what algorithm can we use to optimize the neural network? (The most basic one)

Report (Hand-written questions, 30%)

- Please follow the report template format. (-5pts if not use the template)
- [Link](#)

Submission

- Compress your code, competition summary slides and report into a **.zip file** and submit it to E3.
- All report/slides should be written in English. (-5 pts if not English)
- <STUDENT ID>_HW4.zip
 - [YOUR CODE] (Make sure it is well-organized and can be executed)
 - Training code, Inference code, README, etc. (README should contains how to setup your env and run the code)
 - Model weights: <STUDENT ID>_weight.txt (which contains a link to your google drive & ensure the permission is granted for everyone, so that TAs can download it if needed.)
 - <STUDENT_ID>_CompetitionSlide.pptx (or .pdf)
 - <STUDENT ID>_HW4.pdf (NO .doc, .docx or others format)
- Don't put the data (e.g. train.csv / test.csv) and model weights into submission file
 - -10 pts if you put the wrong things into zip file

Kaggle Submission Reproduction

- Your inference script should be able to reproduce your kaggle submission.
 - Build and load the model (with the model weights you provide) and then generate your kaggle submission file.
- For python file (inference.py)
 - It will be checked on our lab's servers. (NVidia 1080Ti, cuda 11.3)
 - Please provide a **environment setup instruction** which can help us quickly rebuild your environment and accurately reproduce your results.
- For jupyter notebook file (inference.ipynb)
 - It will be checked on [Google Colab](#).
 - Please include the necessary **install instructions** (e.g., pip install) in the first cell.

Other rules

- **Late Policy**: **No Delay allowed** this time. (delayed, a.k.a. submission closed)
 - **You will get 0 pts on HW4**
- **No Plagiarism**: You should complete the assignment by yourself. Students engaged in plagiarism will be penalized heavily. Super serious penalty.
 - **Opt for the assignment or failed this course, etc**
 - **Report to academic integrity office**

AI-Assistant

- Not recommended but no forbidden
- Copy-and-Paste answers from the AI-Assiant will be seen as Plagiarism
 - However, you can have your own answer first then rephrase it by AI-Assiant.
- Some questions might be parts of final exam, make sure you understand the concept



FAQs

- If you have other questions, ask on [E3 forum](#) first! We will reply as soon as possible.
 - Also, feel free to write email to TAs (And remember to cc all TAs).

Have Fun!

