Introduction to Machine Learning and Deep Learning: Final Project

潤羽るしあ🦋ホロライブ3期生

Member

- Oscar Chew (周寬)
  - R10922154
  - Department of Computer Science and Information Engineering

Data Analysis

- The problem of imbalance is not significant. No special treatment was taken.

	Number of examples
Class 0	2775
Class 1	1687
Class 2	1904
Class 3	2913

Data Analysis

- We examined the mean width/height of the images in training/testing set. This helped us in designing the preprocessing pipeline.

	Mean width	Mean height
Training set	559	570
Testing set	596	603

Image Preprocessing

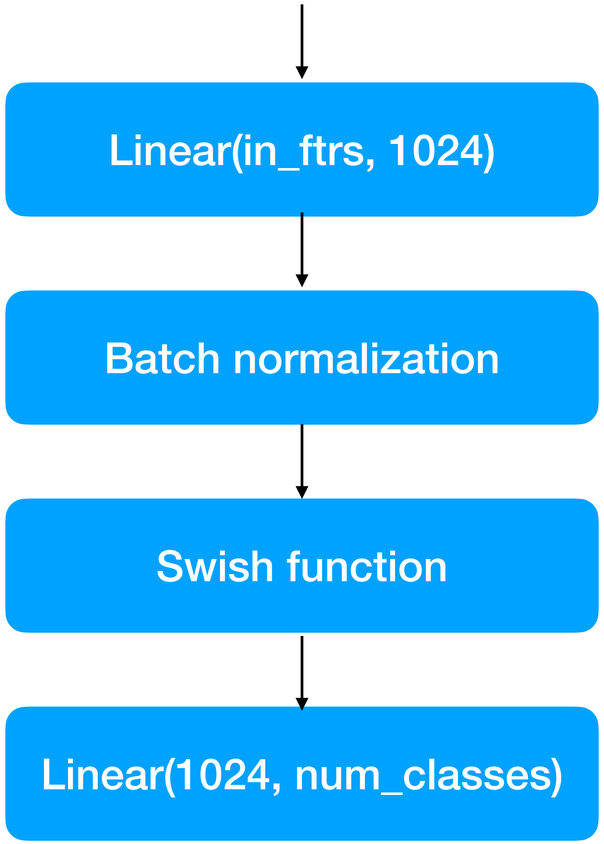
- Resize each image to 600x600 (match our observation earlier).
- Random horizontal flips with probability 0.5 (for training set).
- Normalize the pixel values using the ImageNet statistics.
  - mean=[0.485, 0.456, 0.406], std=[0.229, 0.224, 0.225]
- Random vertical flips does not make much sense in this dataset.
- Center crops may not be good at identifying pants/dresses

Image Preprocessing

- Things to keep in mind:
  - Does the transformation preserve the labels?
  - Does the transformation generate realistic (unseen) data?
  - Does the transformation help to avoid overfitting?

Model Architecture

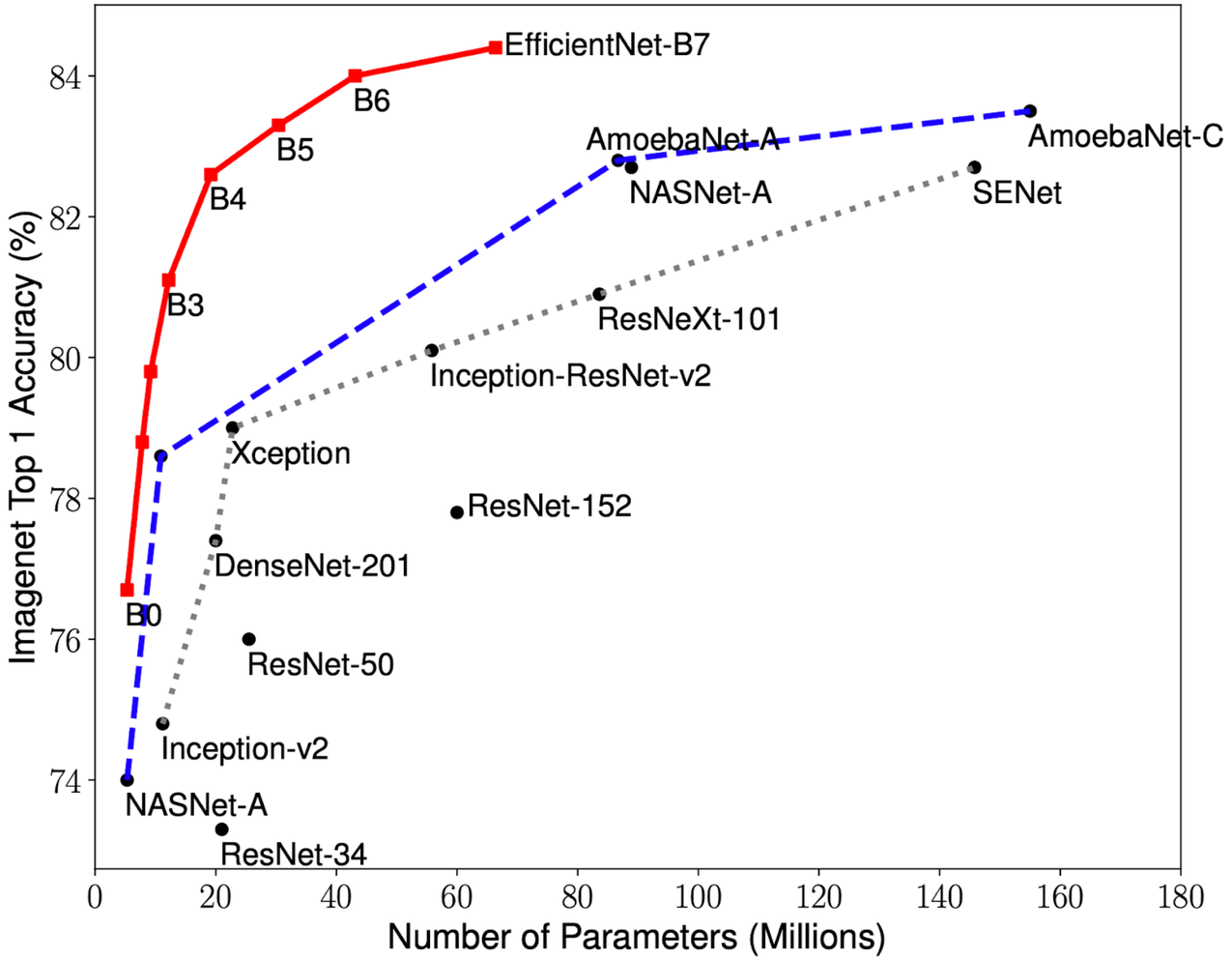
- Backbone: EfficientNet-B7 + self-training with noisy student.
- Classifier:



EfficientNet

- We examined ImageNet classification leaderboard.
  - Looked for a model with good performance and less parameters(since we are poor students 😢).
  - EfficientNet - best performing CNN, small model size.
    Uniformly scales all dimensions of depth/width/resolution.
- Vision Transformers(ViT) is a very strong competitor.

EfficientNet

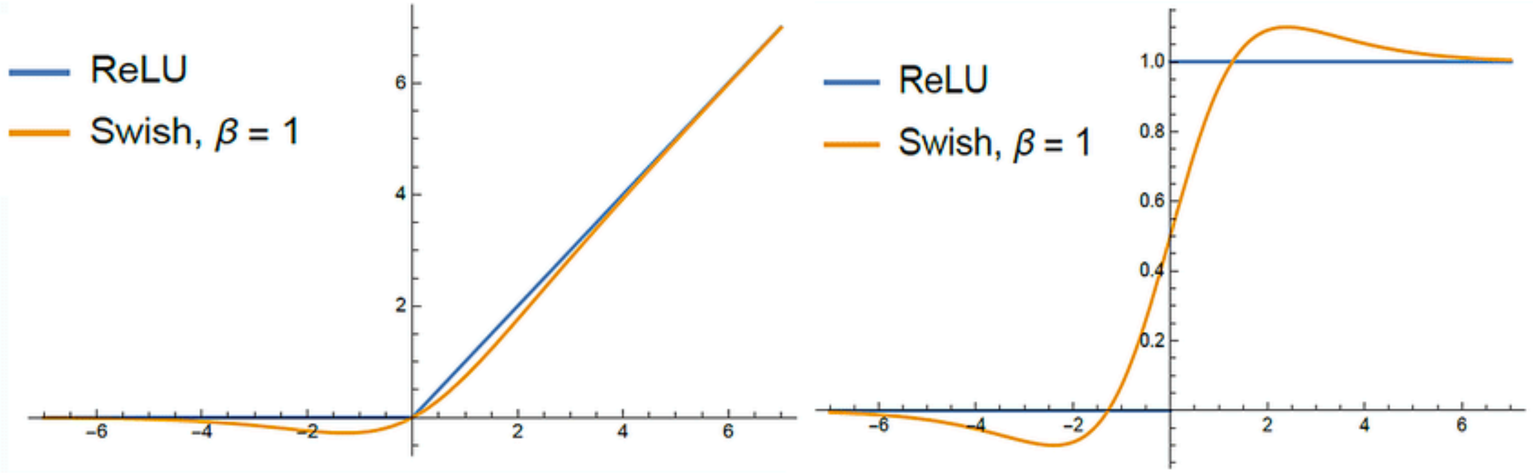


EfficientNet

Base model	resolution
EfficientNetB0	224
EfficientNetB1	240
EfficientNetB2	260
EfficientNetB3	300
EfficientNetB4	380
EfficientNetB5	456
EfficientNetB6	528
EfficientNetB7	600

Swish Function

- Swish is defined as $f(x) = x \cdot \sigma(\beta x)$. We replaced ReLU in our classifier with Swish



- Swish is smooth and non-monotonic.

Semi-supervised Learning

- Self-training with noisy student

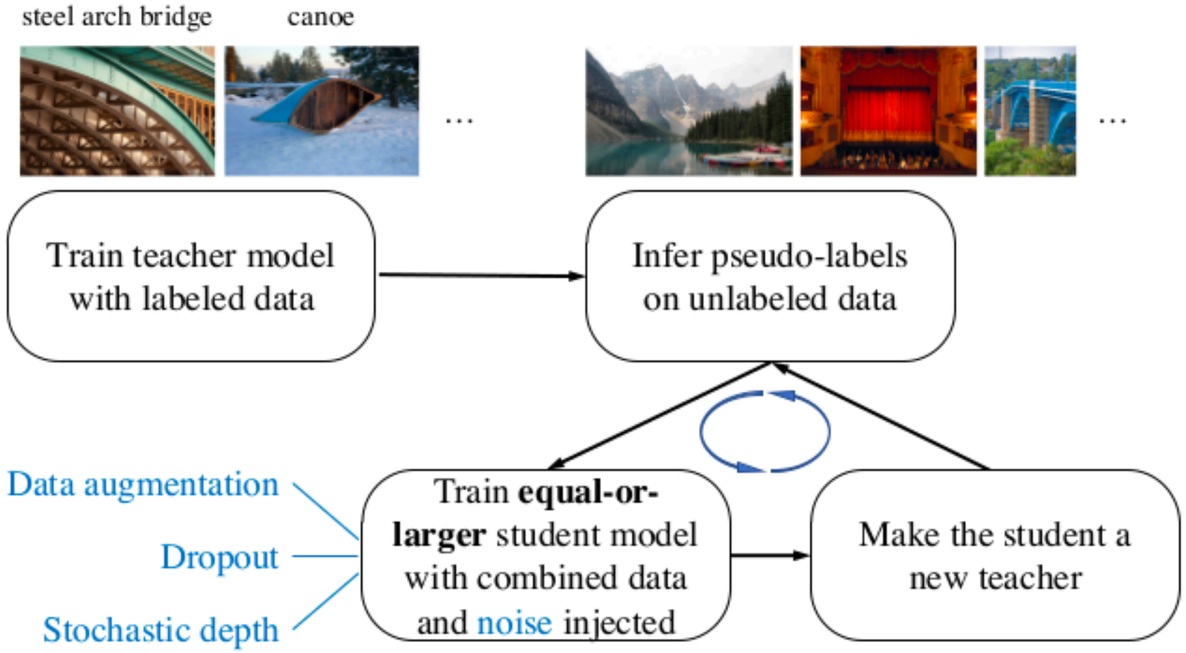


Figure 1: Illustration of the Noisy Student Training. (All shown images are from ImageNet.)

Semi-supervised Learning

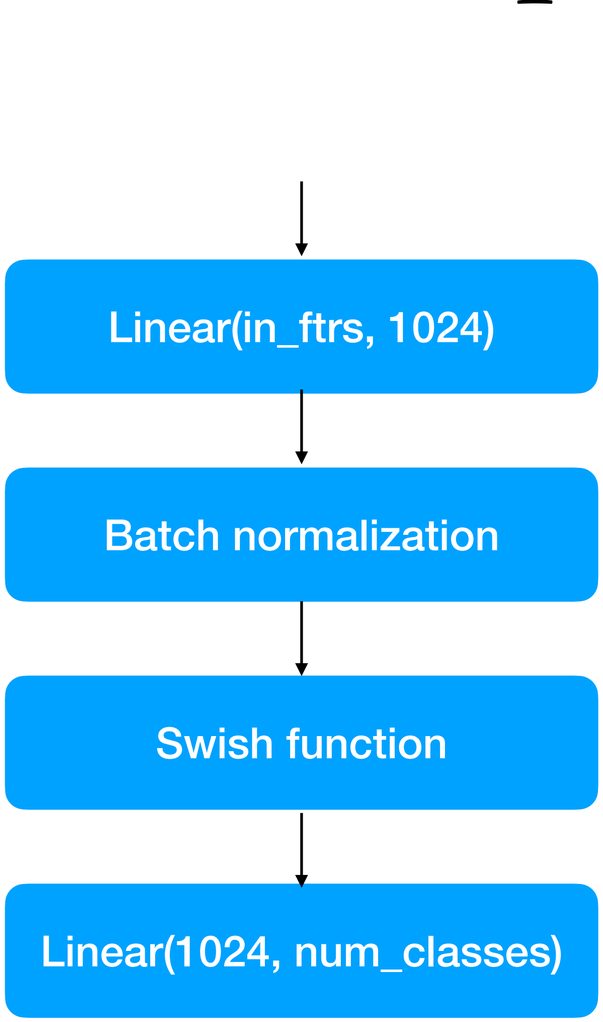
- We leveraged the technique of Noisy Student Training twice:

    1. We loaded an EfficientNetB7 weight that underwent Noisy Student Training.

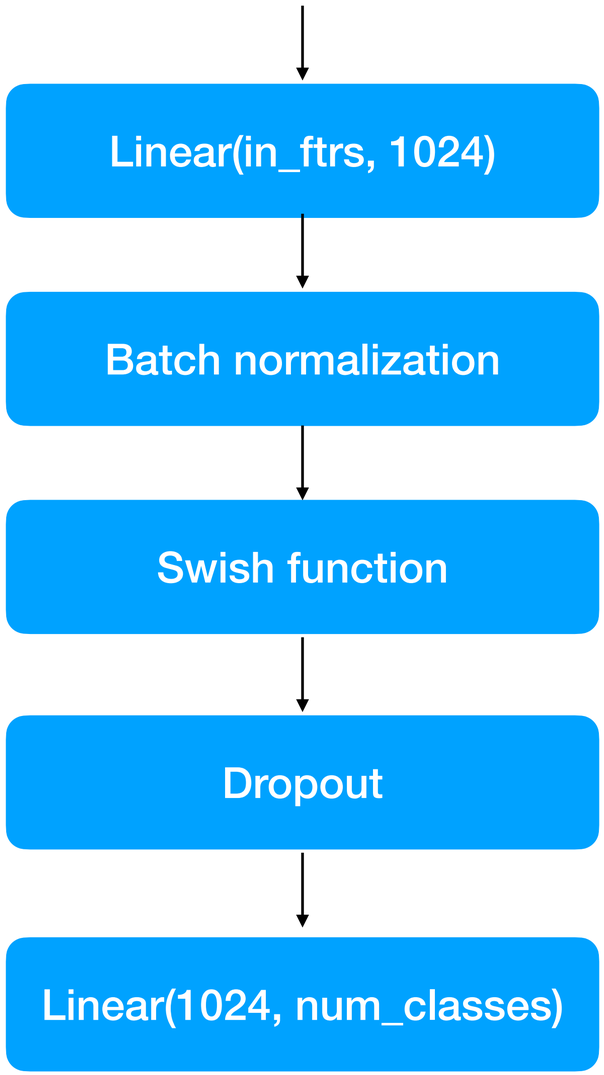
    2. We implemented Noisy Student Training to make good use of the additional unlabeled data.

        - Only dropout(model noise) was added to the student network.

Semi-supervised Learning



Teacher



Student

Semi-supervised Learning

- Confirmation bias
  - Teacher model will inevitably make mistakes when predicting the labels of unlabeled data
  - Student model will train on data with incorrect labels

Label Smoothing

- A regularization technique
- Cross entropy loss:
  - $-\sum_{i=1}^{K} p_i \log q_i$ where $p_i = \begin{cases} 1 & i = y \\ 0 & i \neq y \end{cases}$
- Label smoothing + cross entropy loss:
  - $p_i = \begin{cases} 1 - \varepsilon & i = y \\ \varepsilon / K & i \neq y \end{cases}$

Label Smoothing

- Rule of thumb: $\varepsilon = 0.1$
- Without label smoothing
  - Class 1 label: [1, 0, 0, 0]
- With label smoothing
  - Class 1 label: [0.9, 0.0333, 0.0333, 0.0333]
- Prevent the network from becoming over-confident.

Hyperparameters

- Early-stopping patience = 5
- Batch size = 8
- Learning rate = 1e-4 for AdamW optimizer
  - It is sensible to have a smaller learning rate for finetuning
- Used default values for all other hyperparameters

Ensemble

- We (uniformly) blended 7 models.
  - Vanilla EfficientNetB7
  - Vanilla EfficientNetB7 with pseudo-labeling
  - Vanilla EfficientNetB7 with an additional rotation transformation
  - EfficientNetB7-NS
  - EfficientNetB7-NS with Swish
  - EfficientNetB7-NS with Swish and noisy student training
  - EfficientNetB7-NS with Swish, noisy student training, and label smoothing

Result

- Rank #1 on public leaderboard (0.86785)

1		潤羽るしあ🦋ホロライブ3期生		0.86785	33	2d

- Rank #1 on private leaderboard (0.87357)

1	—	潤羽るしあ🦋ホロライブ3期生		0.87357	33	2d

Room For Improvement

- Search for better hyperparameters.
- Worth trying Vision Transformers.
- The semi-supervised method we mentioned still suffer from confirmation bias
  - Meta Pseudo Labels (CVPR2021) may be a good solution.

Summary

1. Transfer learning
    - EfficientNet with Swish activation function
2. Image preprocessing
3. Noisy student training + label smoothing
4. Uniform blending