# Effect of Stage Training for Long-Tailed Multi-Label Image Classification

Yosuke Yamagishi, Shohei Hanaoka
The University of Tokyo, Japan

## OVERVIEW

**Focus:**

Multi-stage training approach for image classification in the ICCV CVAMD 2023 Shared Task CXR-LT: Multi-Label Long-Tailed Classification on Chest X-Rays.

**Stage Training:**

Adjusting input image size and batch size at each training stage.

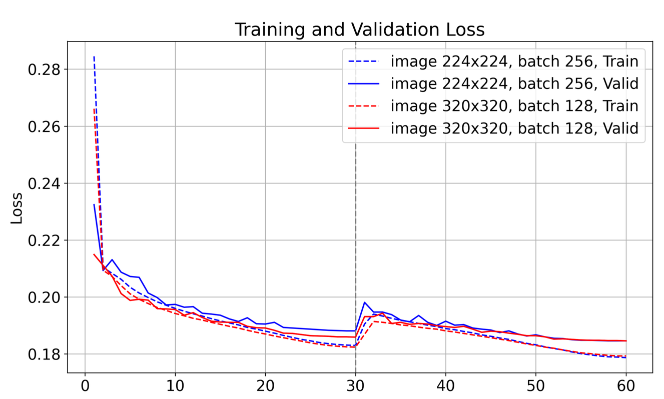
- First stage: Smaller image size, larger batch size.
- Second stage: Increased image size, reduced batch size.

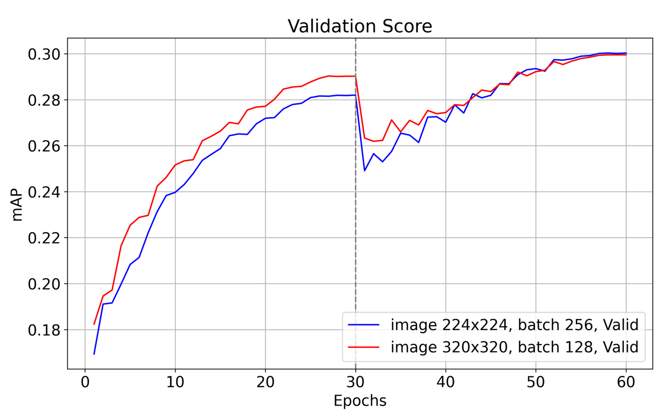
## MODELS

| Model | Stage | Pre-Weight | Image size | Batch size | Max LR | OS-threshold | Val mAP | Dev mAP |
|---|---|---|---|---|---|---|---|---|
| model0 | 1st | ImageNet | 224x224 | 256 | 0.001 | N/A | 0.282 | 0.298 |
| model1 | 1st | ImageNet | 224x224 | 128 | 0.001 | N/A | 0.268 | 0.256 |
| model2 | 2nd | model0 | 320x320 | 128 | 0.001 | N/A | 0.300 | 0.302 |
| **model3** | 2nd | model0 | 320x320 | 128 | 0.001 | 0.10 | 0.303 | 0.304 |
| model4 | 2nd | model1 | 320x320 | 128 | 0.001 | 0.10 | 0.297 | 0.298 |
| **model5** | 2nd | model0 | 320x320 | 128 | 0.001 | 0.15 | 0.299 | N/A |
| **model6** | 2nd | model0 | 384x384 | 80 | 0.001 | 0.10 | 0.301 | 0.306 |
| **model7** | 3rd | model3 | 512x512 | 48 | 0.001 | 0.10 | 0.305 | 0.304 |

All models used EfficientNetV2-S as the backbone [1]. The training stages are indicated by the "Stage" column. Pre-trained weights for stage 1 models come from ImageNet [2], while stage 2 used either model0 or model1. The "OS-threshold" column shows the threshold for oversampling, with "N/A" meaning no oversampling. "Val mAP" and "Dev mAP" represent the mean average precision (mAP) on validation and development data, respectively. Four models in bold were used in the ensemble.

## COMPARISON




The graph displays loss function curves for training and validation, with a separate mAP curve for validation data. The first stage, up to the 30th epoch, used two settings: "image 224x224, batch 256" and "image 320x320, batch 128". The second stage, up to the 60th epoch, standardized both experiments to "image 320x320, batch 128".

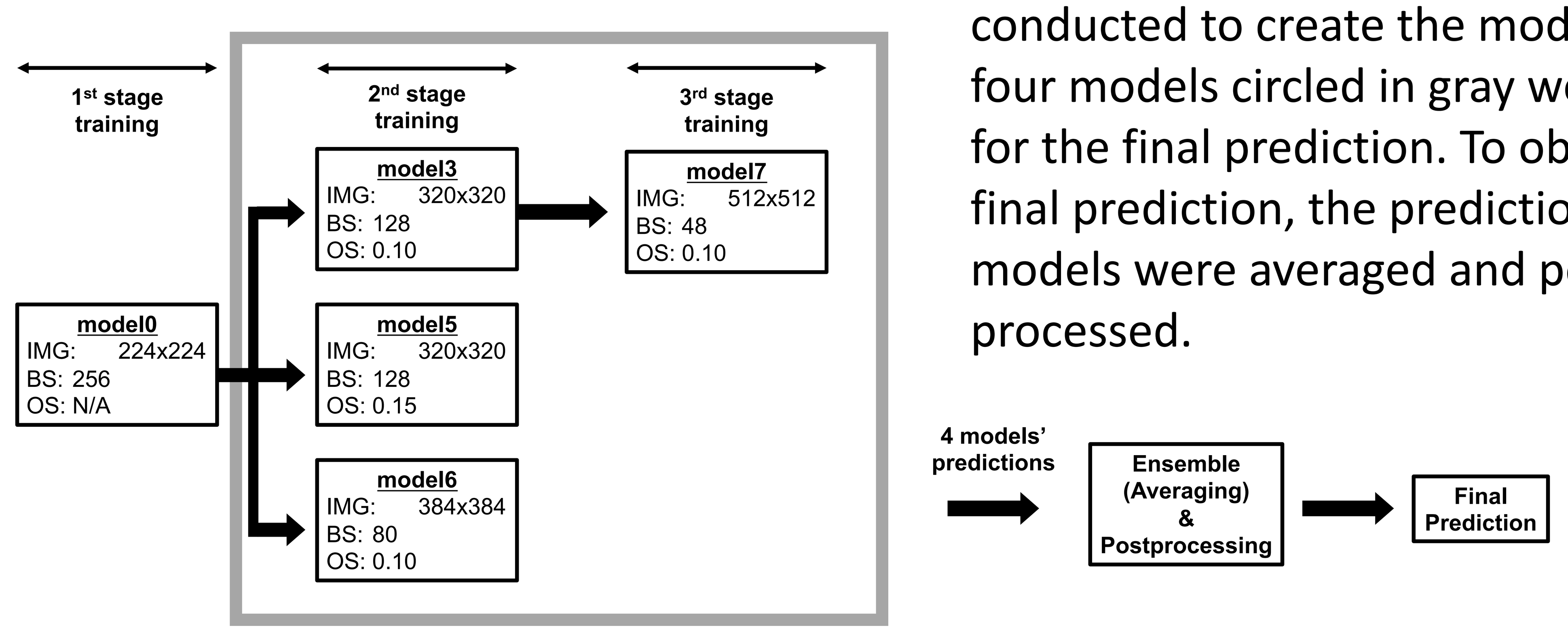
At the start of stage2, the optimizer and scheduler were reset, therefore both loss and mAP showed a temporary deterioration. Under both settings, the final loss and mAP were almost equal.

## RESULT

| Label | 1st stage | 2nd stage | Oversampling | Postprocess | Ensemble |
|---|---|---|---|---|---|
| Support Devices | 0.870 / 0.865 | **0.887 / 0.884** | 0.885 / 0.881 | **0.888 / 0.885** | **0.897 / 0.894** |
| Lung Opacity | 0.580 / 0.561 | **0.593 / 0.571** | 0.588 / 0.569 | **0.600 / 0.581** | **0.611 / 0.590** |
| Cardiomegaly | 0.646 / 0.647 | 0.646 / **0.648** | 0.644 / 0.646 | **0.663 / 0.665** | **0.669 / 0.668** |
| Pleural Effusion | 0.802 / 0.810 | **0.808 / 0.817** | 0.805 / 0.816 | 0.807 / 0.816 | **0.812 / 0.822** |
| Atelectasis | 0.569 / 0.575 | **0.579 / 0.583** | 0.578 / 0.583 | **0.593 / 0.596** | **0.600 / 0.602** |
| Pneumonia | 0.288 / 0.281 | **0.296 / 0.289** | 0.290 / 0.286 | 0.283 / 0.286 | **0.301 / 0.292** |
| No Finding | 0.459 / 0.462 | **0.461 / 0.467** | **0.464** / 0.464 | 0.460 / 0.464 | **0.470 / 0.471** |
| Edema | 0.525 / 0.521 | **0.536 / 0.532** | 0.533 / **0.533** | **0.545 / 0.543** | **0.551 / 0.551** |
| Enlarged Cardiomediastinum | 0.166 / 0.171 | **0.168 / 0.177** | 0.166 / 0.176 | 0.168 / **0.181** | **0.171 / 0.183** |
| Consolidation | 0.210 / 0.203 | 0.210 / **0.206** | 0.205 / 0.202 | **0.212 / 0.216** | **0.216 / 0.225** |
| Pneumothorax | 0.350 / 0.367 | **0.399 / 0.418** | 0.393 / 0.416 | **0.407 / 0.431** | **0.421 / 0.451** |
| Fracture | 0.102 / 0.107 | **0.123 / 0.120** | **0.124** / 0.120 | **0.146 / 0.133** | **0.186 / 0.171** |
| Infiltration | 0.061 / 0.053 | **0.064 / 0.056** | 0.063 / 0.055 | 0.064 / 0.055 | 0.064 / 0.056 |
| Nodule | 0.098 / 0.093 | **0.113 / 0.105** | 0.111 / 0.103 | **0.117 / 0.112** | **0.153 / 0.137** |
| Mass | 0.164 / 0.152 | **0.166 / 0.168** | 0.163 / 0.165 | **0.185 / 0.175** | **0.198 / 0.187** |
| Calcification of the Aorta | 0.081 / 0.075 | **0.110 / 0.103** | 0.109 / 0.098 | **0.141 / 0.112** | **0.156 / 0.130** |
| Emphysema | 0.268 / 0.149 | 0.263 / **0.155** | **0.278** / 0.146 | **0.284** / 0.146 | **0.279 / 0.161** |
| Hernia | 0.478 / 0.457 | **0.506 / 0.487** | **0.513** / 0.481 | **0.491** / 0.481 | **0.531 / 0.499** |
| Pleural Thickening | 0.063 / 0.063 | **0.078 / 0.073** | 0.069 / **0.078** | **0.094 / 0.105** | **0.114 / 0.119** |
| Tortuous Aorta | 0.059 / 0.052 | **0.064 / 0.055** | **0.065** / 0.055 | **0.069 / 0.058** | **0.066 / 0.063** |
| Lung Lesion | 0.055 / 0.029 | **0.068 / 0.030** | 0.062 / 0.028 | 0.055 / 0.028 | 0.064 / **0.031** |
| Subcutaneous Emphysema | 0.338 / 0.457 | **0.427 / 0.488** | 0.416 / 0.457 | **0.450** / 0.471 | **0.465 / 0.507** |
| Fibrosis | 0.087 / 0.105 | **0.117 / 0.108** | 0.114 / **0.110** | 0.117 / 0.106 | **0.148 / 0.116** |
| Pleural Other | 0.017 / 0.024 | **0.022 / 0.036** | **0.037** / 0.026 | 0.028 / 0.032 | **0.050 / 0.059** |
| Pneumomediastinum | 0.075 / 0.223 | 0.075 / **0.266** | **0.139 / 0.296** | **0.152 / 0.319** | 0.149 / **0.326** |
| Pneumoperitoneum | 0.020 / 0.076 | **0.085 / 0.175** | **0.123 / 0.195** | **0.129 / 0.206** | **0.134 / 0.262** |
| Mean | 0.286 / 0.291 | **0.302 / 0.308** | **0.305** / 0.307 | **0.313 / 0.315** | **0.326 / 0.330** |

The table shows the AP for each label and the mAP (Development / Test data). The AP that achieved the best performance up to that point is highlighted in bold. The second stage training improved the APs of almost all labels. The post-processing used the view position in the same study and unifies the prediction of PA or AP images. The post-processing and ensemble (averaging multi predictions) improve the APs remarkably.

## Flowchart




Flowchart of training and prediction. A total of three stage trainings were conducted to create the models. The four models circled in gray were used for the final prediction. To obtain the final prediction, the predictions of the models were averaged and post-processed.

## Conclusion

We utilized stage training, dividing the learning process into multi-stages. By using a smaller image size and a larger batch size in the first stage, we were able to optimize the learning efficiency. In the second stage, we achieved performance equivalent to training with a larger image size from the start. This method has been confirmed to be effective for nearly all labels within long-tailed data.

[1] Mingxing Tan and Quoc V. Le. "EfficientNetV2: Smaller Models and Faster Training." arXiv preprint arXiv:2104.00298 (2021).
[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition, pages 248–2.