INTERMEDIATE LAYERS MATTER IN MOMENTUM CONTRASTIVE SELF SUPERVISED LEARNING

Aakash Kaku, Sahana Upadhya, Narges Razavian
NYU Center for Data Science

SUMMARY
• Bringing intermediate layers’ representations of two augmented versions of an image closer together helps to improve the momentum contrastive (MoCo) method.
• We show this improvement for two loss functions: the mean squared error (MSE) and Barlow Twin’s loss between the intermediate layer representations; and three datasets: NIH-Chest Xrays, Breast Cancer Histopathology, and Diabetic Retinopathy.
• Improved MoCo has large gains (~5%) in the performance especially when we are in a low-labeled regime (1% data is labeled).
• Improved MoCo learns meaningful features earlier in the model and also has high feature reuse.

MoCo FOR MEDICAL DATASETS
• MoCo is widely adopted by medical imaging research community.
• MoCo differentiates from others by having an additional bank of negative examples.
• This helps the model to learn even when the batch sizes are small.
• For medical images, this is important because medical images generally have high resolution and large batch sizes are computationally infeasible.
• Many recent works ([1,2,3,4]) have used MoCo for various tasks therefore we focus our study on MoCo.

PROPOSED METHOD
• The schematic for our proposed method:

PROPOSED METHOD
• Encoder should be encouraged to learn augmentation-invariant representations not only at the end but also for the intermediate layers.
• We use MSE and Barlow Twins loss functions to ensure closeness of intermediate representations.
• Intermediate loss function have shown to help by regularizing the model or improving gradient flow.
• In our case, it helps to learn augmentation-invariant features early in the model thereby learning high quality features for performing the downstream task.
• We run all our experiments with ResNet-50 encoder. We apply the loss function after each of the four ResNet blocks in ResNet-50 architecture.

DATASETS
• We empirically verify our proposed approach with three diverse medical imaging datasets: NIH-Chest Xrays, Breast Cancer Histopathology, and Diabetic Retinopathy.
• Classification is the downstream task for each dataset.

RESULTS
• Performance after fine tuning only the final linear layer. MoCo+MSE approach outperforms the standard MoCo for the three datasets.

ANALYSIS OF FEATURES
• Feature reuse: Higher feature reuse indicates higher feature quality. We measure feature reuse by measuring feature similarity before and after fine tuning them using a labeled dataset.
• Layer-wise Probing: Our approach also learns more useful features earlier in the model.

Label fraction Supervised MoCo MoCo + MSE MoCo + Barlow Twins Supervised
NII Chest Xray (AUC (95% CI))
100% 79.8 (79.2-80.3) 82.4 (81.7-83.0) 81.5 (80.9-82.1) 80.0 (79.5-80.7)
6% 65.2 (64.6-65.8) 69.8 (69.3-70.4) 70.5 (69.7-71.0) 70.0 (69.2-70.6)
1% 57.8 (57.2-58.4) 59.2 (58.6-59.9) 61.4 (60.7-62.0) 62.9 (62.3-63.5)
NiCo + Diabetic Retinopathy (AUC (95% CI))
100% 94.1 (94.1-94.2) 94.6 (94.3-94.9) 96.6 (96.6-97.4) 95.5 (95.7-95.8)
6% 69.1 (68.6-69.2) 72.4 (72.2-72.6) 95.1 (94.8-95.5) 94.0 (94.0-94.3)
1% 65.5 (65.4-65.6) 68.5 (68.1-68.9) 95.6 (95.6-96.4) 95.5 (95.2-96.2)

MoCo + Breast Cancer Histopathology (AUC (95% CI))
100% 82.7 (82.4-83.1) 82.9 (82.6-83.3) 85.7 (85.4-86.0) 86.4 (86.1-86.7)
6% 82.7 (82.4-83.1) 82.8 (82.4-83.2) 85.6 (85.2-86.0) 84.8 (84.2-84.8)
1% 80.6 (80.3-81.0) 81.8 (82.5-83.3) 85.1 (84.7-85.4) 84.4 (84.1-85.7)

REFERENCES
2. Dehaene, D et al. Self-supervision closes the gap between weak & strong supervision in histology.