# INTERMEDIATE LAYERS MATTER IN MOMENTUM CONTRASTIVE SELF SUPERVISED LEARNING

NYU School of Medicine
NYU LANGONE MEDICAL CENTER

NYU CENTER FOR DATA SCIENCE

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

## SUMMARY

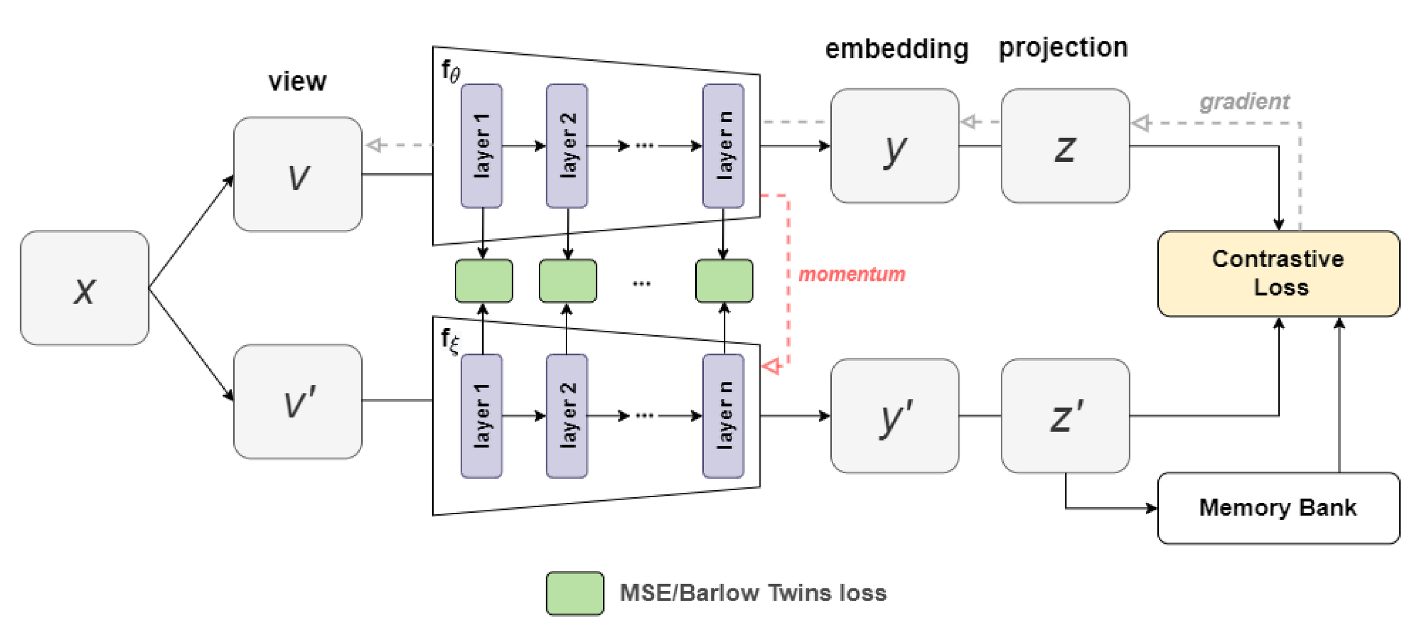
- Bringing intermediate layers' representations of two augmented versions of an image closer together helps to improve the momentum contrastive (MoCo) method
- We show this improvement for two loss functions: the mean squared error (MSE) and Barlow Twin's loss between the intermediate layer representations; 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.

| Dataset / Method | MoCo | MoCo + MSE | MoCo + Barlow Twins | Supervised |
|---|---|---|---|---|
| NIH Chest X-ray (AUC (95% CI)) | 74.4 (73.9-75.0) | **74.8 (74.2-75.4)** | 73.5 (72.9-74.0) | 79.8 (79.2-80.3) |
| Diabetic Retinopathy (AUC (95% CI)) | 74.6 (74.5-74.7) | **84.8 (84.6-85.0)** | 79.7 (79.6-79.7) | 94.1 (94.1-94.2) |
| Breast Cancer Histopathology (F1-score (95% CI)) | 80.7 (80.4-81.1) | **82.5 (82.2 - 82.9)** | 82.3 (82.0-82.7) | 82.7 (82.4-83.1) |

- Performance after **fine tuning the entire encoder** for different percentage of label data. For **low labeled data regime** (1% and 6% of labeled data), our proposed approach **improves** over standard MoCo performance by ~5% when averaged across three datasets.

| Label fraction | Supervised | MoCo | MoCo + MSE | MoCo + Barlow Twins |
|---|---|---|---|---|
| NIH Chest X-ray (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 (69.9-71.0)** | 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)** |
| Diabetic Retinopathy (AUC (95% CI)) | | | | |
| 100% | 94.1 (94.1-94.2) | 94.6 (94.3-94.6) | **96.6 (96.6-96.7)** | 95.7 (95.7-95.8) |
| 6% | 69.1 (69.0-69.2) | 92.4 (92.2-92.6) | **95.1 (94.8-95.2)** | 94.0 (94.0-94.3) |
| 1% | 65.5 (65.4-65.6) | 88.1 (88.1-88.4) | **93.6 (93.2-93.6)** | 92.5 (92.2-92.7) |
| Breast Cancer Histopathology (F1-score (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) | **84.6 (84.2-84.9)** | 84.5 (84.2-84.8) |
| 1% | 80.6 (80.3-81.0) | 82.8 (82.5-83.2) | **85.1 (84.7-85.4)** | 84.4 (84.1-84.7) |

## 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. **Our approach shows higher feature reuse** as compared to the standard MoCo for all three datasets.
- We also see that initial layers have higher feature reuse which indicate SSL helps to learn low-level image statistics.

| | 1% labeled data | | | | | 6% labeled data | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Method | Block 1 | Block 2 | Block 3 | Block 4 | Performance | Block 1 | Block 2 | Block 3 | Block 4 | Performance |
| NIH Chest X-ray (Performance in AUC) | | | | | | | | | | |
| MoCo | 0.81 | 0.80 | 0.57 | 0.41 | 59.2 | 0.72 | 0.75 | 0.66 | 0.39 | 69.8 |
| MoCo + MSE | 0.97 | 0.83 | 0.65 | **0.42** | 61.4 | **0.97** | 0.84 | 0.55 | **0.43** | 70.5 |
| MoCo + Barlow Twins | **0.99** | **0.98** | **0.76** | 0.38 | 62.9 | **0.97** | **0.92** | **0.79** | 0.41 | 70.0 |
| Diabetic Retinopathy (Performance in AUC) | | | | | | | | | | |
| MoCo | 0.87 | 0.80 | **0.51** | 0.19 | 88.1 | 0.81 | 0.69 | **0.50** | 0.14 | 92.4 |
| MoCo + MSE | 0.96 | 0.78 | 0.33 | **0.26** | 93.6 | **0.95** | 0.73 | 0.25 | 0.12 | 95.1 |
| MoCo + Barlow Twins | **0.98** | **0.83** | 0.58 | 0.24 | 92.5 | 0.93 | **0.75** | **0.50** | **0.15** | 94.0 |
| Breast Cancer Histopathology (Performance in F1-score) | | | | | | | | | | |
| MoCo | 0.50 | 0.55 | **0.98** | 0.16 | 82.8 | 0.57 | 0.48 | **0.97** | 0.19 | 82.8 |
| MoCo + MSE | **0.77** | **0.82** | 0.58 | **0.42** | 85.1 | 0.75 | **0.79** | 0.55 | 0.31 | 84.6 |
| MoCo + Barlow Twins | **0.77** | 0.74 | 0.54 | 0.36 | 84.4 | **0.76** | 0.76 | 0.58 | **0.38** | 84.5 |

- **Layer-wise Probing:** Our approach also learns more useful features earlier in the model.

| | Block 1 | Block 2 | Block 3 | Block 4 |
|---|---|---|---|---|
| NIH Chest X-ray (AUC (95% CI)) | | | | |
| MoCo | **58.8 (58.4-59.3)** | 59.5 (59.0-60.0) | 65.3 (64.8-65.8) | 74.4 (73.9-75.0) |
| MoCo + MSE | 57.6 (56.9-58.3) | **59.9 (59.4-60.4)** | **69.2 (68.70-69.7)** | **74.8 (74.2-75.4)** |
| MoCo + Barlow Twins | 56.6 (56.2-57.0) | 56.5 (56.1-56.9) | 64.2 (63.7-64.6) | 73.5 (72.9-74.0) |
| Diabetic Retinopathy (AUC (95% CI)) | | | | |
| MoCo | 68.1 (68.0-68.1) | 68.2 (68.2-68.3) | 69.2 (69.2-69.5) | 74.6 (74.5-74.7) |
| MoCo + MSE | **68.3 (68.2-68.3)** | **70.1 (70.0-70.1)** | **71.2 (71.1-71.3)** | **84.8 (84.6-85.0)** |
| MoCo + Barlow Twins | 67.2 (67.2-67.3) | 68.6 (68.5-68.7) | 69.9 (69.4-69.9) | 79.7 (79.6-79.7) |
| Breast Cancer Histopathology (F1-score (95% CI)) | | | | |
| MoCo | 80.9 (80.5-81.3) | 81.1 (80.8-81.5) | 81.1 (80.7-81.5) | 80.7 (80.4-81.1) |
| MoCo + MSE | 80.6 (80.2-81.0) | **81.3 (81.0-81.7)** | **82.7 (82.4-83.0)** | **82.5 (82.2-82.9)** |
| MoCo + Barlow Twins | **81.0 (90.7-81.4)** | 81.1 (80.7-81.5) | 82.2 (81.9-82.6) | 82.3 (82.0-82.7) |

References
1. Ciga, O., Martel, A. L., and Xu, T. Self supervised contrastive learning for digital histopathology.
2. Dehaene, O et al. Self-supervision closes the gap between weak & strong supervision in histology.
3. Sowrirajan, H et al. Mococxr:Moco pretraining improves representation and transferability of chest x-ray models.
4. Azizi, S., et al. Big self-supervised models advance medical image classification.