Contrastive Micrograph-Metadata Pre-Training

Henrik Eliasson[1,2*] and Angus Lothian[2]

[1]Center for Visualizing Catalytic Processes (VISION), Department of Physics, Technical University of Denmark, 2800 Kongens Lyngby, Denmark

[2]Stemson, stemson.ai

*haoel@dtu.dk

Abstract

Most machine learning workflows for electron microscopy do not generalize well beyond specific instrument settings. Conditioning networks on metadata could be a way to provide crucial information in the training process to realize truly robust applications. However, best practices for incorporating image metadata in AI for microscopy pipelines are not well established. In this hackathon contribution, we explore if it is possible to learn a joint embedding for HAADF-STEM images and a selection of important STEM parameters found in their paired metadata, with the hope that such a smooth learned latent space would be highly useful for conditioning downstream tasks. In more detail, we implement a version of CLIP (Contrastive Language-Image Pre-training) but replace the natural language captions with metadata vectors. After training, we find that the model can successfully separate metadata combinations that do not fit with the style of the presented HAADF image, and that matching metadata vectors align closely with the image representation in the learned space.

Methodology

A dataset was created from 2000 HAADF-STEM images acquired on the same 300kV Titan S/TEM microscope and with the same HAADF detector in the years 2022-2025. These images are recorded with a range of different acquisition parameters. The dataset consists of image-metadata pairs where the images are 224x224 random crops from the 2000 original images. We extract 10 such patches from each of the 2000 images, making the dataset contain 20000 224x224 images. For the metadata, we chose to focus on 14 parameters, {*Pixel Size*, *Dwell Time*, *C1 aperture*, *C2 aperture*, *Detector Gain*, *Detector Offset*, *Inner HAADF Collection Angle*, *Outer HAADF Collection Angle*, *Beam Convergence Angle*, *Camera Length*, *Extractor Voltage*, *Gun Lens Setting*, *Spot Index*, *Beam Current*}, so each of the 20000 images patches are paired with a 14 element long metadata vector. The 10 patches from the same original image naturally have the same metadata vector. All metadata were normalized between 0 and 1 based on the global min and max of that metadata, and all images were normalized by dividing by 65536. The dataset split was 90-10, keeping 10% of examples for validation, ensuring also that all 10 patches of an image are in the same set to avoid data leakage.

For our CLIP [1] variant which we call CMMP (Contrastive Micrograph-Metadata Pre-Training), we utilize a simple MLP as our metadata encoder, taking the 14 element vector and outputting a 128 element embedding. For our image encoder, we start from a pretrained ResNet18, but change the final layer to an *nn.Linear* layer that also gives us a 128 element embedding. Since the ResNet18 expects RGB input, we simply duplicate our greyscale channel 3 times. We use the Sigmoid loss [2] for contrastive training, with the modification that, because a batch may contain duplicates of the same metadata vector, we assign a target of +1 to all matching pairs rather than only to the diagonal of the target matrix. We train the CMMP model for 200 epochs with a batch size of 512 and initial temperature and bias of 20 and -6.24, respectively. Training was done in Google Colab with an A100 GPU.

Results

The first thing to check after training the model is to see if the cosine similarity between correctly matched image-metadata embeddings is on average larger than when incorrectly matched image-metadata embeddings are compared. We find that matching image-metadata pairs have a 0.4 larger cosine similarity on average compared to mismatched pairs (0.1754 compared to -0.2330). The model can now be used for tasks like retrieval and metadata prediction. Given a certain metadata vector, we can return images from our dataset with similar metadata, or given an image we can predict the metadata it was recorded with by looking at its nearest neighbors in the embedding space and taking a weighted average of their real metadata. For each of the 14 metadata elements, we achieve the following mean absolute percentage error using the 5 nearest neighbors in the embedding space: {68.28%,15.57%, 0.66%, 5.44%, 1.75%, 0.18%, 4.71%, 0.00%, 7.12%, 5.10%, 0.06%, 1.48%, 3.43%, 40.59%}. An interesting observation is that the prediction error for pixel size is very high which is reasonable considering there is very little scale information in small 224x224 patches. Gain and Offset prediction works particularly well, likely due to these parameters having a large effect on the actual pixel values of the images.

A potential use-case: Physics-Aware Denoising

Imagine a live denoiser that runs on your microscope and works even when you change settings like beam current, magnification, etc. Our smooth latent space could be ideal to condition an image-image translator to convert noisy low dose images to high signal ones. We can for instance encode the current acquisition metadata e_real, and encode a fictional metadata with higher electron dose e_target, the difference between these vectors in the latent space is e_transformation = e_target - e_real. We can now use a U-Net for image-image translation, conditioning it on both e_real and e_transformation via FiLM [3] or other. For the loss, we can encode the output image with our image encoder to e_output and calculate the loss as L2(e_output-e_target) ensuring that the style of the generated image is the style that would be expected for our chosen new metadata. With this we would have a relatively lightweight network with inference times low enough to run live on a S/TEM, showing the user a virtual “high-dose” image, aiding instrument operation while still keeping the actual dose low.

[1] A. Radford *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," *arXiv preprint arXiv:2103.00020*, 2021.

[2] X. Zhai *et al.*, "Sigmoid Loss for Language Image Pre-Training," *arXiv preprint arXiv:2303.15343*, 2023.

[3] E. Perez *et al.*, "FiLM: Visual Reasoning with a General Conditioning Layer," *arXiv preprint arXiv:1709.07871*, 2017.