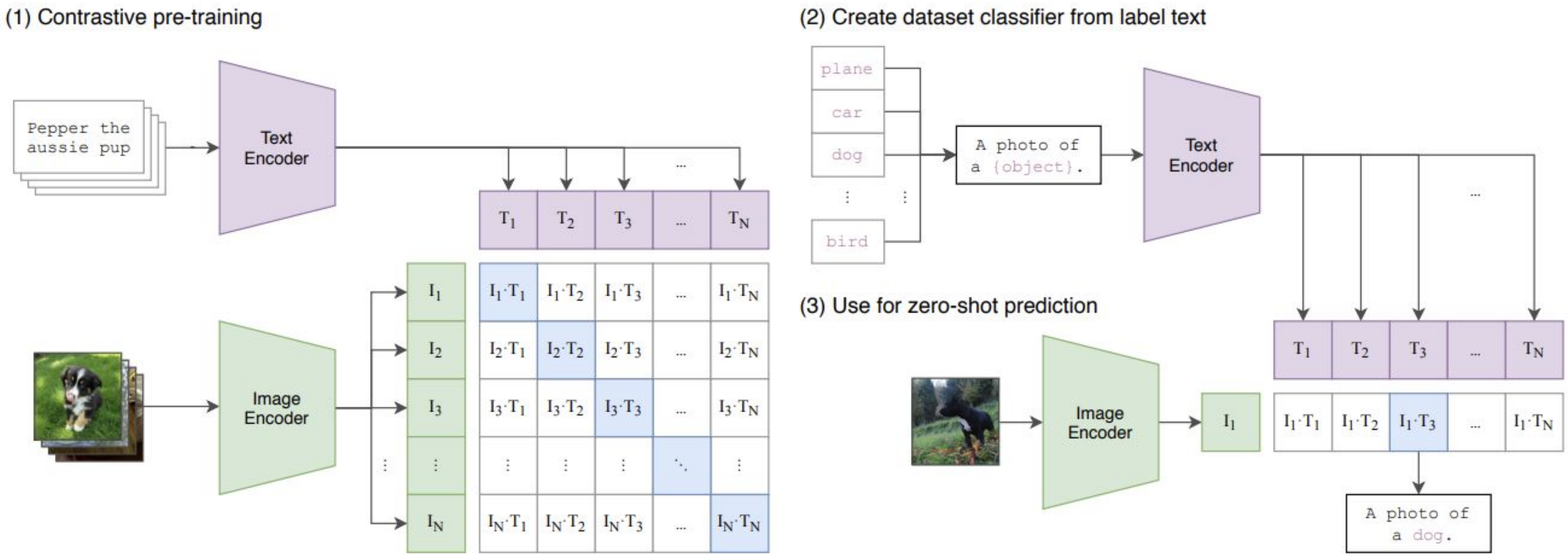
CLIP, MedCLIP, MedCLIP - SAM

CLIP



(1) Contrastive pre-training
Pepper the aussie pup
Text Encoder
Image Encoder
T1 T2 T3 ... TN
I1 I2 I3 ... IN
(2) Create dataset classifier from label text
plane
car
dog
bird
A photo of a {object}.
Text Encoder
(3) Use for zero-shot prediction
Image Encoder
A photo of a dog.

```
# image_encoder - ResNet or Vision Transformer
# text_encoder  - CBOW or Text Transformer
# I[n, h, w, c] - minibatch of aligned images
# T[n, l]       - minibatch of aligned texts
# W_i[d_i, d_e] - learned proj of image to embed
# W_t[d_t, d_e] - learned proj of text to embed
# t             - learned temperature parameter

# extract feature representations of each modality
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T)  #[n, d_t]

# joint multimodal embedding [n, d_e]
I_e = l2_normalize(np.dot(I_f, W_i), axis=1)
T_e = l2_normalize(np.dot(T_f, W_t), axis=1)

# scaled pairwise cosine similarities [n, n]
logits = np.dot(I_e, T_e.T) * np.exp(t)

# symmetric loss function
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss   = (loss_i + loss_t)/2
```

Testing:

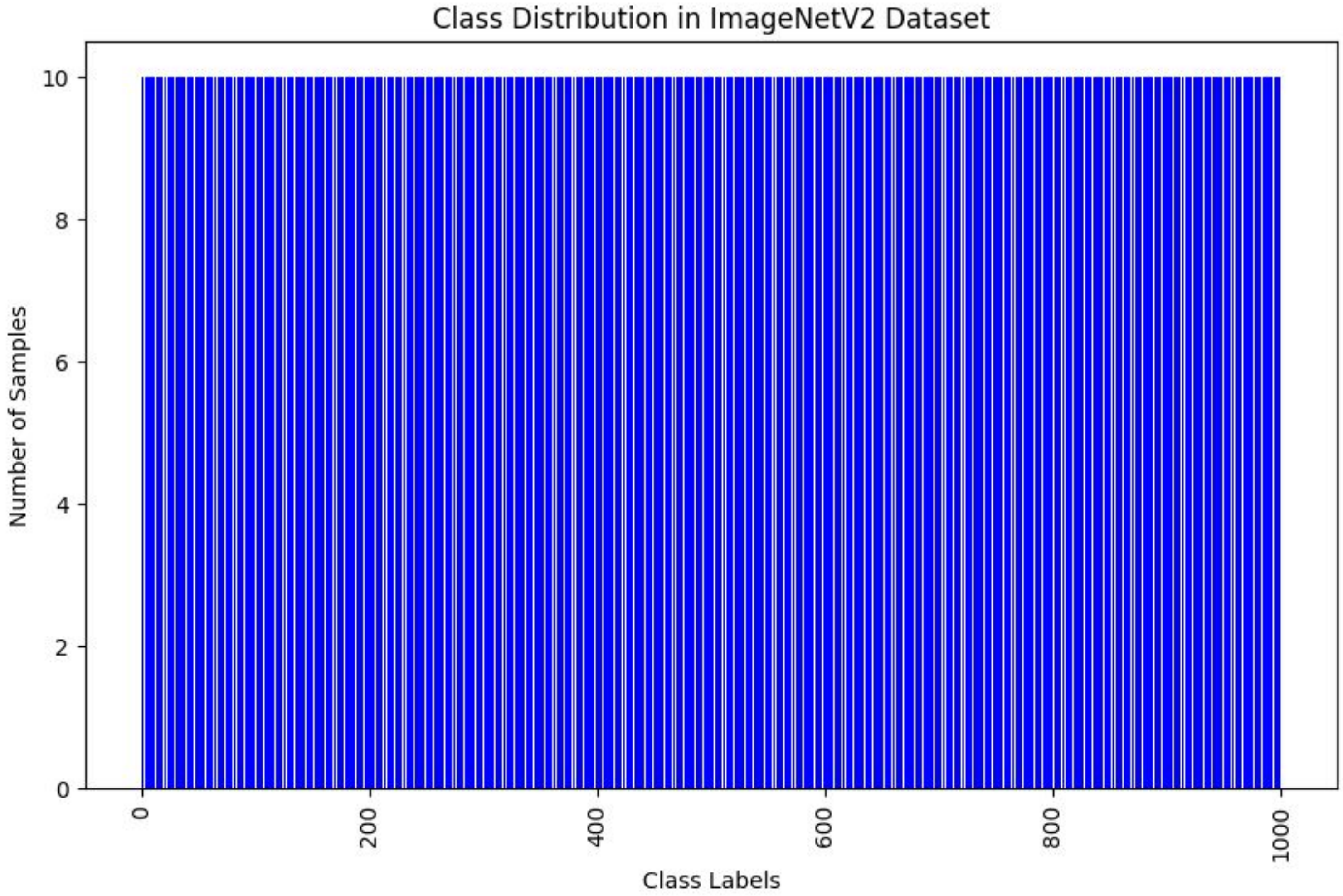
CLIP Model - ViT-B/32

Text - 1000 classes with 80 templates

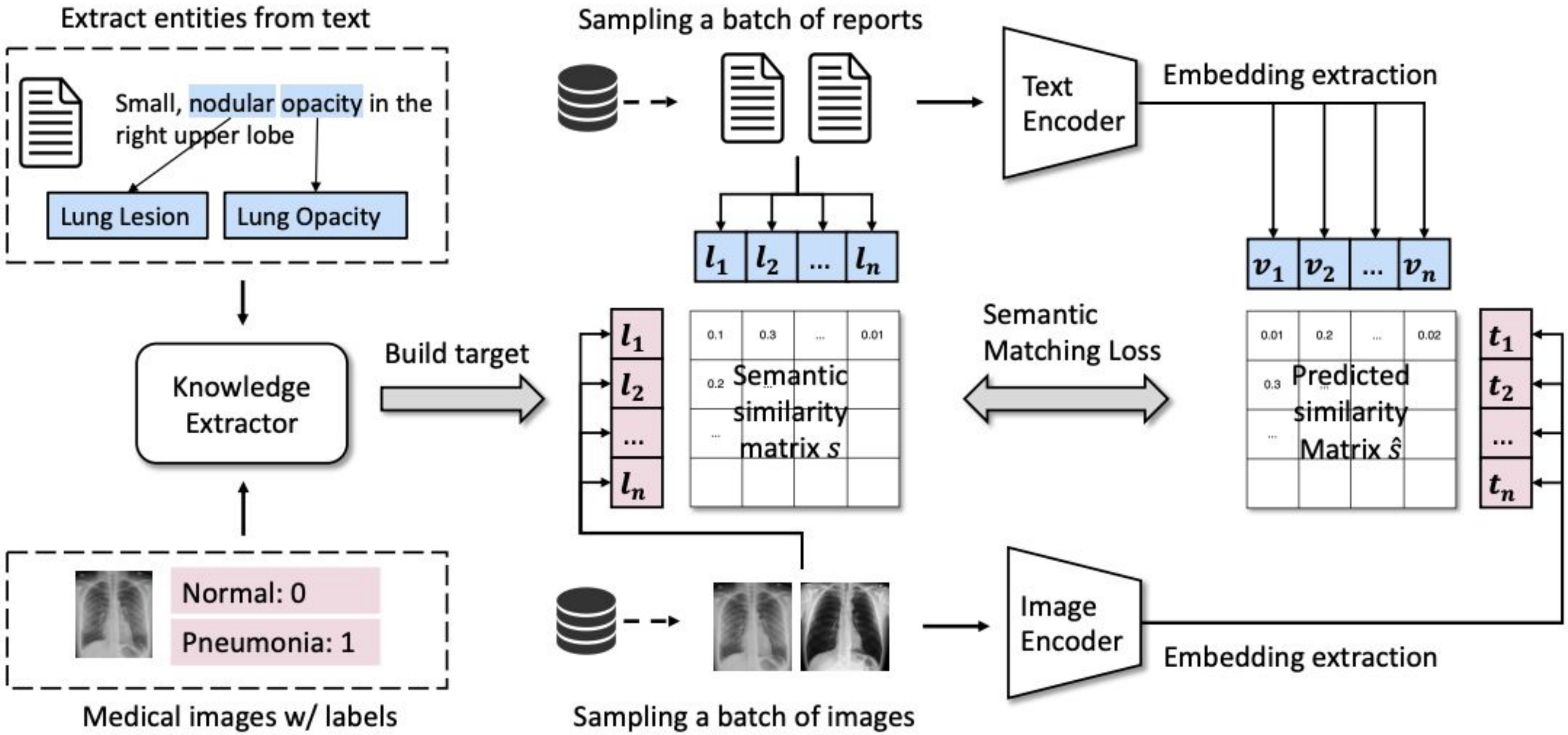
Images - ImageNetV2Dataset - 10,000 images

Top-1 accuracy: 55.95

Top-5 accuracy: 83.41


Class Distribution in ImageNetV2 Dataset
Number of Samples
0
2
4
6
8
10
0
200
400
600
800
1000
Class Labels

MedCLIP: Contrastive Learning from Unpaired Medical Images and Text



$$s = \frac{1_{\text{img}}^{\top} \cdot l_{\text{txt}}}{\|l_{\text{img}}\| \cdot \|l_{\text{txt}}\|}$$

$$y_{ij}^{v \to t} = \frac{\exp(s_{ij})}{\sum_{j=1}^{N_{\text{batch}}} \exp(s_{ij})}$$

$$y_{ji}^{t \to v} = \frac{\exp(s_{ji})}{\sum_{i=1}^{N_{\text{batch}}} \exp(s_{ji})}$$

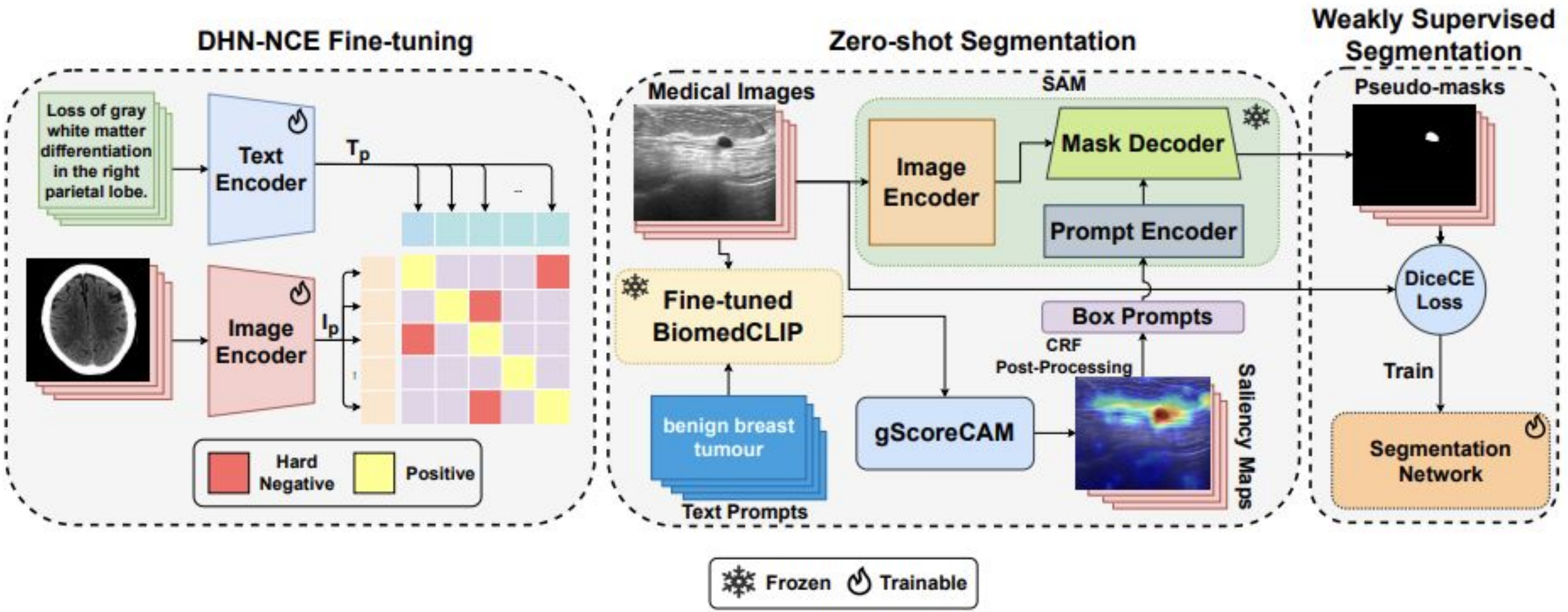
$$\hat{s}_{ij} = \tilde{v_i}^{\top} \cdot \tilde{t_j}$$

$$\hat{y}_{ij} = \frac{\exp(\hat{s}_{ij}/\tau)}{\sum_{i=1}^{N_{\text{batch}}} \exp(\hat{s}_{ij}/\tau)}$$

$$\mathcal{L}^{v \to l} = -\frac{1}{N_{\text{batch}}} \sum_{i=1}^{N_{\text{batch}}} \sum_{j=1}^{N_{\text{batch}}} y_{ij} \log \hat{y}_{ij}$$

$$\mathcal{L} = \frac{\mathcal{L}^{v \to l} + \mathcal{L}^{l \to v}}{2}$$

MedCLIP-SAM: Bridging Text and Image Towards Universal Medical Image Segmentation



$$\mathcal{L}^{v\to t} = -\sum_{i=1}^{B} \frac{\mathbf{I}_{p,i}\mathbf{T}_{p,i}^{\top}}{\tau} + \sum_{i=1}^{B} log\left(\sum_{j\neq i} e^{\mathbf{I}_{p,i}\mathbf{T}_{p,j}^{\top}/\tau}\mathcal{W}^{v\to t}_{\mathbf{I}_{p,i}\mathbf{T}_{p,j}}\right) \tag{1}$$

$$\mathcal{L}^{t\to v} = -\sum_{i=1}^{B} \frac{\mathbf{T}_{p,i}\mathbf{I}_{p,i}^{\top}}{\tau} + \sum_{i=1}^{B} log\left(\sum_{j\neq i} e^{\mathbf{T}_{p,i}\mathbf{I}_{p,j}^{\top}/\tau}\mathcal{W}^{t\to v}_{\mathbf{T}_{p,i}\mathbf{I}_{p,j}}\right) \tag{2}$$

$$\mathcal{L}_{DHN-NCE} = \mathcal{L}^{v\to t} + \mathcal{L}^{t\to v} \tag{3}$$

where B is the batch size, τ is the temperature parameter, and the hardness weighting formulas are as follows:

$$\mathcal{W}^{v\to t}_{\mathbf{I}_{p,i}\mathbf{T}_{p,j}} = (B-1) \times \frac{e^{\beta_1 \mathbf{I}_{p,i}\mathbf{T}_{p,j}/\tau}}{\sum_{k\neq i} e^{\beta_1 \mathbf{I}_{p,i}\mathbf{T}_{p,k}/\tau}} \tag{4}$$

$$\mathcal{W}^{t\to v}_{\mathbf{T}_{p,i}\mathbf{I}_{p,j}} = (B-1) \times \frac{e^{\beta_2 \mathbf{T}_{p,i}\mathbf{I}_{p,j}/\tau}}{\sum_{k\neq i} e^{\beta_2 \mathbf{T}_{p,i}\mathbf{I}_{p,k}/\tau}} \tag{5}$$

Thank you.