CS480/680: Introduction to Machine Learning

Lec 16: Contrastive Learning

Yaoliang Yu

UNIVERSITY OF WATERLOO | FACULTY OF MATHEMATICS
DAVID R. CHERITON SCHOOL
OF COMPUTER SCIENCE

March 21, 2024

Self-supervised Pre-training

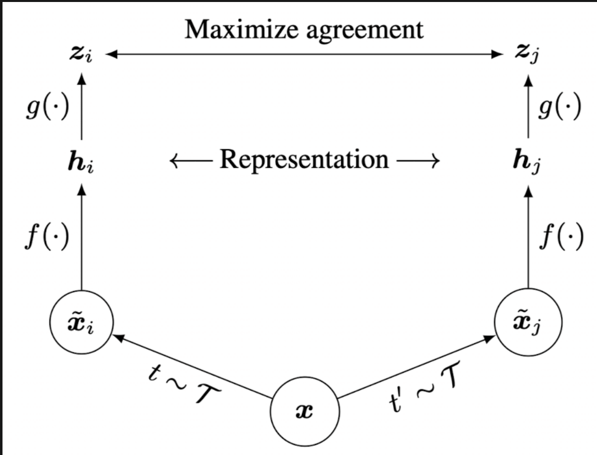
- Self-supervised pre-training of a language model by predicting the next token:

$$\min_{\Theta} \ \hat{\mathbb{E}} - \log \prod_{j=1}^{m} p(\mathbf{x}_j | \mathbf{x}_1, \ldots, \mathbf{x}_{j-1}; \Theta)$$

- Self-supervised pre-training of a generative model by predicting the (next) pixel?
  - works ok for representation learning but not as competitive
  - perhaps the task of predicting the next pixel is too difficulty and unnecessary

SimCLR: Simple Contrastive Learning of visual Representation

- Stochastic data augmentation
- Encoder network f to learn representation (e.g., ResNet)
- Projection head g for self-supervised learning (e.g., simple 2-layer MLP)
- Contrastive loss to pull positive pairs and push negative pairs
  - anything but "my twin" is negative



T. Chen et al. "A Simple Framework for Contrastive Learning of Visual Representations". In: *Proceedings of the 37th International Conference on Machine Learning*. 2020, pp. 1597–1607.

Algorithm 1: SimCLR

Input: batch size b, constant τ, encoder f, projection g, augmentation $\mathcal{T}$

Output: only the encoder f

for $t = 0, 1, \ldots$ **do**

  sample a minibatch B_t with size b `// large` b `for many negative pairs`

  for $i = 1, \ldots, b$ **do**

    draw two augmentations $T, T' \sim \mathcal{T}$

    $\mathbf{z}_{2i-1} \leftarrow g(f(T\,(B_t[i])))$

    $\mathbf{z}_{2i} \quad \leftarrow g(f(T'(B_t[i])))$

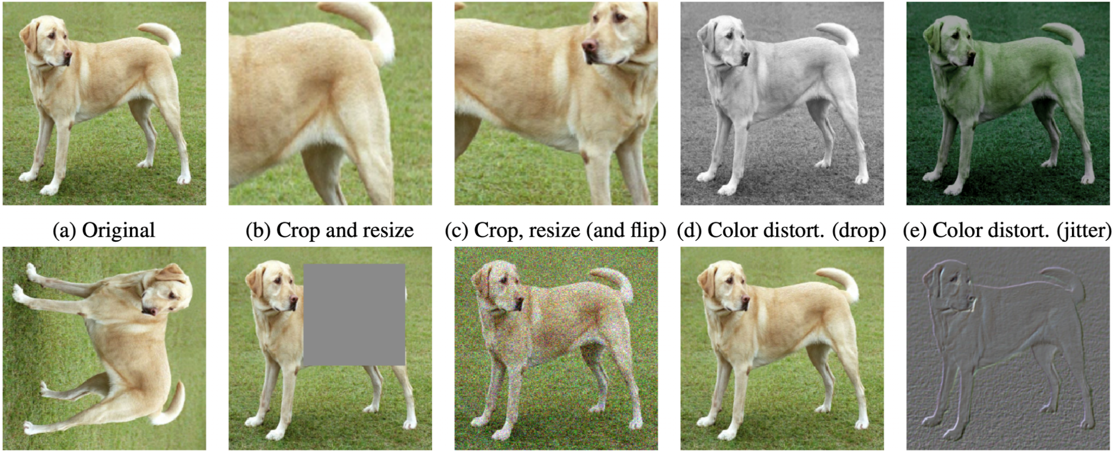
  for $i = 1, \ldots, 2b$ **do**

    for $j = 1, \ldots, 2b$ **do**

      $s_{ij} \leftarrow \mathtt{sim}(\mathbf{z}_i, \mathbf{z}_j)$ `// e.g.,` $\mathtt{sim}(\mathbf{z}_i, \mathbf{z}_j) = \frac{\mathbf{z}_i}{\|\mathbf{z}_i\|_2} \frac{\mathbf{z}_j}{\|\mathbf{z}_j\|_2}$, `cosine similarity`

  $$\min_{f,g} \quad \frac{1}{2b} \sum_{i=1}^{b} \left[-\log \frac{\exp(s_{2i-1,2i}/\tau)}{\sum_{k \neq 2i-1} \exp(s_{2i-1,k}/\tau)} - \log \frac{\exp(s_{2i,2i-1}/\tau)}{\sum_{k \neq 2i} \exp(s_{2i,k}/\tau)} \right]$$

Data Augmentation



(a) Original (b) Crop and resize (c) Crop, resize (and flip) (d) Color distort. (drop) (e) Color distort. (jitter)

(f) Rotate $\{90°, 180°, 270°\}$ (g) Cutout (h) Gaussian noise (i) Gaussian blur (j) Sobel filtering

Scaling

Linear probing (evaluation): fix the representation and train a linear classifier (e.g., logistic regression) on top

Self-supervised pre-training benefits from

- stronger data augmentation
- bigger (wider and deeper) models
- longer (more epochs) training
- larger batch size (more neg pairs)

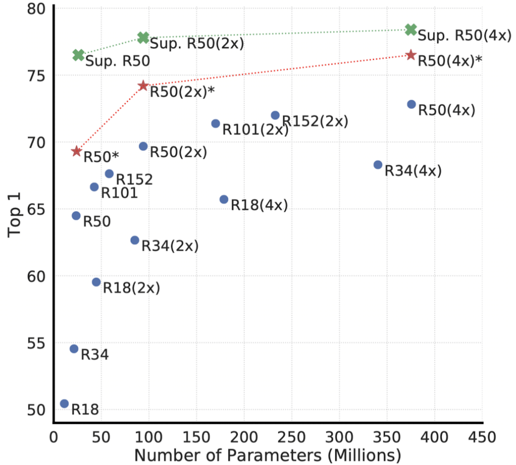


Figure 7. Linear evaluation of models with varied depth and width. Models in blue dots are ours trained for 100 epochs, models in red stars are ours trained for 1000 epochs, and models in green crosses are supervised ResNets trained for 90 epochs[7] (He et al., 2016).

Projection

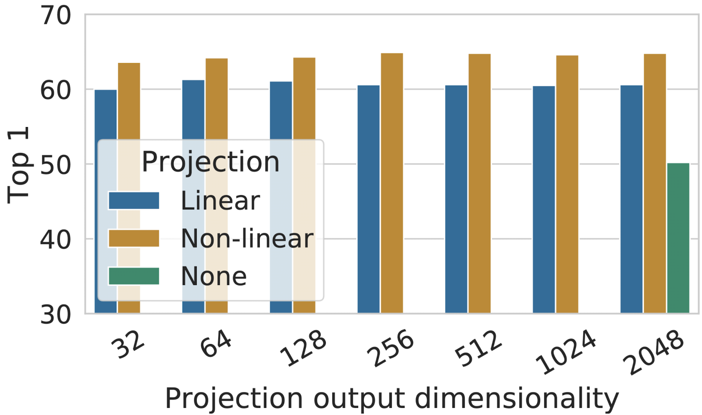


Figure 8. Linear evaluation of representations with different projection heads $g(\cdot)$ and various dimensions of $\boldsymbol{z} = g(\boldsymbol{h})$. The representation $\boldsymbol{h}$ (before projection) is 2048-dimensional here.

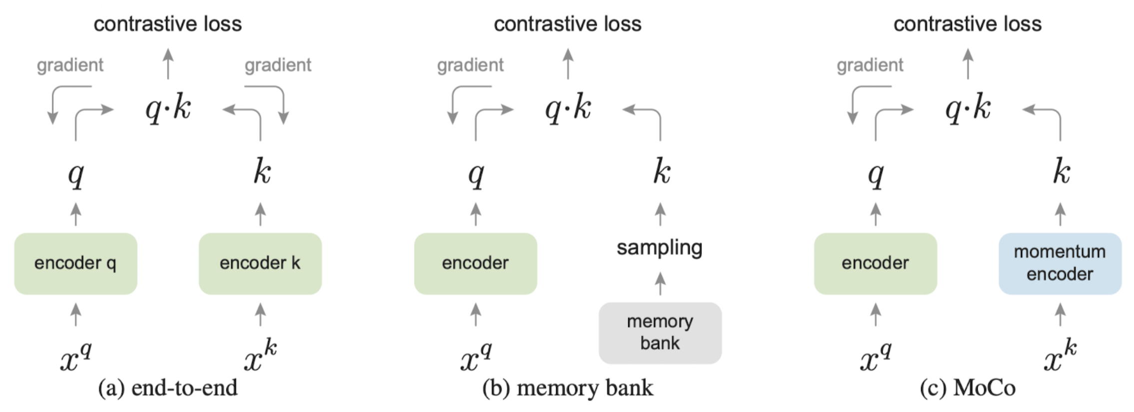
Comparison: ResNet50(4x) vs. ResNet50

Both SimCLR and Supervised are trained on ImageNet to extract feature representation

	Food	CIFAR10	CIFAR100	Birdsnap	SUN397	Cars	Aircraft	VOC2007	DTD	Pets	Caltech-101	Flowers
Linear evaluation:												
SimCLR (ours)	**76.9**	**95.3**	80.2	48.4	**65.9**	60.0	61.2	**84.2**	**78.9**	89.2	**93.9**	**95.0**
Supervised	75.2	**95.7**	**81.2**	**56.4**	64.9	**68.8**	**63.8**	83.8	**78.7**	**92.3**	**94.1**	94.2
Fine-tuned:												
SimCLR (ours)	**89.4**	**98.6**	**89.0**	**78.2**	**68.1**	**92.1**	**87.0**	**86.6**	**77.8**	92.1	**94.1**	97.6
Supervised	88.7	98.3	**88.7**	**77.8**	67.0	91.4	**88.0**	86.5	**78.8**	**93.2**	**94.2**	**98.0**
Random init	88.3	96.0	81.9	**77.0**	53.7	91.3	84.8	69.4	64.1	82.7	72.5	92.5

	Food	CIFAR10	CIFAR100	Birdsnap	SUN397	Cars	Aircraft	VOC2007	DTD	Pets	Caltech-101	Flowers
Linear evaluation:												
SimCLR (ours)	68.4	90.6	71.6	37.4	58.8	50.3	50.3	80.5	**74.5**	83.6	90.3	91.2
Supervised	**72.3**	**93.6**	**78.3**	**53.7**	**61.9**	**66.7**	**61.0**	**82.8**	**74.9**	**91.5**	**94.5**	**94.7**
Fine-tuned:												
SimCLR (ours)	**88.2**	**97.7**	**85.9**	**75.9**	63.5	91.3	**88.1**	84.1	**73.2**	89.2	92.1	97.0
Supervised	**88.3**	**97.5**	**86.4**	**75.8**	**64.3**	**92.1**	86.0	**85.0**	**74.6**	**92.1**	**93.3**	**97.6**
Random init	86.9	95.9	80.2	**76.1**	53.6	91.4	85.9	67.3	64.8	81.5	72.6	92.0

Moment Contrastive (MoCo)



K. He et al. "Momentum contrast for unsupervised visual representation learning". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 9729–9738.

Algorithm 2: MoCo

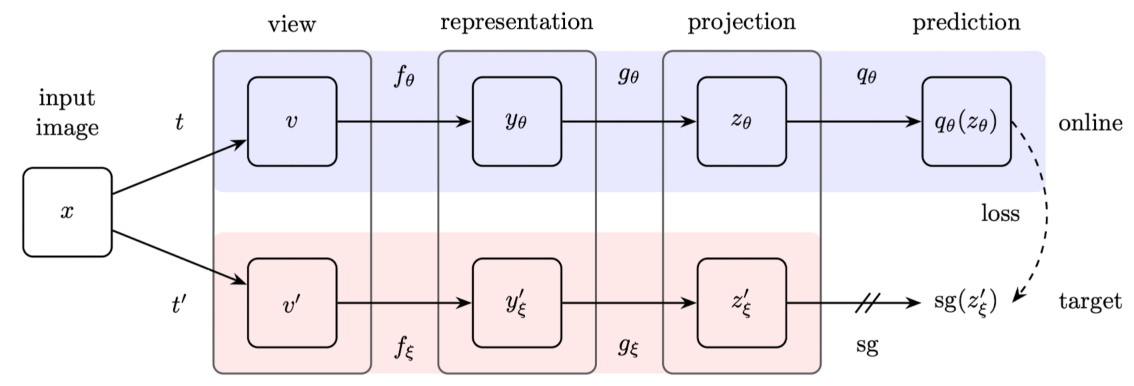
Input: batch size b, constant τ, d-dim queue for s keys, momentum $\lambda \in [0, 1]$

$f_k = f_q$ `// initialize query and key encoders`

for $t = 0, 1, \ldots$ **do**

- sample a minibatch B_t with size b
- $q = f_q(\texttt{aug}(B_t))$ `// randomly augmented query`
- $k = f_k(\texttt{aug}(B_t))$ `// randomly augmented key; detach k: no gradient to keys`
- $l_{\text{pos}} = \texttt{bmm}(q.\text{view}(b, 1, d),\ k.\text{view}(b, d, 1))$ `// batch matrix multiplication`
- $l_{\text{neg}} = \texttt{mm}(q.\text{view}(b, d),\ queue.\text{view}(d, s))$ `// matrix multiplication`
- $logits = \texttt{cat}([l_{\text{pos}},\ l_{\text{neg}}], \text{dim} = 1)$ `// ` $b \times (s+1)$
- $labels = \texttt{zeros}(b)$ `// positives are the 0-th`
- $loss = \texttt{CrossEntropyLoss}(logits/\tau,\ labels)$ `// contrastive loss`
- $loss.\text{backward}()$
- $\text{update}(f_q)$ `// SGD update on the query network`
- $f_k = \lambda f_k + (1 - \lambda) f_q$ `// momentum update of the key network; e.g.,` $\lambda = 0.999$
- $\texttt{enqueue}(queue,\ k)$ `// enqueue the current minibatch`
- $\texttt{dequeue}(queue)$ `// dequeue the earliest minibatch`

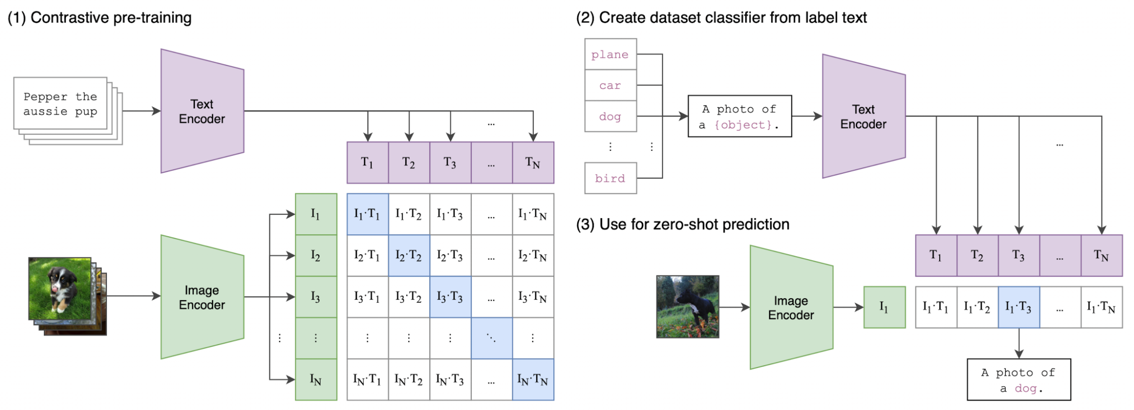
Bootstrapping Your Own Latent (BYOL)



$$\min_\theta \ \mathrm{sim}(q_\theta(z_\theta),\ z'_\xi), \qquad \xi \leftarrow \lambda\xi + (1-\lambda)\theta$$

J.-B. Grill et al. "Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning". In: *Advances in Neural Information Processing Systems 33*. 2020.

Contrastive Language Image Pre-training (CLIP)



A. Radford et al. "Learning Transferable Visual Models From Natural Language Supervision". In: *Proceedings of the 38th International Conference on Machine Learning*. 2021.

Algorithm 3: CLIP

Input: batch size b, temperature τ, imageEncoder, textEncoder, W_i, W_t

for $t = 0, 1, \ldots$ **do**

  sample a minibatch of images I_t and texts T_t with size b

  $I_f = \texttt{imageEncoder}(I_t)$ // $b \times d_i$

  $T_f = \texttt{textEncoder}(T_t)$ // $b \times d_t$

  $I_e = l2\_\mathrm{normalize}(I_f * W_i)$ // $W_i \in \mathbb{R}^{d_i \times d_e}$, learned

  $T_e = l2\_\mathrm{normalize}(T_f * W_t)$ // $W_t \in \mathbb{R}^{d_t \times d_e}$, learned

  $logits = I_e * T_e^\top * \exp(\tau)$ // $b \times b$

  $labels = [0, 1, \ldots, b-1]$

  $loss_i = \texttt{CrossEntropyLoss}(logits,\ labels, \mathrm{axis} = 0)$ // image loss

  $loss_t = \texttt{CrossEntropyLoss}(logits,\ labels, \mathrm{axis} = 1)$ // text loss

  $loss = (loss_i + loss_t)/2$ // loss to be minimized

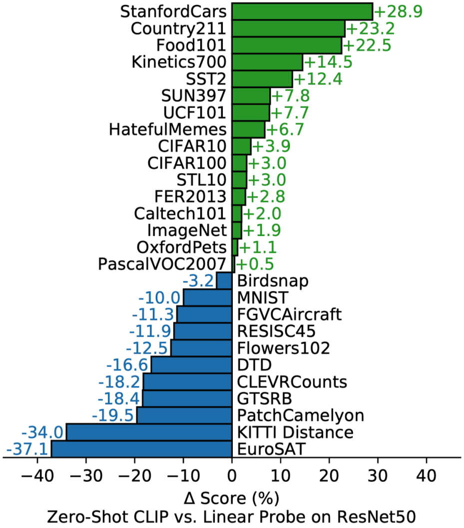
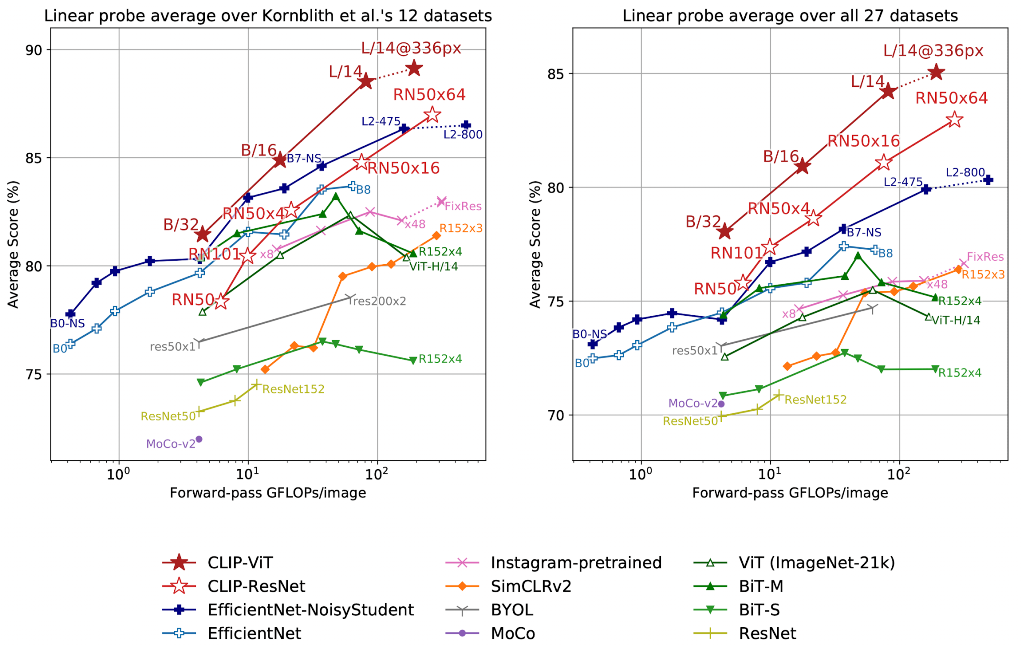
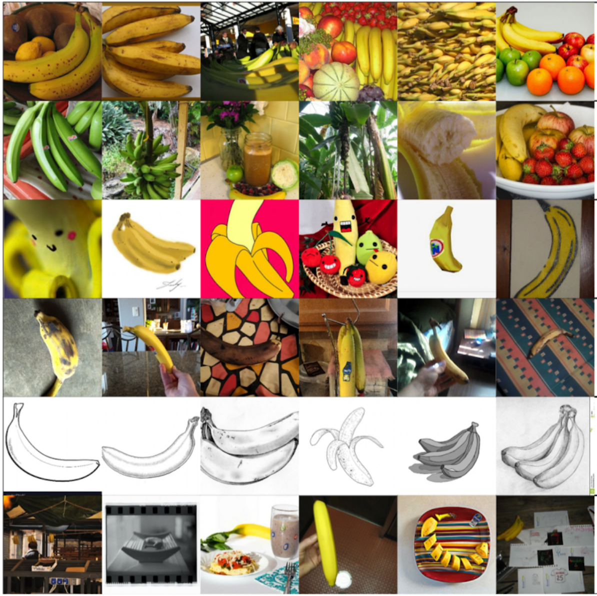


Figure 4. **Zero-shot CLIP is competitive with a fully supervised baseline.** Across a 27 dataset eval suite, a zero-shot CLIP classifier outperforms a fully supervised linear classifier fitted on ResNet50 features on 16 datasets, including ImageNet.


Linear probe average over Kornblith et al.'s 12 datasets
Linear probe average over all 27 datasets
Average Score (%)
Forward-pass GFLOPs/image
L/14@336px
L/14
RN50x64
RN50x16
RN50x4
RN101
RN50
B/16
B/32
L2-475
L2-800
B7-NS
B8
B0-NS
B0
FixRes
x48
x8
R152x3
R152x4
ViT-H/14
res200x2
res50x1
ResNet152
ResNet50
MoCo-v2
CLIP-ViT
CLIP-ResNet
EfficientNet-NoisyStudent
EfficientNet
Instagram-pretrained
SimCLRv2
BYOL
MoCo
ViT (ImageNet-21k)
BiT-M
BiT-S
ResNet

Dataset Examples



	ImageNet ResNet101	Zero-Shot CLIP	Δ Score
ImageNet	**76.2**	**76.2**	0%
ImageNetV2	64.3	**70.1**	+5.8%
ImageNet-R	37.7	**88.9**	+51.2%
ObjectNet	32.6	**72.3**	+39.7%
ImageNet Sketch	25.2	**60.2**	+35.0%
ImageNet-A	2.7	**77.1**	+74.4%

Food101

guacamole (90.1%) Ranked 1 out of 101 labels

✓ a photo of **guacamole**, a type of food.

× a photo of **ceviche**, a type of food.

× a photo of **edamame**, a type of food.

× a photo of **tuna tartare**, a type of food.

× a photo of **hummus**, a type of food.

SUN397

television studio (90.2%) Ranked 1 out of 397 labels

✓ a photo of a **television studio.**

× a photo of a **podium indoor.**

× a photo of a **conference room.**

× a photo of a **lecture room.**

× a photo of a **control room.**

EuroSAT

annual crop land (46.5%) Ranked 4 out of 10 labels

✕ a centered satellite photo of **permanent crop land.**

✕ a centered satellite photo of **pasture land.**

✕ a centered satellite photo of **highway or road.**

✓ a centered satellite photo of **annual crop land.**

✕ a centered satellite photo of **brushland or shrublan**

Youtube-BB

airplane, person (89.0%) Ranked 1 out of 23 labels

✓ a photo of a **airplane.**

× a photo of a **bird.**

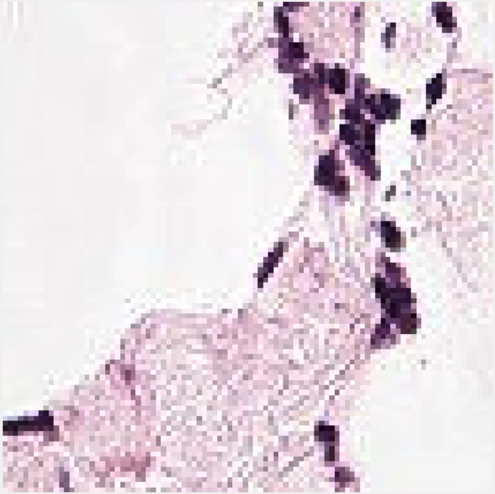
× a photo of a **bear.**

× a photo of a **giraffe.**

× a photo of a **car.**

PatchCamelyon (PCam)

healthy lymph node tissue (77.2%) Ranked 2 out of 2 labels



× this is a photo of **lymph node tumor tissue**

✓ this is a photo of **healthy lymph node tissue**