DSC291: Machine Learning with Few Labels

Self-supervised Learning

Zhiting Hu

Lecture 10, February 1, 2023

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

Recap: Self-Supervised Learning

- Given an observed data instance $\boldsymbol{t}$
- One could derive various supervision signals based on the structure of the data
- By applying a "split" function that artificially partition $\boldsymbol{t}$ into two parts
  - $(\boldsymbol{x}, \boldsymbol{y}) = split(\boldsymbol{t})$
  - sometimes split in a stochastic way
- Treat $\boldsymbol{x}$ as the input and $\boldsymbol{y}$ as the output
- Train a model $p_\theta(\boldsymbol{y}|\boldsymbol{x})$

Self-Supervised Learning (SSL): Examples

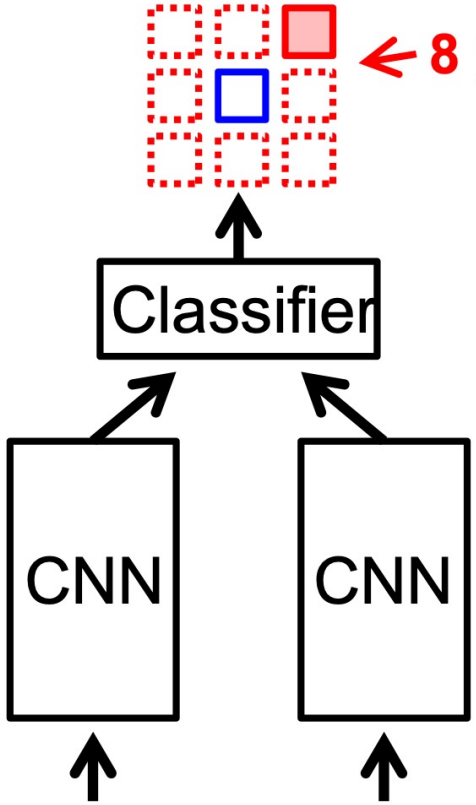
- SSL from text
  - Language models: next-word prediction (GPT-3)
  - Learning contextual text representations: masked language model (BERT)

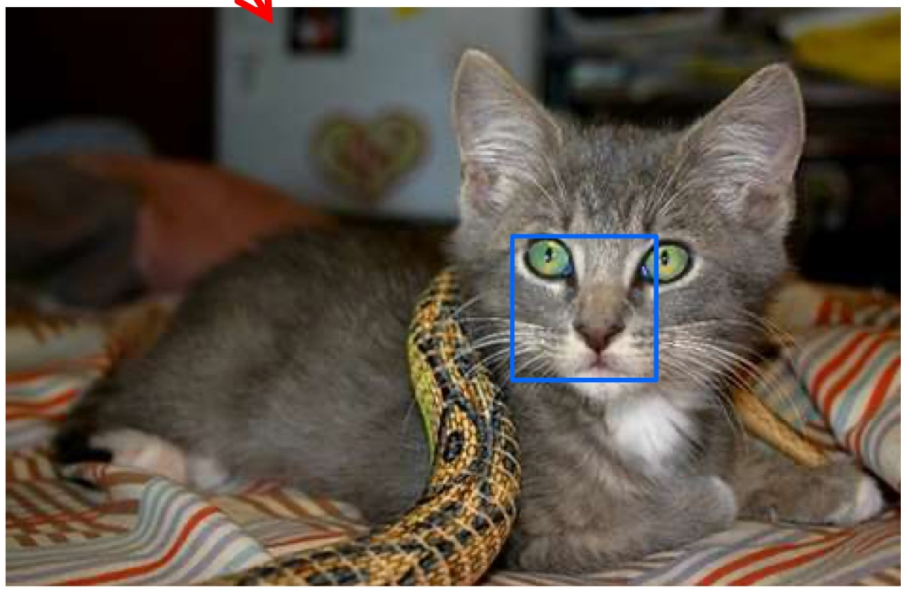
- SSL from images

- SSL from videos

SSL from Images, EX (I): relative positioning

Train network to predict relative position of two regions in the same image



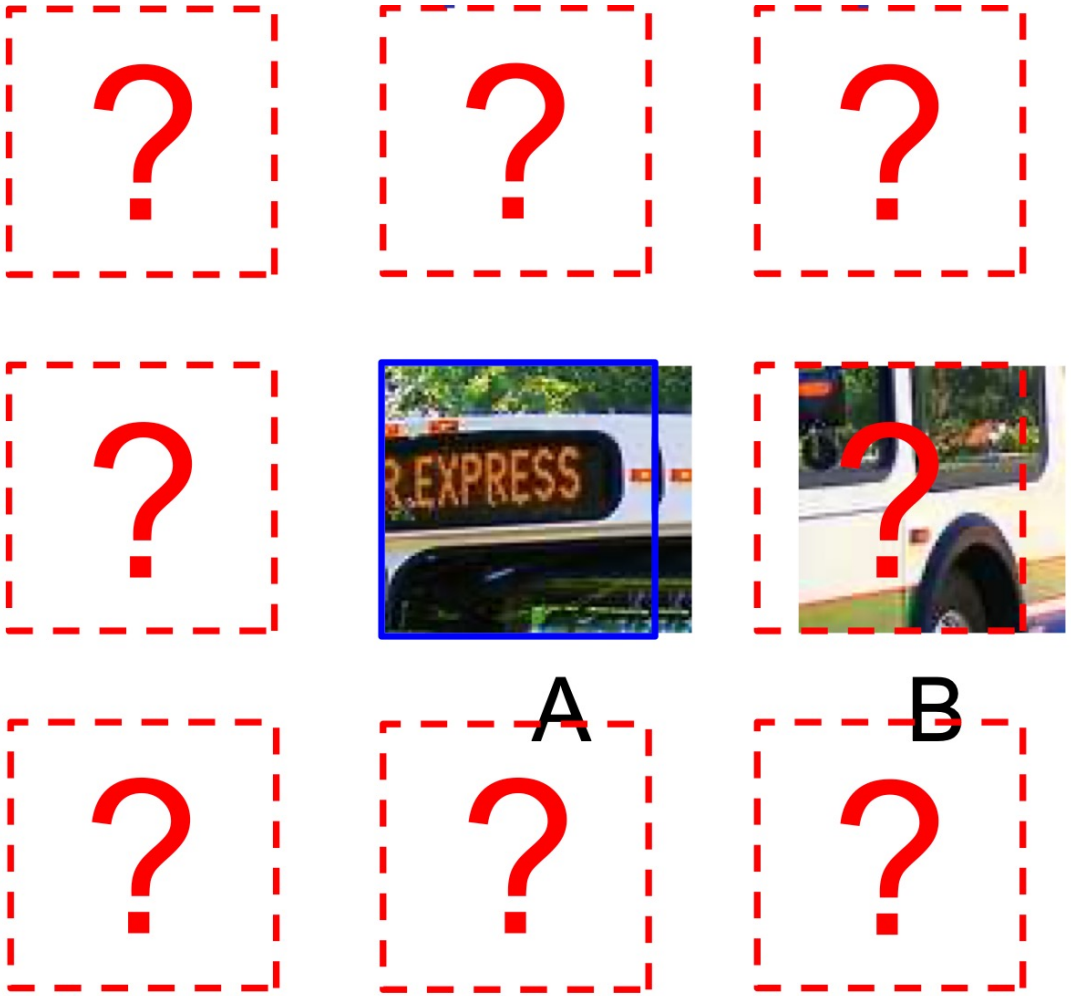


Randomly Sample Patch
Sample Second Patch

Unsupervised visual representation learning by context prediction,
Carl Doersch, Abhinav Gupta, Alexei A. Efros, ICCV 2015

[Courtesy: Zisserman “Self-supervised Learning”]

SSL from Images, EX (I): relative positioning



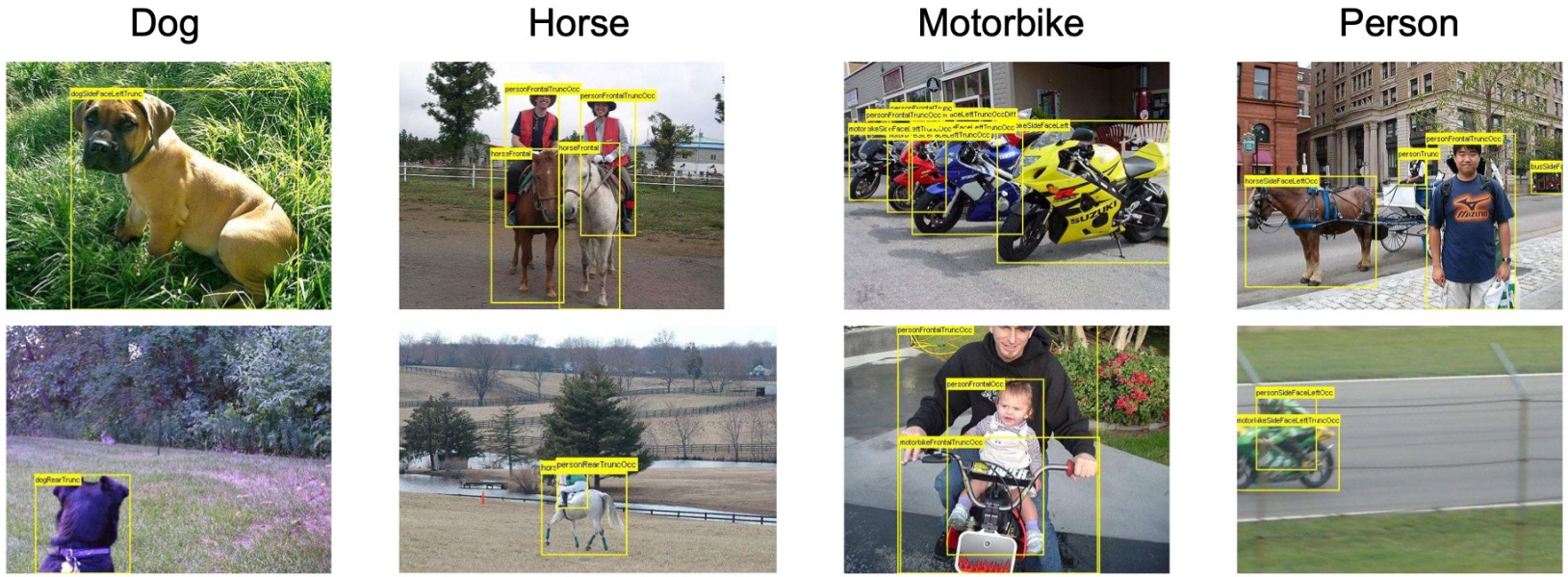
Unsupervised visual representation learning by context prediction,
Carl Doersch, Abhinav Gupta, Alexei A. Efros, ICCV 2015

[Courtesy: Zisserman “Self-supervised Learning”]

SSL from Images, EX (I): relative positioning

Evaluation: PASCAL VOC Detection

- 20 object classes (car, bicycle, person, horse …)
- Predict the bounding boxes of all objects of a given class in an image (if any)



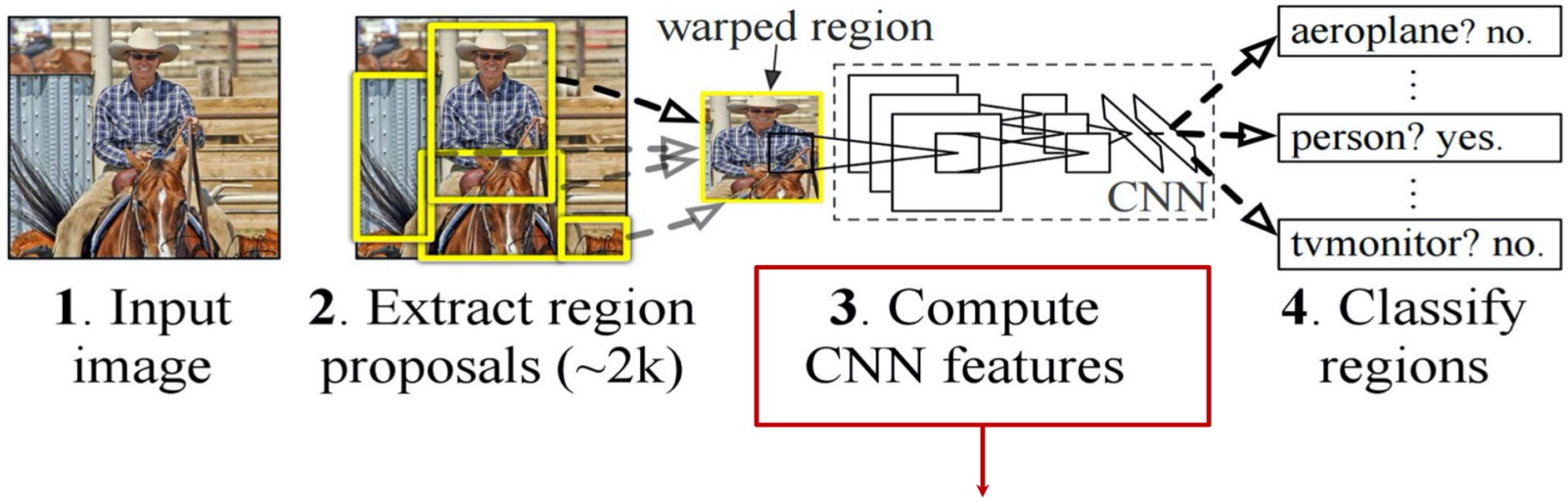
[Courtesy: Zisserman “Self-supervised Learning”]

SSL from Images, EX (I): relative positioning

Evaluation: PASCAL VOC Detection

- Pre-train CNN using self-supervision (no labels)
- Train CNN for detection in R-CNN object category detection pipeline

R-CNN



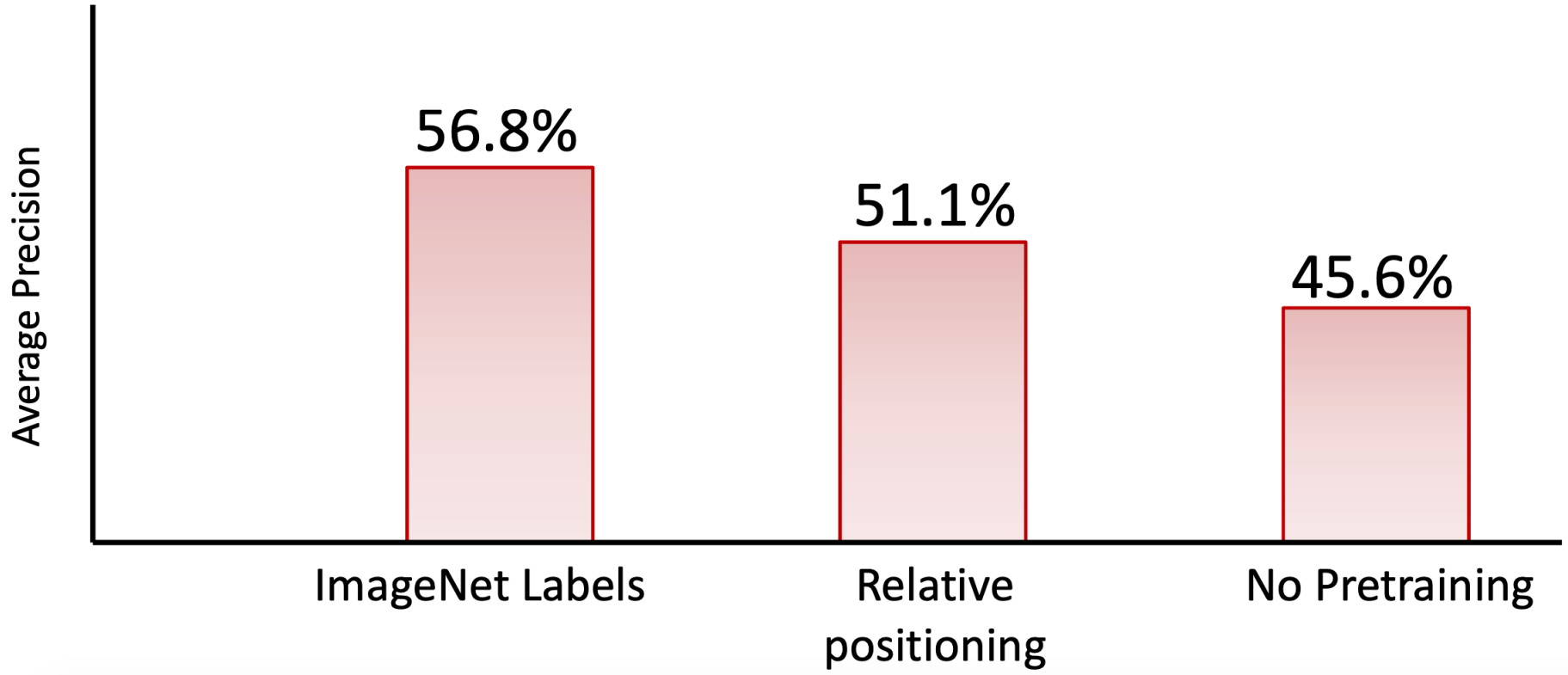
Pre-train on relative-position task, w/o labels

[Girshick et al. 2014]

[Courtesy: Zisserman "Self-supervised Learning"]

SSL from Images, EX (I): relative positioning

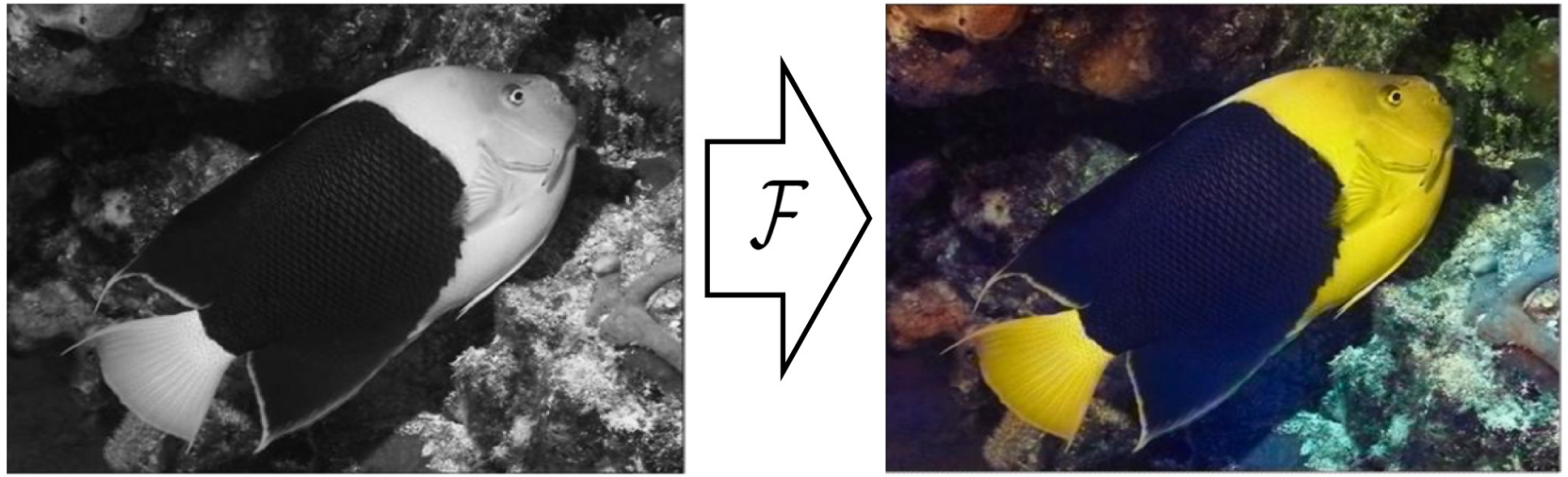
Evaluation: PASCAL VOC Detection



[Courtesy: Zisserman “Self-supervised Learning”]

SSL from Images, EX (II): colorization

Train network to predict pixel colour from a monochrome input

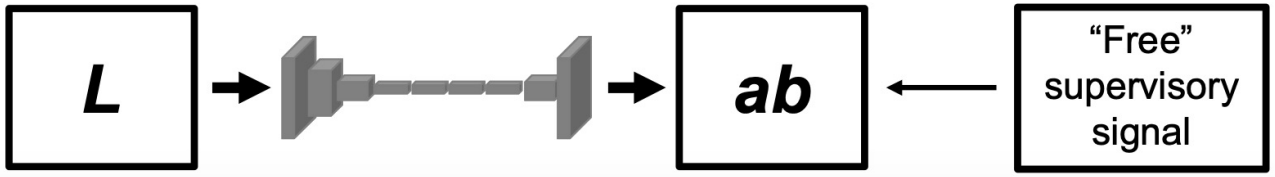


Grayscale image: *L* channel

$$\mathbf{X} \in \mathbb{R}^{H \times W \times 1}$$

Concatenate (L,ab)

$$(\mathbf{X}, \hat{\mathbf{Y}})$$



[Courtesy: Zisserman "Self-supervised Learning"]

Colorful Image Colorization, Zhang et al., ECCV 2016

SSL from Images, EX (II): colorization

Train network to predict pixel colour from a monochrome input

[Courtesy: Zisserman “Self-supervised Learning”]

Colorful Image Colorization, Zhang et al., ECCV 2016

SSL from Images, EX (III): exemplar networks

- Exemplar Networks (Dosovitskiy et al., 2014)
- Perturb/distort image patches, e.g. by cropping and affine transformations
- Train to classify these exemplars as same class

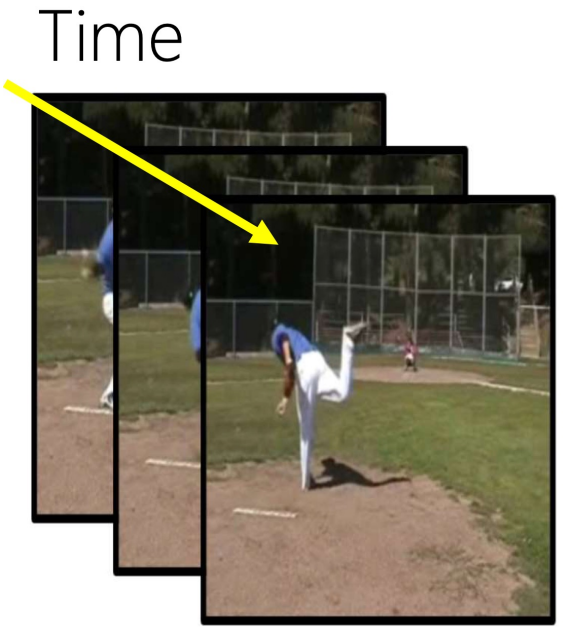
[Courtesy: Zisserman “Self-supervised Learning”]

SSL from Videos

Three example tasks:

- Video sequence order
  - Sequential Verification: Is this a valid sequence?





"Sequence" of data

[Courtesy: Zisserman "Self-supervised Learning"]

Wei et al., 2018 Arrow of Time

SSL from Videos

Three example tasks:

- Video sequence order
  - Sequential Verification: Is this a valid sequence?
- Video direction
  - Predict if video playing forwards or backwards

[Courtesy: Zisserman “Self-supervised Learning”]

Wei et al., 2018 Arrow of Time

SSL from Videos

Three example tasks:

- Video sequence order
  - Sequential Verification: Is this a valid sequence?
- Video direction
  - Predict if video playing forwards or backwards
- Video tracking
  - Given a color video, colorize all frames of a gray scale version using a reference frame

[Courtesy: Zisserman “Self-supervised Learning”]

Vondrick et al., 2018

Key Takeaways

- Self supervision learning
  - Predicting any part of the observations given any available information
  - The prediction task forces models to learn semantic representations
  - Massive/unlimited data supervisions
- SSL for text:
  - Language models: next word prediction
  - BERT text representations: masked language model (MLM)
- SSL for images/videos:
  - Various ways of defining the prediction task

Contrastive Learning

Contrastive learning

- Take a data example x, sample a "positive" sample x_{pos} and "negative" samples x_{neg} in some way
- Then try fit a scoring model such that

$$score(x, x_{pos}) > score(x, x_{neg})$$

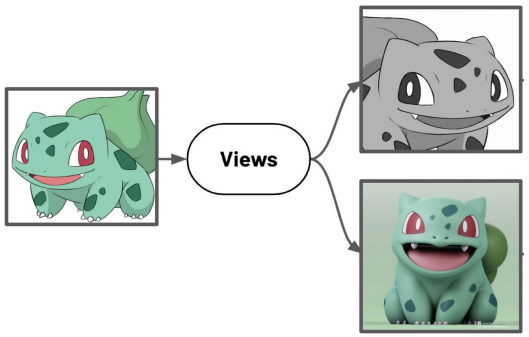
Credit: [CVPR 2021 Tutorial] Leave Those Nets Alone: Advances in Self-Supervised Learning

Contrastive learning

- Take a data example x, sample a "positive" sample x_{pos} and "negative" samples x_{neg} in some way

"positive" sample:

- Data of the same labels
- Data of the same pseudo-labels
- Augmented/distorted version of x
- Data that captures the same target from different views



"negative" sample:

- Randomly sampled data
- Hard negative sample mining

Contrastive learning

- Take a data example x, sample a "positive" sample x_{pos} and "negative" samples x_{neg} in some way
- Then try fit a scoring model such that

$$score(x, x_{pos}) > score(x, x_{neg})$$

Credit: [CVPR 2021 Tutorial] Leave Those Nets Alone: Advances in Self-Supervised Learning

Contrastive learning: Ex 1

Learning a similarity metric discriminatively

Sample a pair of images and compute their distance:

$$D_i = \|x, x_i\|_2$$

If **positive** sample:

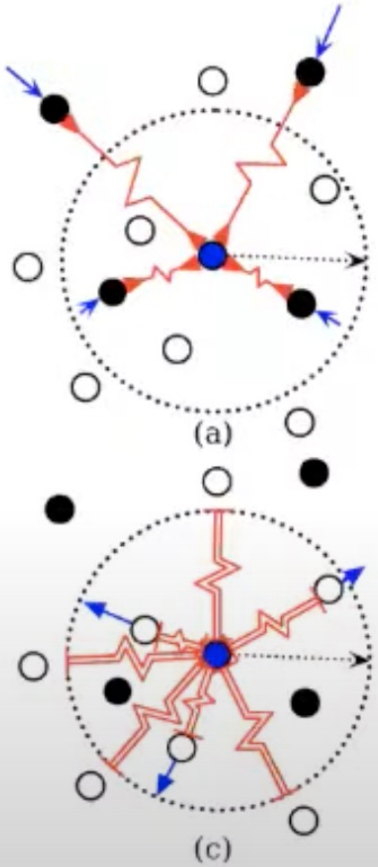
$$L_i = D_i^2$$

x pos

If **negative** sample:

$$L_i = \max\left(0, \epsilon - D_i\right)^2$$

x neg



[Chopra et al., 2005; Hadsell et al., 2006]

Credit: [CVPR 2021 Tutorial] Leave Those Nets Alone: Advances in Self-Supervised Learning

Common contrastive learning functions

- Contrastive loss (Chopra et al. 2005)
- Triplet loss (Schroff et al. 2015; FaceNet)
- Lifted structured loss (Song et al. 2015)
- Multi-class n-pair loss (Sohn 2016)
- Noise contrastive estimation (“NCE”; Gutmann & Hyvarinen 2010)
- InfoNCE (van den Oord, et al. 2018)
- Soft-nearest neighbors loss (Salakhutdinov & Hinton 2007, Frosst et al. 2019)

[Courtesy: Weng & Kim, NeurIPS 2021 tutorial]

Contrastive learning: Ex 2

- SimCSE ("Simple Contrastive learning of Sentence Embeddings"; Gao et al. 2021)
  - Predict a sentence from itself with only dropout noise
  - One sentence gets two different versions of dropout augmentations

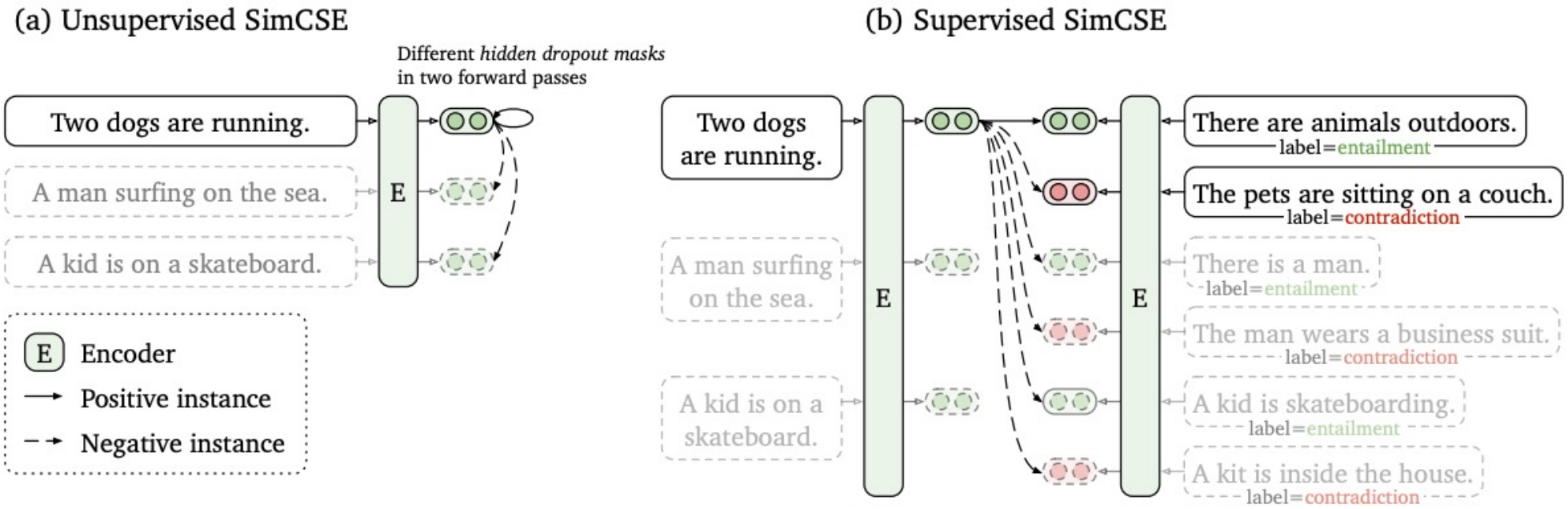
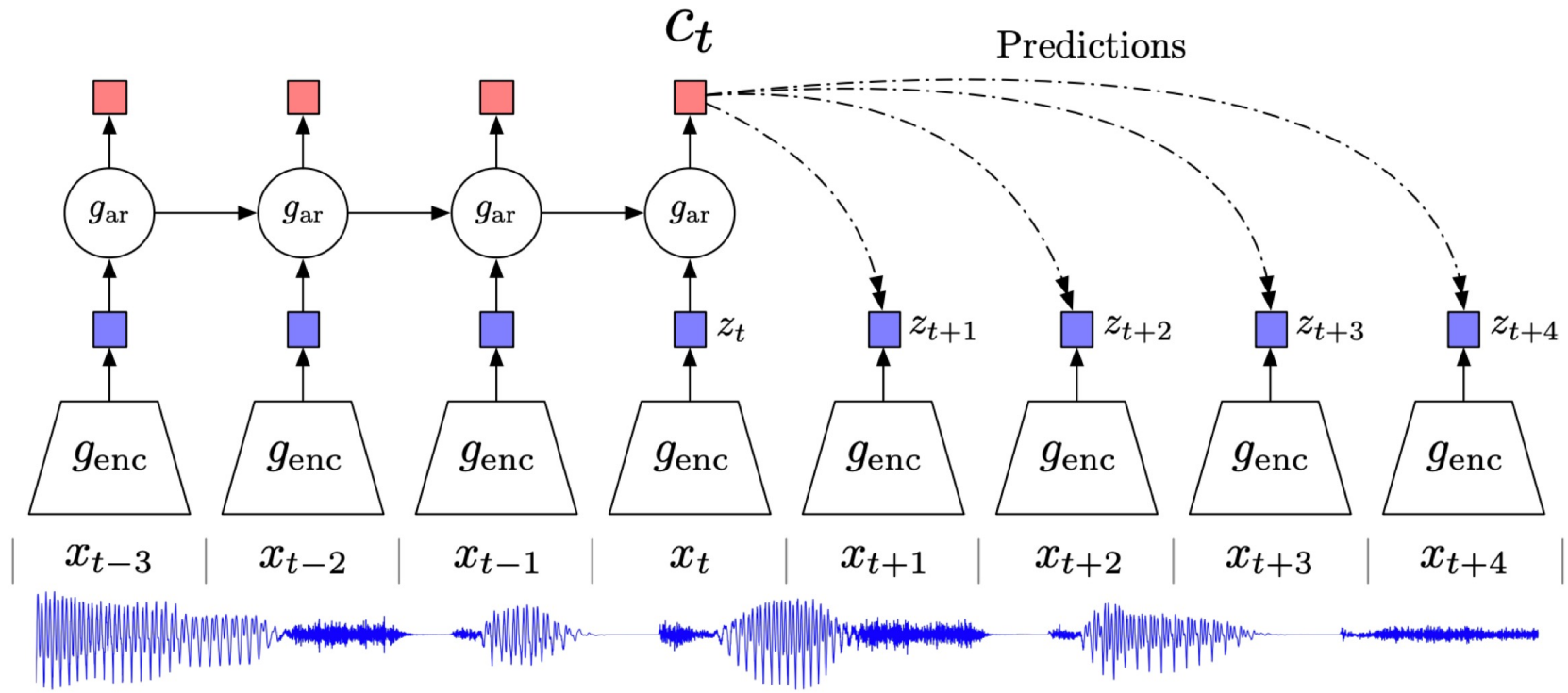


Figure 1: (a) Unsupervised SimCSE predicts the input sentence itself from in-batch negatives, with different hidden dropout masks applied. (b) Supervised SimCSE leverages the NLI datasets and takes the entailment (premise-hypothesis) pairs as positives, and contradiction pairs as well as other in-batch instances as negatives.

Contrastive learning: Ex 3 - InfoNCE

- The CPC model
  - c_t: context representation from history
  - x_{t+k} (or z_{t+k}): future target



[van den Oord et al., “Representation Learning with Contrastive Predictive Coding”]

InfoNCE loss

- Define scoring function $f_k > 0$
- The InfoNCE loss:
  - Given $X = \{$ one positive sample from $p(x_{t+k} | c_t)$, $N - 1$ negative samples from the negative sampling distribution $p(x_{t+k}) \}$

$$\mathcal{L}_{\mathrm{N}} = - \mathop{\mathbb{E}}_{X} \left[\log \frac{f_k(x_{t+k}, c_t)}{\sum_{x_j \in X} f_k(x_j, c_t)} \right]$$

- InfoNCE is interesting because it's effectively maximizing the mutual information between c_t and x_{t+k}

[van den Oord et al., "Representation Learning with Contrastive Predictive Coding"]

Mutual Information (MI)

- How much is our uncertainty about x reduced by knowing c ?

$$I(x;c) = \sum_{x,c} p(x,c) \log \frac{p(x,c)}{p(x)p(c)} = \sum_{x,c} p(x,c) \log \frac{p(x|c)}{p(x)}$$

$$= H(x) + H(c) - H(x,c)$$

$$= H(x) - H(x|c)$$

$$= KL(p(x,c) \,||\, p(x)p(c))$$

[van den Oord et al., “Representation Learning with Contrastive Predictive Coding”]

Minimizing InfoNCE <=> Maximzing MI

- InfoNCE loss

$$\mathcal{L}_{\mathrm{N}} = -\mathop{\mathbb{E}}_{X}\left[\log \frac{f_k(x_{t+k}, c_t)}{\sum_{x_j \in X} f_k(x_j, c_t)}\right]$$

- The loss is optimized when

$$f_k(x_{t+k}, c_t) \propto \frac{p(x_{t+k}|c_t)}{p(x_{t+k})}$$

  - Proof:

$$p(sample\ i\ is\ positive|X, c_t) = \frac{p(x_i|c_t)\prod_{l\neq i} p(x_l)}{\sum_{j=1}^{N} p(x_j|c_t)\prod_{l\neq j} p(x_l)}$$

$$= \frac{\frac{p(x_i|c_t)}{p(x_i)}}{\sum_{j=1}^{N} \frac{p(x_j|c_t)}{p(x_j)}}.$$

[van den Oord et al., “Representation Learning with Contrastive Predictive Coding”]

- How does this loss maximize the mutual information?

$$\mathcal{L}_{\mathrm{N}} = -\underset{X}{\mathbb{E}}\left[\log \frac{f_k(x_{t+k}, c_t)}{\sum_{x_j \in X} f_k(x_j, c_t)}\right]$$

[van den Oord et al., “Representation Learning with Contrastive Predictive Coding”]

- How does this loss maximize the mutual information?

$$\mathcal{L}_{\mathrm{N}} = -\mathop{\mathbb{E}}_{X}\left[\log \frac{f_k(x_{t+k}, c_t)}{\sum_{x_j \in X} f_k(x_j, c_t)}\right]$$

$$\mathcal{L}_{\mathrm{N}}^{\mathrm{opt}} = -\mathop{\mathbb{E}}_{X}\log\left[\frac{\frac{p(x_{t+k}|c_t)}{p(x_{t+k})}}{\frac{p(x_{t+k}|c_t)}{p(x_{t+k})} + \sum_{x_j \in X_{\mathrm{neg}}} \frac{p(x_j|c_t)}{p(x_j)}}\right]$$

Use proportionality condition

[van den Oord et al., “Representation Learning with Contrastive Predictive Coding”]

- How does this loss maximize the mutual information?

$$\mathcal{L}_{\mathrm{N}} = -\mathop{\mathbb{E}}_{X}\left[\log \frac{f_k(x_{t+k}, c_t)}{\sum_{x_j \in X} f_k(x_j, c_t)}\right]$$

$$\mathcal{L}_{\mathrm{N}}^{\mathrm{opt}} = -\mathop{\mathbb{E}}_{X} \log\left[\frac{\frac{p(x_{t+k}|c_t)}{p(x_{t+k})}}{\frac{p(x_{t+k}|c_t)}{p(x_{t+k})} + \sum_{x_j \in X_{\mathrm{neg}}} \frac{p(x_j|c_t)}{p(x_j)}}\right]$$

$$= \mathop{\mathbb{E}}_{X} \log\left[1 + \frac{p(x_{t+k})}{p(x_{t+k}|c_t)} \sum_{x_j \in X_{\mathrm{neg}}} \frac{p(x_j|c_t)}{p(x_j)}\right]$$

Take -ve inside log

[van den Oord et al., “Representation Learning with Contrastive Predictive Coding”]

- How does this loss maximize the mutual information?

$$\mathcal{L}_{\mathrm{N}} = -\underset{X}{\mathbb{E}}\left[\log \frac{f_k(x_{t+k}, c_t)}{\sum_{x_j \in X} f_k(x_j, c_t)}\right]$$

$$\begin{aligned}
\mathcal{L}_{\mathrm{N}}^{\mathrm{opt}} &= -\underset{X}{\mathbb{E}} \log \left[\frac{\frac{p(x_{t+k}|c_t)}{p(x_{t+k})}}{\frac{p(x_{t+k}|c_t)}{p(x_{t+k})} + \sum_{x_j \in X_{\mathrm{neg}}} \frac{p(x_j|c_t)}{p(x_j)}}\right] \\
&= \underset{X}{\mathbb{E}} \log \left[1 + \frac{p(x_{t+k})}{p(x_{t+k}|c_t)} \sum_{x_j \in X_{\mathrm{neg}}} \frac{p(x_j|c_t)}{p(x_j)}\right] \\
&\approx \underset{X}{\mathbb{E}} \log \left[1 + \frac{p(x_{t+k})}{p(x_{t+k}|c_t)} (N-1) \underset{x_j}{\mathbb{E}} \frac{p(x_j|c_t)}{p(x_j)}\right]
\end{aligned}$$

This approximation becomes more accurate as N increases, so it is preferable to use large negative samples

[van den Oord et al., “Representation Learning with Contrastive Predictive Coding”]

- How does this loss maximize the mutual information?

$$\mathcal{L}_{\mathrm{N}} = -\mathop{\mathbb{E}}_{X}\left[\log \frac{f_k(x_{t+k}, c_t)}{\sum_{x_j \in X} f_k(x_j, c_t)}\right]$$

$$\mathcal{L}_{\mathrm{N}}^{\mathrm{opt}} = -\mathop{\mathbb{E}}_{X}\log\left[\frac{\frac{p(x_{t+k}|c_t)}{p(x_{t+k})}}{\frac{p(x_{t+k}|c_t)}{p(x_{t+k})} + \sum_{x_j \in X_{\mathrm{neg}}} \frac{p(x_j|c_t)}{p(x_j)}}\right]$$

$$= \mathop{\mathbb{E}}_{X}\log\left[1 + \frac{p(x_{t+k})}{p(x_{t+k}|c_t)} \sum_{x_j \in X_{\mathrm{neg}}} \frac{p(x_j|c_t)}{p(x_j)}\right]$$

$$\approx \mathop{\mathbb{E}}_{X}\log\left[1 + \frac{p(x_{t+k})}{p(x_{t+k}|c_t)}(N-1)\mathop{\mathbb{E}}_{x_j}\frac{p(x_j|c_t)}{p(x_j)}\right] \quad = 1$$

$$= \mathop{\mathbb{E}}_{X}\log\left[1 + \frac{p(x_{t+k})}{p(x_{t+k}|c_t)}(N-1)\right]$$

[van den Oord et al., “Representation Learning with Contrastive Predictive Coding”]

- How does this loss maximize the mutual information?

$$\mathcal{L}_{\mathrm{N}} = -\mathop{\mathbb{E}}_{X}\left[\log \frac{f_k(x_{t+k}, c_t)}{\sum_{x_j \in X} f_k(x_j, c_t)}\right]$$

$$\begin{aligned}
\mathcal{L}_{\mathrm{N}}^{\mathrm{opt}} &= -\mathop{\mathbb{E}}_{X} \log\left[\frac{\frac{p(x_{t+k}|c_t)}{p(x_{t+k})}}{\frac{p(x_{t+k}|c_t)}{p(x_{t+k})} + \sum_{x_j \in X_{\mathrm{neg}}} \frac{p(x_j|c_t)}{p(x_j)}}\right] \\
&= \mathop{\mathbb{E}}_{X} \log\left[1 + \frac{p(x_{t+k})}{p(x_{t+k}|c_t)} \sum_{x_j \in X_{\mathrm{neg}}} \frac{p(x_j|c_t)}{p(x_j)}\right] \\
&\approx \mathop{\mathbb{E}}_{X} \log\left[1 + \frac{p(x_{t+k})}{p(x_{t+k}|c_t)}(N-1) \mathop{\mathbb{E}}_{x_j} \frac{p(x_j|c_t)}{p(x_j)}\right] \\
&= \mathop{\mathbb{E}}_{X} \log\left[1 + \frac{p(x_{t+k})}{p(x_{t+k}|c_t)}(N-1)\right] \\
&\geq \mathop{\mathbb{E}}_{X} \log\left[\frac{p(x_{t+k})}{p(x_{t+k}|c_t)} N\right] \\
&= -I(x_{t+k}, c_t) + \log(N),
\end{aligned}$$

[van den Oord et al., “Representation Learning with Contrastive Predictive Coding”]

- How does this loss maximize the mutual information?

$$\mathcal{L}_{\mathrm{N}} = -\underset{X}{\mathbb{E}}\left[\log \frac{f_k(x_{t+k}, c_t)}{\sum_{x_j \in X} f_k(x_j, c_t)}\right]$$

$$I(x_{t+k}, c_t) \geq \log(N) - \mathcal{L}_{\mathrm{N}}$$

[van den Oord et al., “Representation Learning with Contrastive Predictive Coding”]

Key Takeaways: Contrastive learning

- Contrastive learning is a way of doing self-supervised learning
- Positive samples, negative samples
- Mutual information

$$I(x;c) = \sum_{x,c} p(x,c) \log \frac{p(x,c)}{p(x)p(c)} = \sum_{x,c} p(x,c) \log \frac{p(x|c)}{p(x)}$$

$$= H(x) + H(c) - H(x,c)$$

$$= H(x) + H(x|c)$$

$$= KL(p(x,c) \,||\, p(x)p(c))$$

  - InfoNCE ⇔ MI

Questions?