DSC291: Machine Learning with Few Labels

Contrastive Learning
Data Manipulation

Zhiting Hu
Lecture 9, April 29, 2025

UC San Diego
**HALICIOĞLU DATA SCIENCE INSTITUTE**

Logistics

- Homework 1 released
  - Due: May 23, Friday

Outline

- Contrastive learning

- Paper presentation:
  - Andrii Dovhaniuk, Wendi Tan: "Machine learning center-specific models"

Recap: Self-Supervised Learning: Examples

► **Predict any part of the input from any other part.**

► **Predict the future from the past.**

► **Predict the future from the recent past.**

► **Predict the past from the present.**

► **Predict the top from the bottom.**

► **Predict the occluded from the visible**

► **Pretend there is a part of the input you don't know and predict that.**
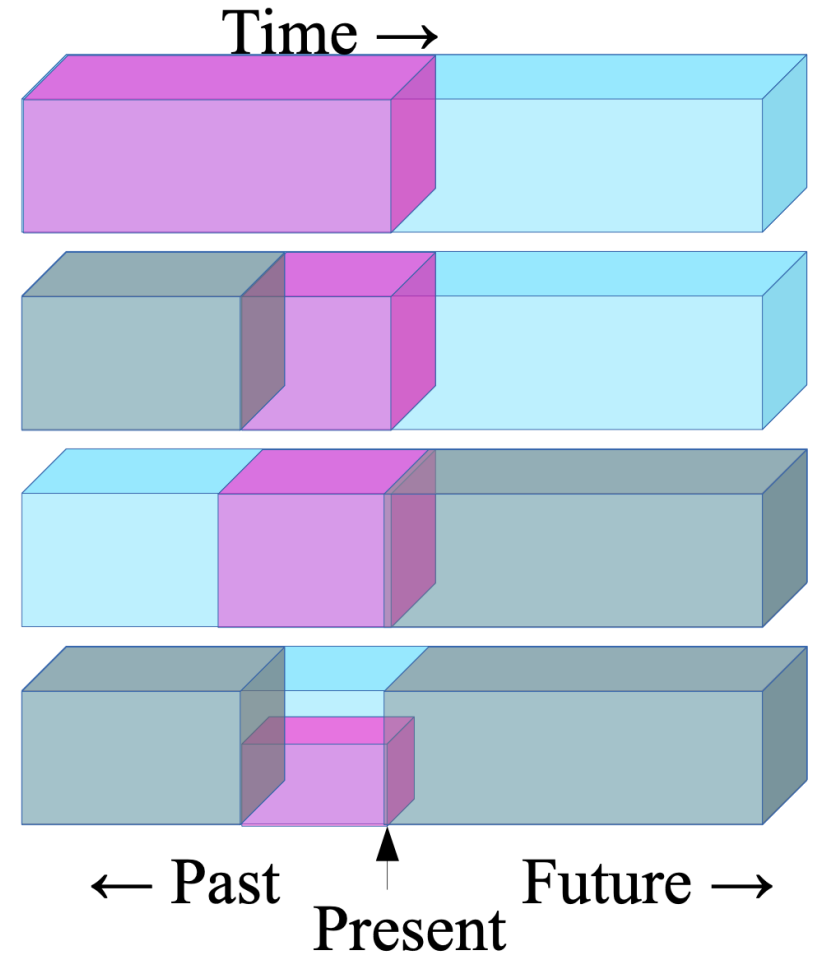
Time →

← Past Present Future →

4

Contrastive learning

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

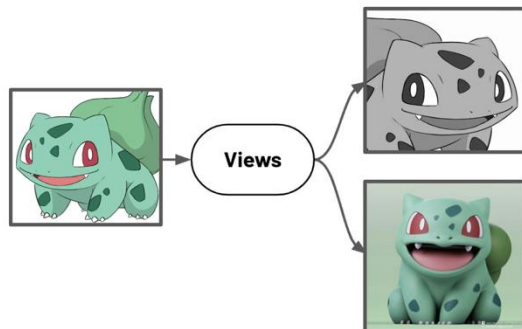- Then try fit a scoring model such that

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

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: Ex 1

Learning a similarity metric discriminatively

Sample a pair of images and compute their distance:

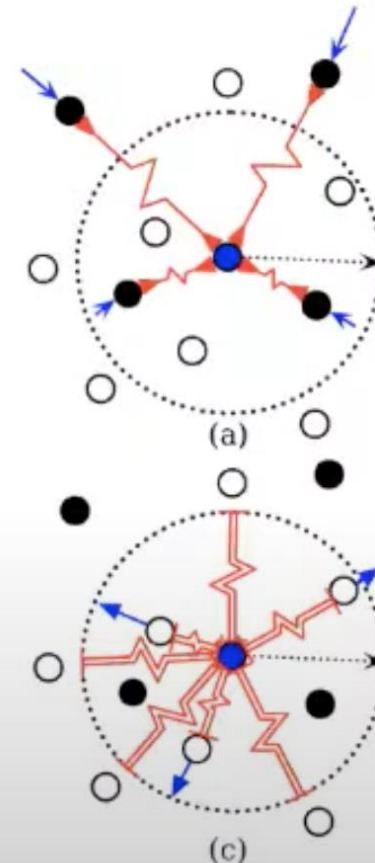$$D_i = \|x, x_i\|_2$$

If **positive** sample:

$$L_i = D_i^2$$

x   pos

If **negative** sample:

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

x   neg

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

Credit: [CVPR 2021 Tutorial] Leave Those Nets Alone: Advances in Self-Supervised Learning   8

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)

Contrastive learning: Ex 2

- SimCSE ("Simple Contrastive learning of Sentence Embeddings"; Gao et al. 2021)

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

(a) Unsupervised SimCSE

Figure 1: (a) Unsupervised SimCSE predicts the input sentence itself from in-batch negatives, with different hidden dropout masks applied.

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
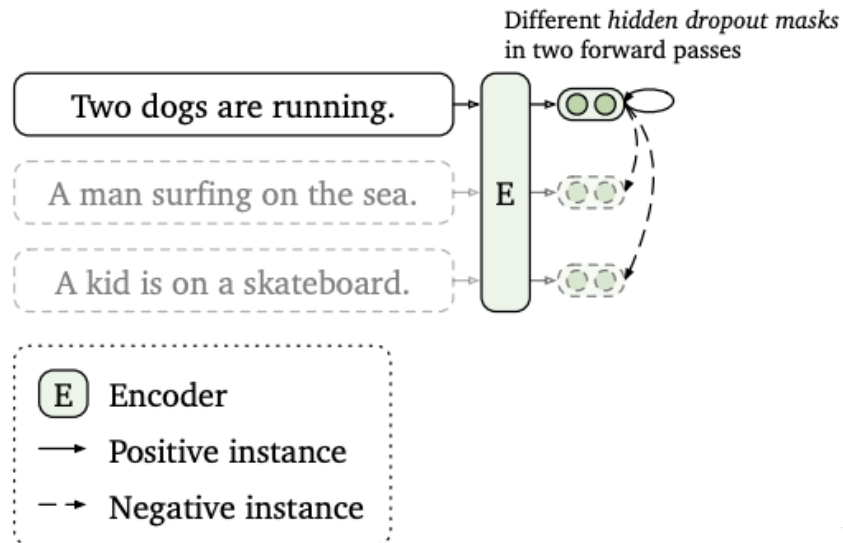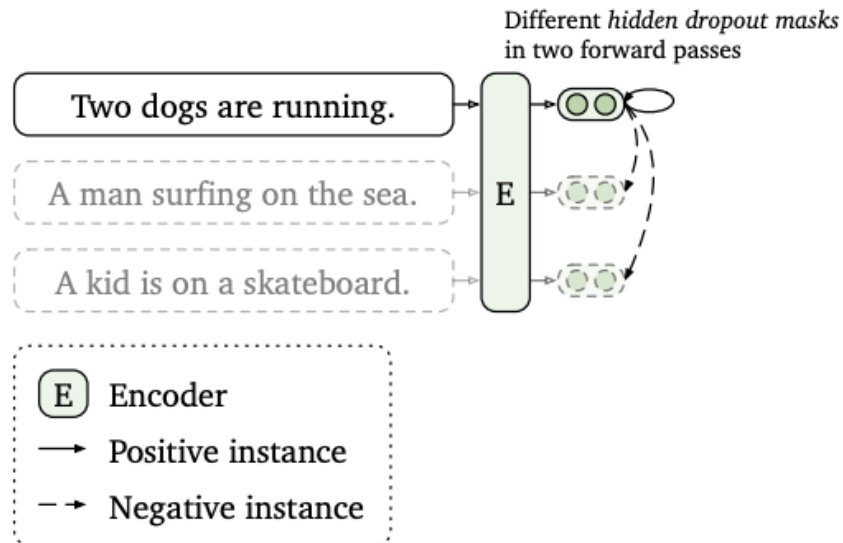
(a) Unsupervised SimCSE

(b) Supervised SimCSE

Different *hidden dropout masks* in two forward passes

Two dogs are running.

A man surfing on the sea.

A kid is on a skateboard.

E

Question

E   Encoder

→   Positive instance

--→   Negative instance

Figure 1: (a) Unsupervised SimCSE predicts the input sentence itself from in-batch negatives, with different hidden dropout masks applied.

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
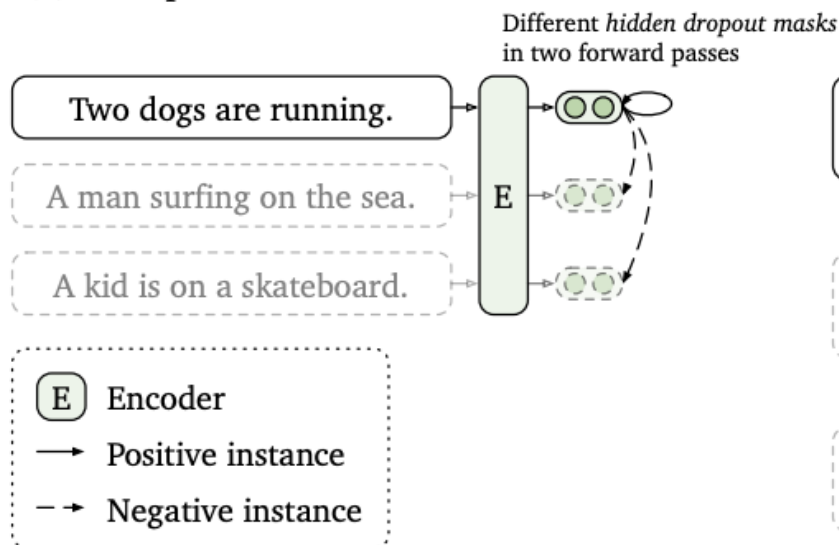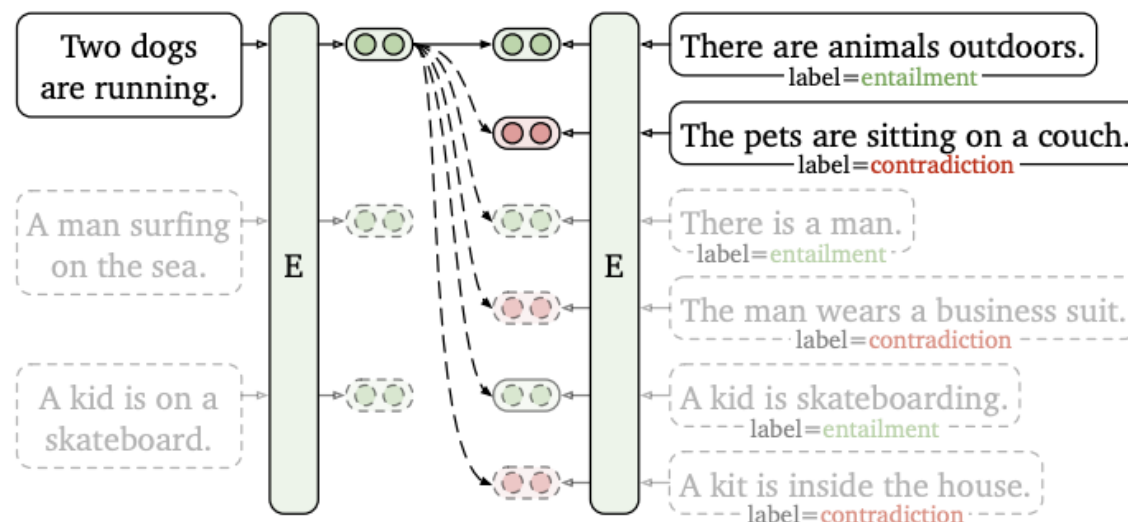

Figure 1: (a) Unsupervised SimCSE predicts the input sentence itself from in-batch negatives, with different hidden dropout masks applied. (b) Supervised SimCSE leverages the NLI datasets and takes the entailment (premise-hypothesis) pairs as positives, and contradiction pairs as well as other in-batch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

Contrastive learning: Ex 3 – InfoNCE (Noise-Contrastive Estimation)

- 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"]
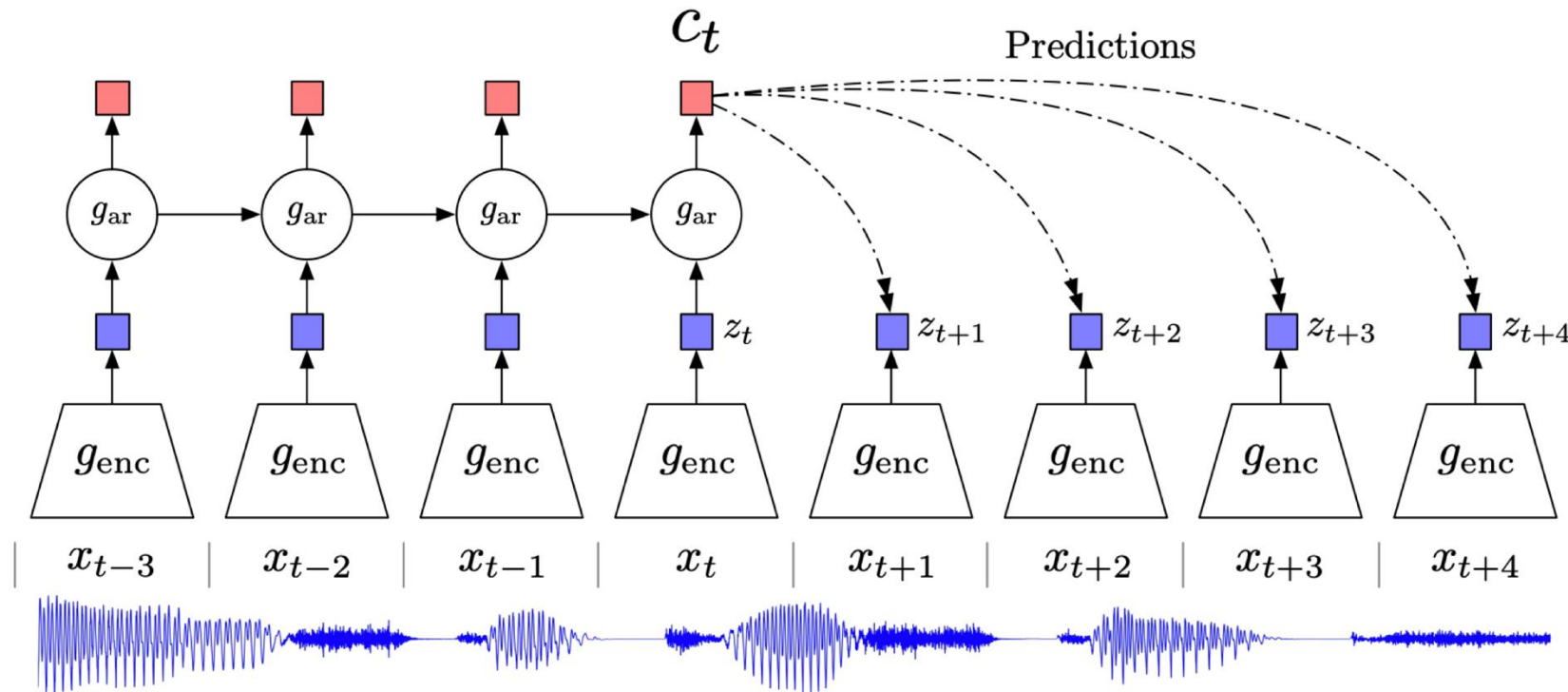
InfoNCE loss

- 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 $p(x_{t+k})$ $\}$

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

- 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"]   16

Minimizing InfoNCE <=> Maximzing MI

- InfoNCE loss

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

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

Minimizing InfoNCE <=> Maximzing MI

- InfoNCE loss

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

- 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?

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

- How does this loss maximize the mutual information?

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

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

Use proportionality condition

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

- How does this loss maximize the mutual information?

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

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

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

Take -ve inside log

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

- How does this loss maximize the mutual information?

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

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

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

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

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?

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

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

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

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

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

[van den Oord et al., "Representation Learning with Contrastive Predictive Coding"]   23

- How does this loss maximize the mutual information?

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

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

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

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

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

$$\geq \mathbb{E}_X \log \left[\frac{p(x_{t+k})}{p(x_{t+k}|c_t)} N \right]$$

$$= -I(x_{t+k}, c_t) + \log(N),$$

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

- How does this loss maximize the mutual information?

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

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

[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

Data Manipulation

Data manipulation

- Data augmentation
  - Applies label-preserving transformations on original data points to expand the data size

- Data reweighting
  - Assigns an importance weight to each instance to adapt its effect on learning

- Data synthesis
  - Generates entire artificial examples

- Curriculum learning
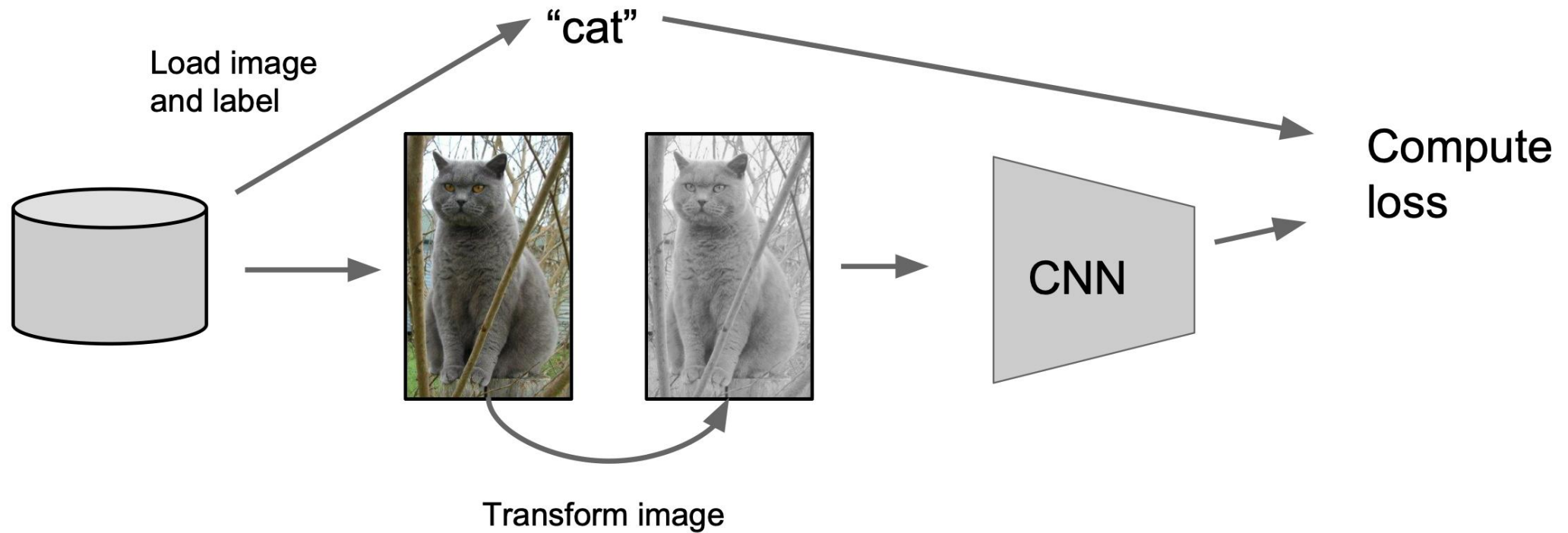 - Makes use of data instances in an order based on "difficulty"

- …

Data augmentation

- Applies **label-preserving transformations** on original data points to expand the data size
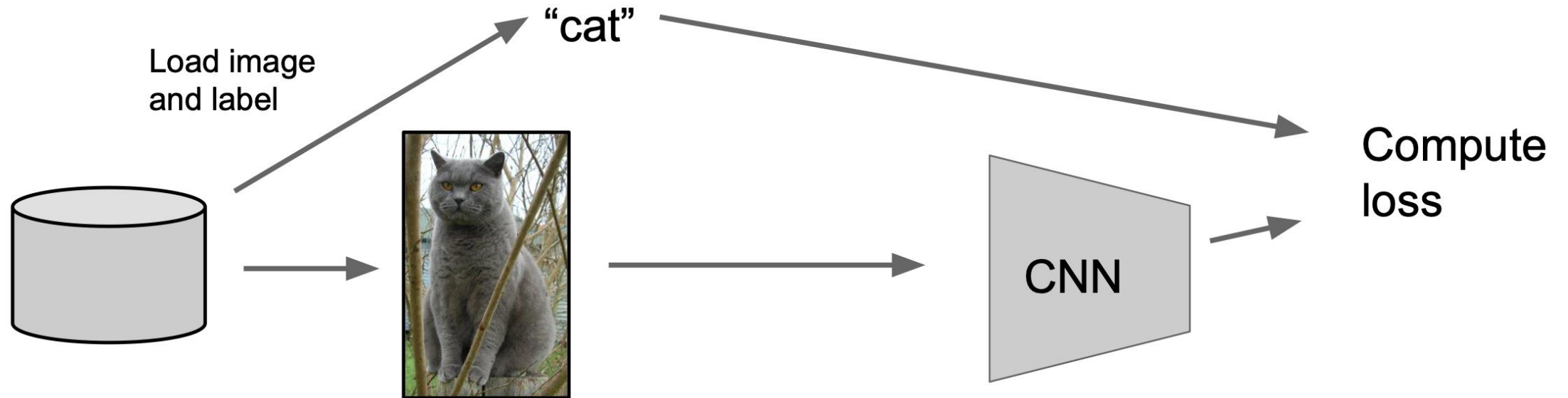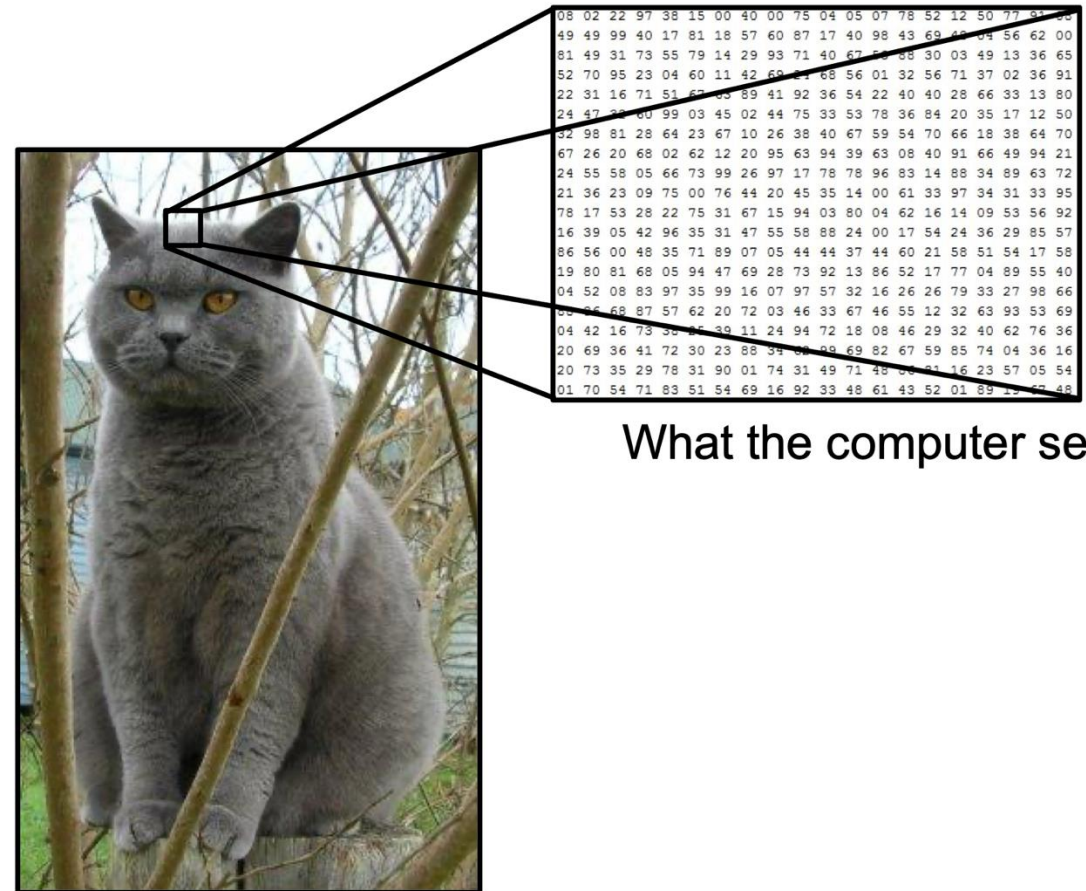
Data augmentation

- Applies **label-preserving transformations** on original data points to expand the data size
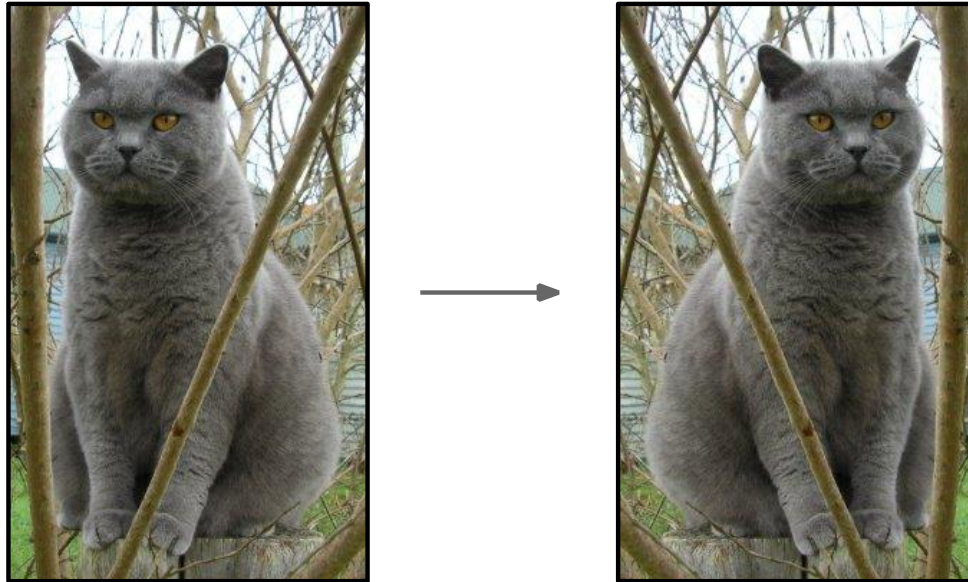
Data augmentation for image

- Change the pixels without changing the label
- Train on transformed data
- VERY widely used

What the computer sees
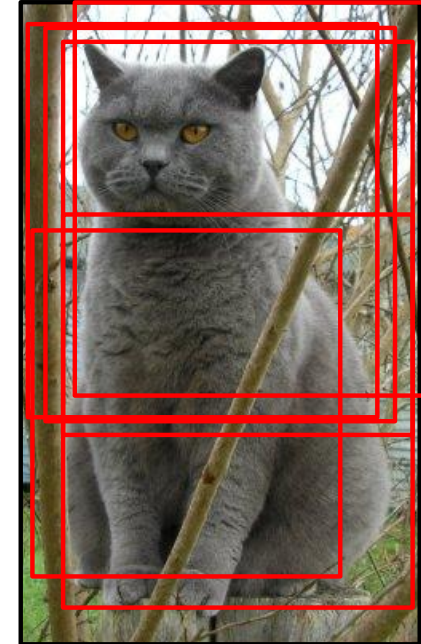
Data augmentation for image

1. Horizontal flips

Data augmentation for image

2. Random crops/scales

Training: sample random crops / scales

Data augmentation for image

2. Random crops/scales

Training: sample random crops / scales
ResNet:
1. Pick random L in range [256, 480]
2. Resize training image, short side = L
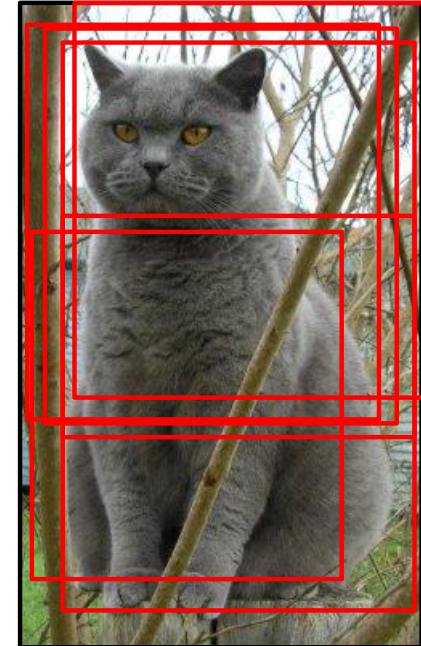3. Sample random 224 x 224 patch

Data augmentation for image

2. Random crops/scales

Training: sample random crops / scales

ResNet:
1. Pick random L in range [256, 480]
2. Resize training image, short side = L
3. Sample random 224 x 224 patch

Testing: average a fixed set of crops
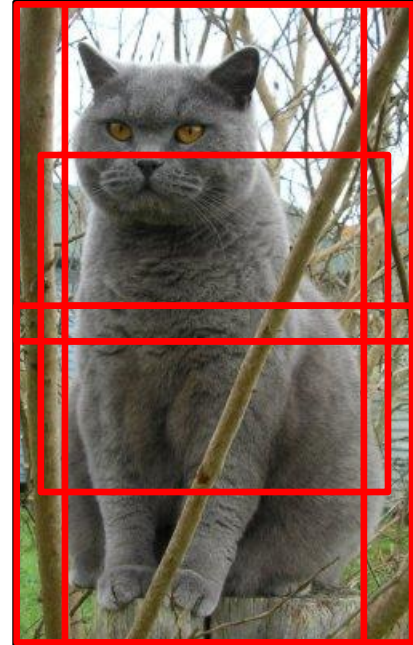
Data augmentation for image

2. Random crops/scales

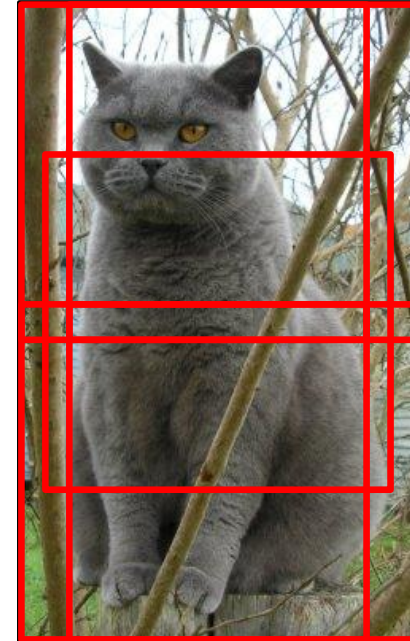**Training**: sample random crops / scales
ResNet:
1. Pick random L in range [256, 480]
2. Resize training image, short side = L
3. Sample random 224 x 224 patch

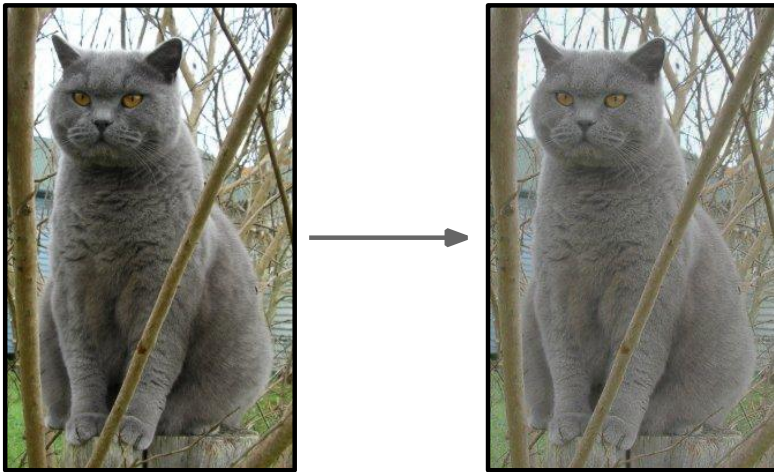**Testing**: average a fixed set of crops
ResNet:
1. Resize image at 5 scales: {224, 256, 384, 480, 640}
2. For each size, use 10 224 x 224 crops: 4 corners + center, + flips

Data augmentation for image

3. Color jitter

Randomly jitter contrast

Data augmentation for image

4. Mixup

- **Training:** Train on random blends of images

- **Testing:** Use original images

CNN → Target label:
cat: 0.4
dog: 0.6

Randomly blend the pixels of pairs of training images, e.g. 40% cat, 60% dog

[Zhang et al., "*mixup*: Beyond Empirical Risk Minimization", ICLR 2018]

Data augmentation for image

5. Get creative!

Random mix/combinations of :
- translation
- rotation
- stretching
- shearing
- lens distortions, ...

Data augmentation for text

Methods	Level	Diversity	Tasks	Related Work
Synonym replacement	Token	Low	Text classification Sequence labeling	Kolomiyets et al. (2011), Zhang et al. (2015a), Yang (2015), Miao et al. (2020), Wei and Zou (2019)
Word replacement via LM	Token	Medium	Text classification Sequence labeling Machine translation	Kolomiyets et al. (2011), Gao et al. (2019) Kobayashi (2018), Wu et al. (2019a) Fadaee et al. (2017)
Random insertion, deletion, swapping	Token	Low	Text classification Sequence labeling Machine translation Dialogue generation	Iyyer et al. (2015), Xie et al. (2017) Artetxe et al. (2018), Lample et al. (2018) Xie et al. (2020), Wei and Zou (2019)
Compositional Augmentation	Token	High	Semantic Parsing Sequence labeling Language modeling Text generation	Jia and Liang (2016) , Andreas (2020) Nye et al. (2020), Feng et al. (2020) Furrer et al. (2020) , Guo et al. (2020)
Paraphrasing	Sentence	High	Text classification Machine translation Question answering Dialogue generation Text summarization	Yu et al. (2018), Xie et al. (2020) Chen et al. (2019), He et al. (2020) Chen et al. (2020c), Cai et al. (2020)
Conditional generation	Sentence	High	Text classification Question answering	Anaby-Tavor et al. (2020), Kumar et al. (2020) Zhang and Bansal (2019), Yang et al. (2020)

Chen et al., "An Empirical Survey of Data Augmentation for Limited Data Learning in NLP"

Data augmentation for text

White-box attack	Token or Sentence	Medium	Text classification Sequence labeling Machine translation	Miyato et al. (2017), Ebrahimi et al. (2018b) Ebrahimi et al. (2018a), Cheng et al. (2019), Chen et al. (2020d)
Black-box attack	Token or Sentence	Medium	Text classification Sequence labeling Machine translation Textual entailment Dialogue generation Text Summarization	Jia and Liang (2017) Belinkov and Bisk (2017), Zhao et al. (2017) Ribeiro et al. (2018), McCoy et al. (2019) Min et al. (2020), Tan et al. (2020)
Hidden-space perturbation	Token or Sentence	High	Text classification Sequence labeling Speech recognition	Hsu et al. (2017), Hsu et al. (2018) Wu et al. (2019b), Chen et al. (2021) Malandrakis et al. (2019), Shen et al. (2020)
Interpolation	Token	High	Text classification Sequence labeling Machine translation	Miao et al. (2020), Chen et al. (2020c) Cheng et al. (2020b), Chen et al. (2020a) Guo et al. (2020)

Chen et al., "An Empirical Survey of Data Augmentation for Limited Data Learning in NLP"

Data reweighting

- Assigns an importance weight to each instance to adapt its effect on learning
  - Weighting by inverse class frequency
  - Weighting by the magnitude of loss

$$\min_{\theta} -\mathbb{E}_{x_i \sim \mathcal{D}} \left[\phi_i \log p_\theta(x_i) \right]$$

Automatically learn the data weights

- Can we learn ϕ_i automatically?

$$\min_\theta - \mathbb{E}_{x_i \sim \mathcal{D}} [\, \phi_i \log p_\theta(x_i) \,]$$

$$\mathbb{E}_{x_i \sim \mathcal{D}_v} [\, \log p_{\theta'}(x_i) \,]$$

- Training set $\mathcal{D}$, a held-out "validation" set $\mathcal{D}_v$

- Intuition: after training the model θ on the weighted data, the model gets better <u>performance on the validation set</u>

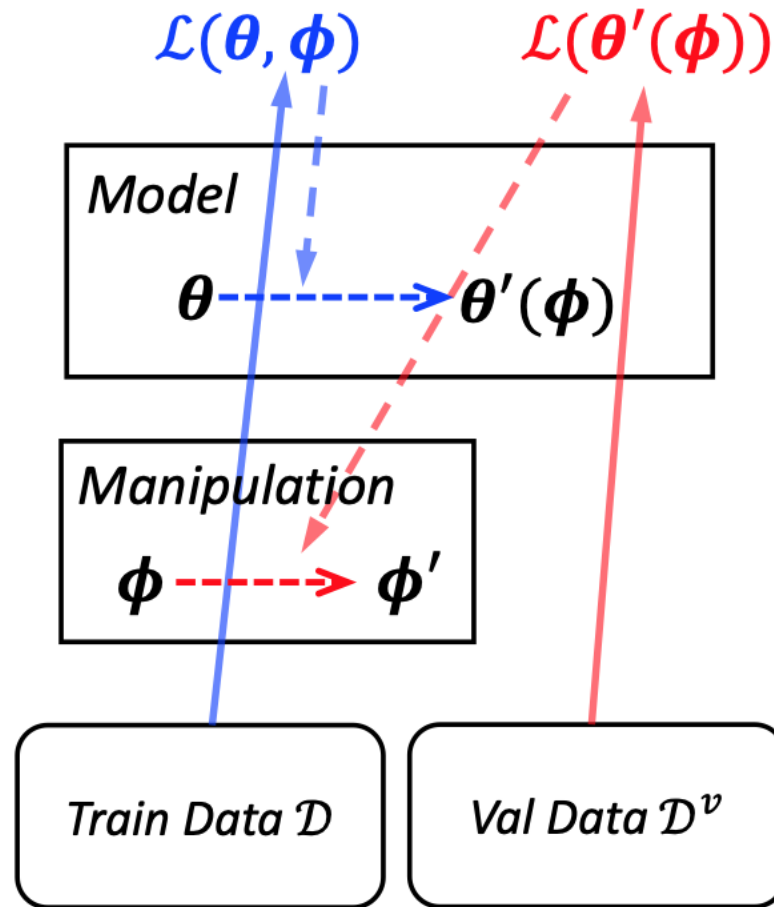$$\theta' = \operatorname*{argmin}_\theta - \mathbb{E}_{x_i \sim \mathcal{D}} [\, \phi_i \log p_\theta(x_i) \,]$$

  - θ' is a function of ϕ, i.e., $\theta' = \theta'(\phi)$

$$\phi' = \operatorname*{argmin}_\phi - \mathbb{E}_{x_i \sim \mathcal{D}_v} [\, \log p_{\theta'(\phi)}(x_i) \,]$$

Ren et al., "Learning to reweight examples for robust deep learning"

Hu et al., "Learning Data Manipulation for Augmentation and Weighting"

Automatically learn the data weights

Hu et al., "Learning Data Manipulation for Augmentation and Weighting"

Apply the same algorithm to learn data augmentation

- Augmentation function $x' = g_\phi(x)$. Can we learn ϕ automatically?

$$\min_\theta - \mathbb{E}_{x_i \sim \mathcal{D}} \left[\log p_\theta(g_\phi(x_i)) \right]$$

- Training set $\mathcal{D}$, a held-out "validation" set $\mathcal{D}_v$

- Intuition: after training the model θ on the augmented data, the model gets better performance on the validation set
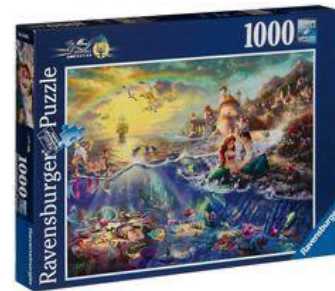
$$\theta' = \underset{\theta}{\mathrm{argmin}} - \mathbb{E}_{x_i \sim \mathcal{D}} \left[\log p_\theta(g_\phi(x_i)) \right]$$

  - θ' is a function of ϕ, i.e., $\theta' = \theta'(\phi)$

$$\phi' = \mathrm{argmin}_\phi - \mathbb{E}_{x_i \sim \mathcal{D}_v} \left[\log p_{\theta'(\phi)}(x_i) \right]$$

Hu et al., "Learning Data Manipulation for Augmentation and Weighting"

Curriculum learning

Not my first Jigsaw puzzle

Curriculum learning

My first Jigsaw puzzle

47

Curriculum learning

LEARNING COGNITIVE TASKS (CURRICULUM):

Curriculum learning

- Standard supervised learning:
  - Data is sampled randomly

- Curriculum learning:
  - Instead of randomly selecting training points, select easier examples first, slowly exposing the more difficult examples from easiest to the most difficult
  - Key: definition of "difficulty"

Curriculum learning

- (Bengio et al, 2009): setup of paradigm, object recognition of geometric shapes using a perceptron; *difficulty is determined by user from geometric shape*

- (Zaremba 2014): LSTMs used to evaluate short computer programs; *difficulty is automatically evaluated from data – nesting level of program*.

- (Amodei et al, 2016): End-to-end speech recognition in english and mandarin; *difficulty is automatically evaluated from utterance length.*

- (Jesson et al, 2017): deep learning segmentation and detection; *human teacher (user/programmer) determins difficulty.*

Key Takeaways

- Data manipulation
  - Augmentation
  - Reweighting
  - Curriculum learning
  - Synthesis (later)

Questions?