DSC291: Machine Learning with Few Labels

Data Manipulation

Zhiting Hu

Lecture 10, May 1, 2025

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

Outline

- Data Manipulation
  - (meta learning)

- Paper presentation:
  - Akbota Assan, Derrick Yao: “Reasoning Models Don't Always Say What They Think”

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

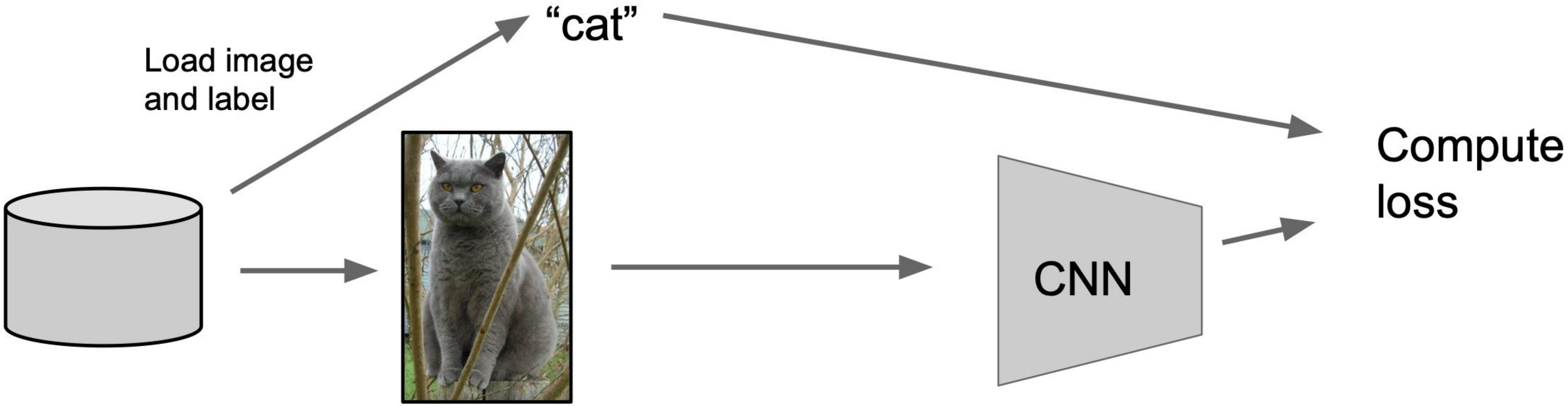


Figure credit: http://cs231n.stanford.edu/slides/2016/winter1516_lecture11.pdf

Data augmentation

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

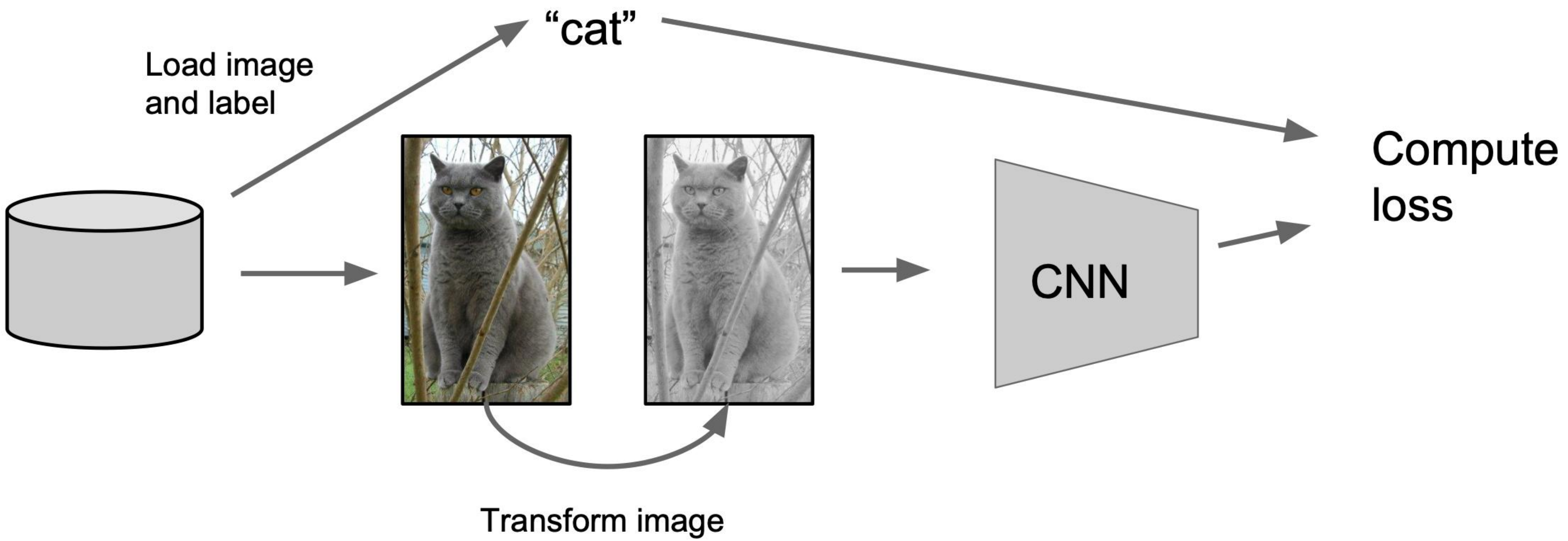
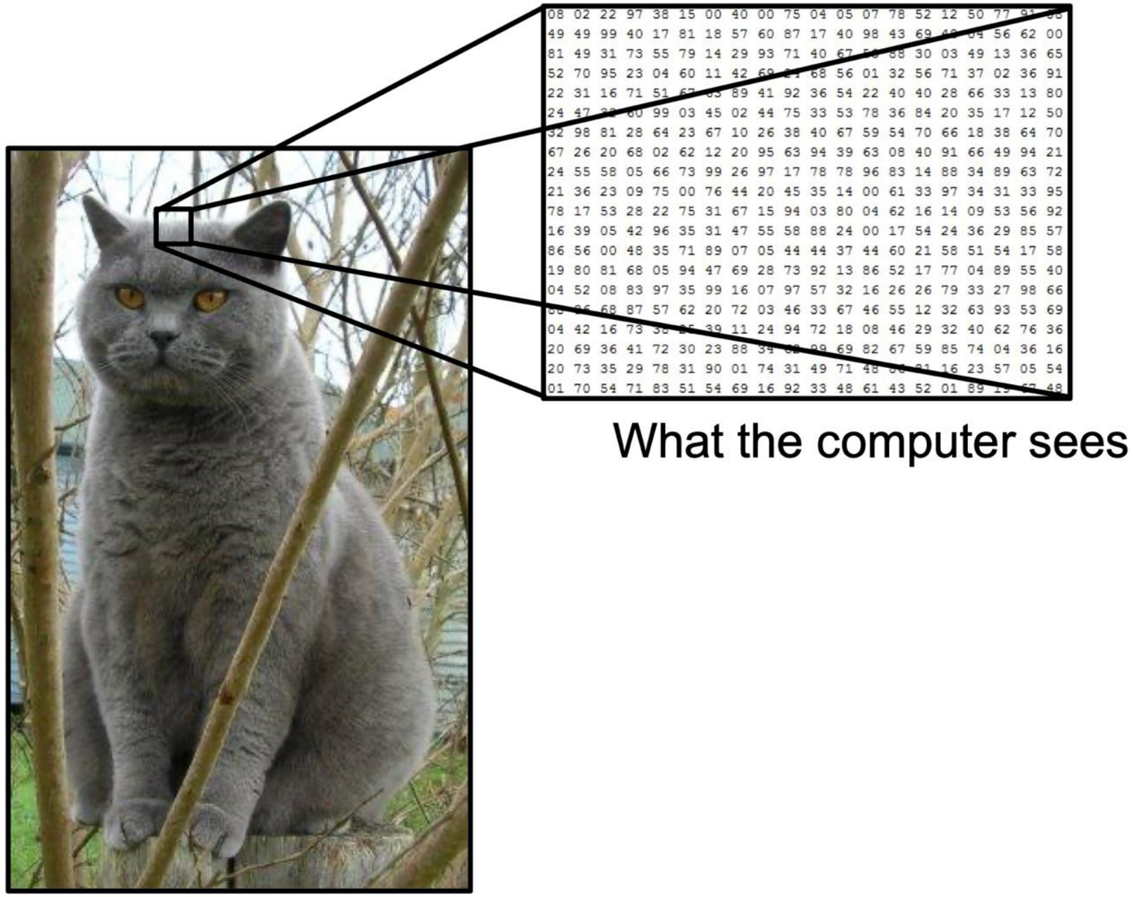


Figure credit: http://cs231n.stanford.edu/slides/2016/winter1516_lecture11.pdf

Data augmentation for image

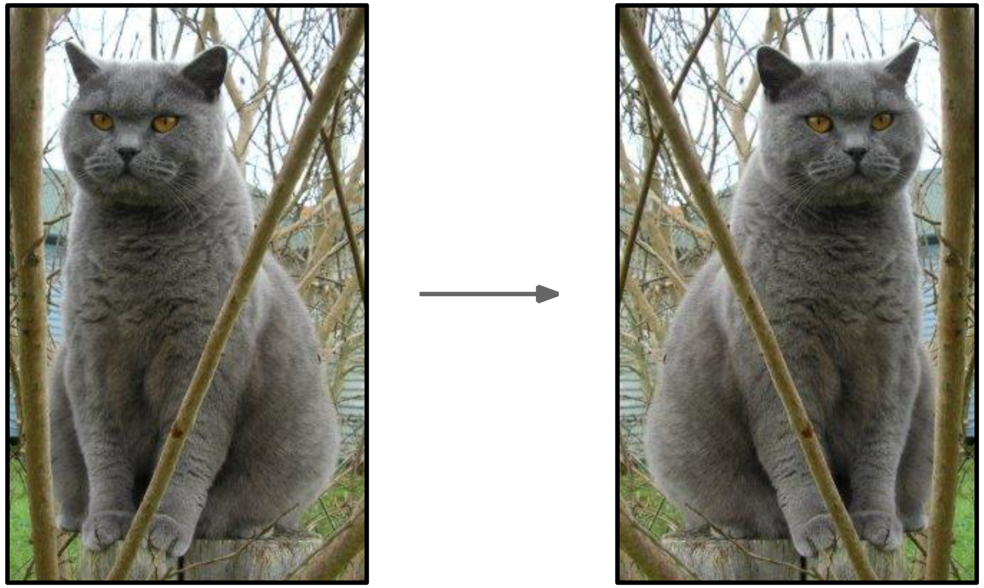
- Change the pixels without changing the label
- Train on transformed data
- VERY widely used



Credit: http://cs231n.stanford.edu/slides/2016/winter1516_lecture11.pdf

Data augmentation for image

1. Horizontal flips

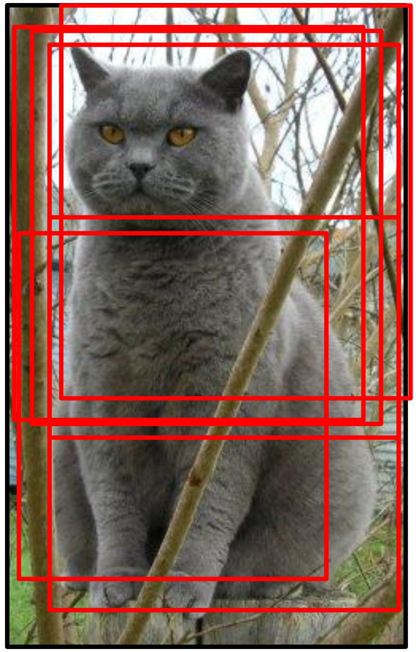


Credit: http://cs231n.stanford.edu/slides/2016/winter1516_lecture11.pdf

Data augmentation for image

2. Random crops/scales

Training: sample random crops / scales



Credit: http://cs231n.stanford.edu/slides/2016/winter1516_lecture11.pdf

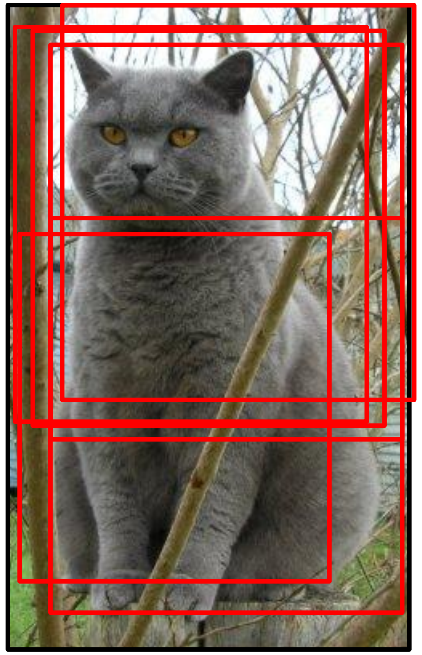
Data augmentation for image

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



Credit: http://cs231n.stanford.edu/slides/2016/winter1516_lecture11.pdf

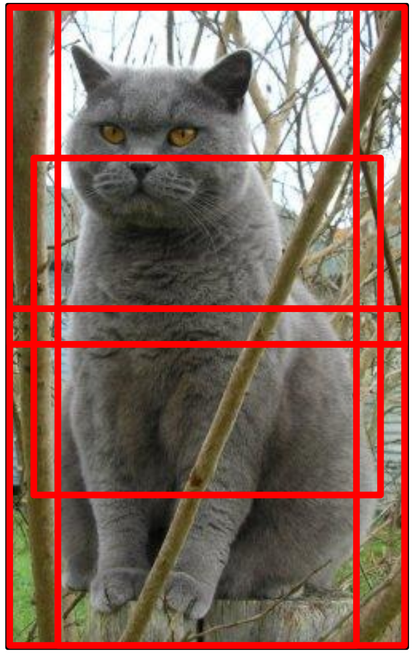
Data augmentation for image

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

Credit: http://cs231n.stanford.edu/slides/2016/winter1516_lecture11.pdf

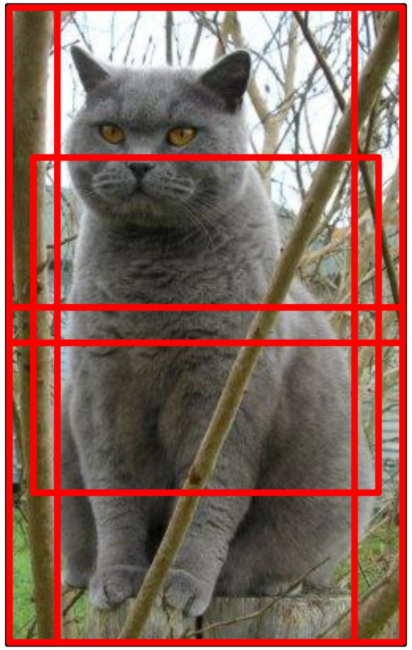
Data augmentation for image

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:

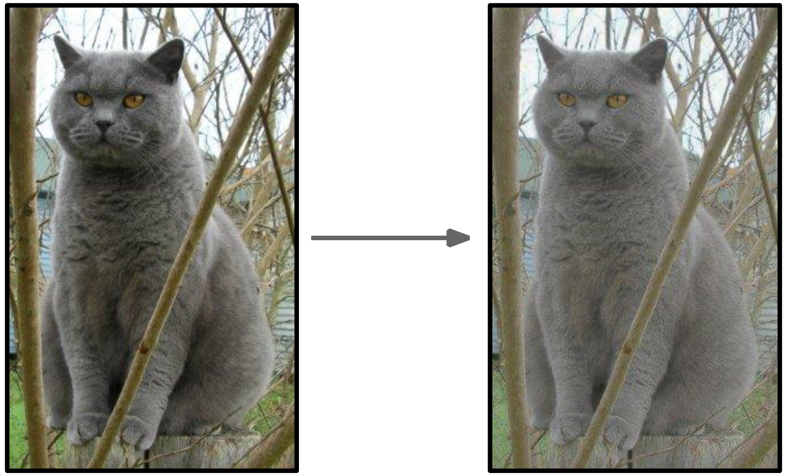
1. Resize image at 5 scales: {224, 256, 384, 480, 640}
2. For each size, use 10 224 x 224 crops: 4 corners + center, + flips

Credit: http://cs231n.stanford.edu/slides/2016/winter1516_lecture11.pdf

Data augmentation for image

3. Color jitter

Randomly jitter contrast

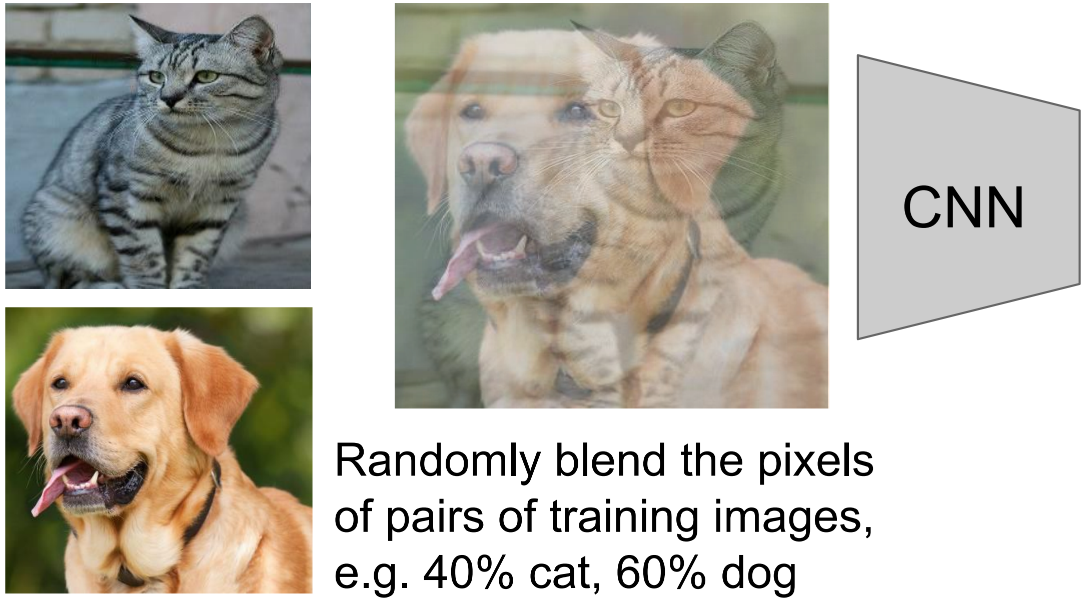


Credit: http://cs231n.stanford.edu/slides/2016/winter1516_lecture11.pdf

Data augmentation for image

4. Mixup

- **Training:** Train on random blends of images
- **Testing:** Use original images



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]

Credit: http://cs231n.stanford.edu/slides/2019/cs231n_2019_lecture08.pdf

Data augmentation for image

5. Get creative!

Random mix/combinations of :

- translation
- rotation
- stretching
- shearing
- lens distortions, …

Credit: http://cs231n.stanford.edu/slides/2016/winter1516_lecture11.pdf

Data augmentation for text

- Token-level augmentation

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)

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

Data augmentation for text

- Sentence-level augmentation

Methods	Level	Diversity	Tasks	Related Work
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

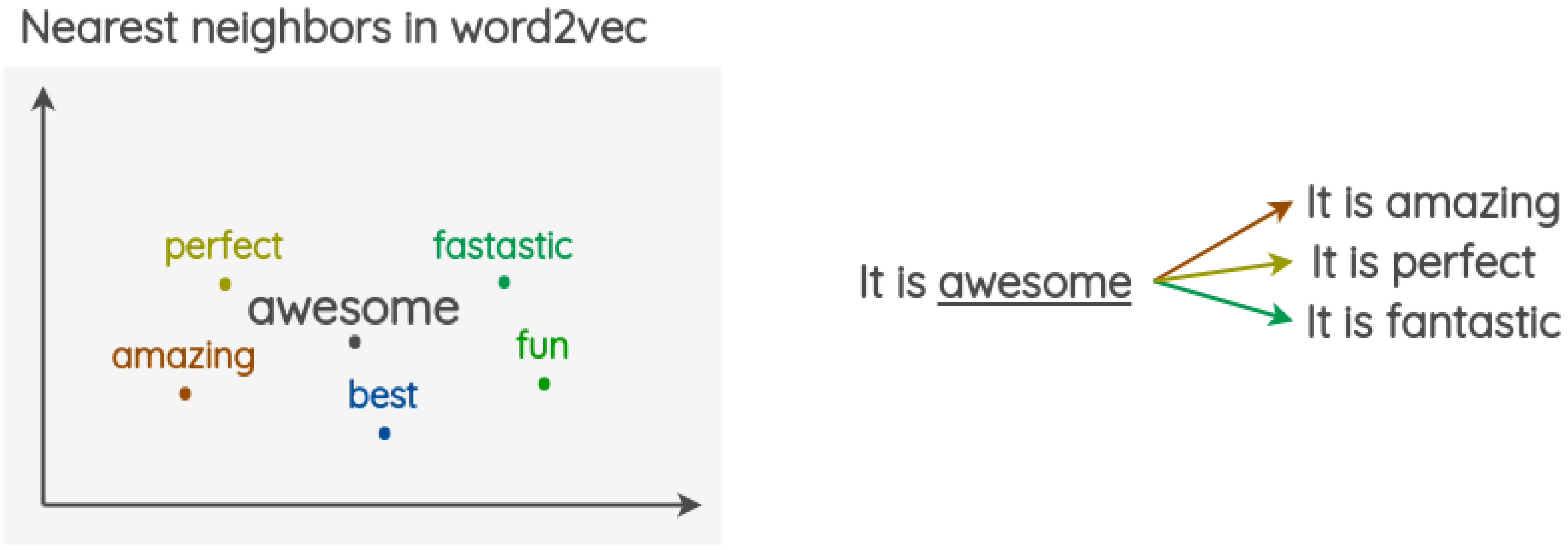
- Others

Methods	Level	Diversity	Tasks	Related Work
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 augmentation for text: Examples

- **Lexical Substitution**
  - **Word-embedding substitution**



[Courtesy: Amit Chaudhary https://amitness.com/2020/05/data-augmentation-for-nlp/]

Data augmentation for text: Examples

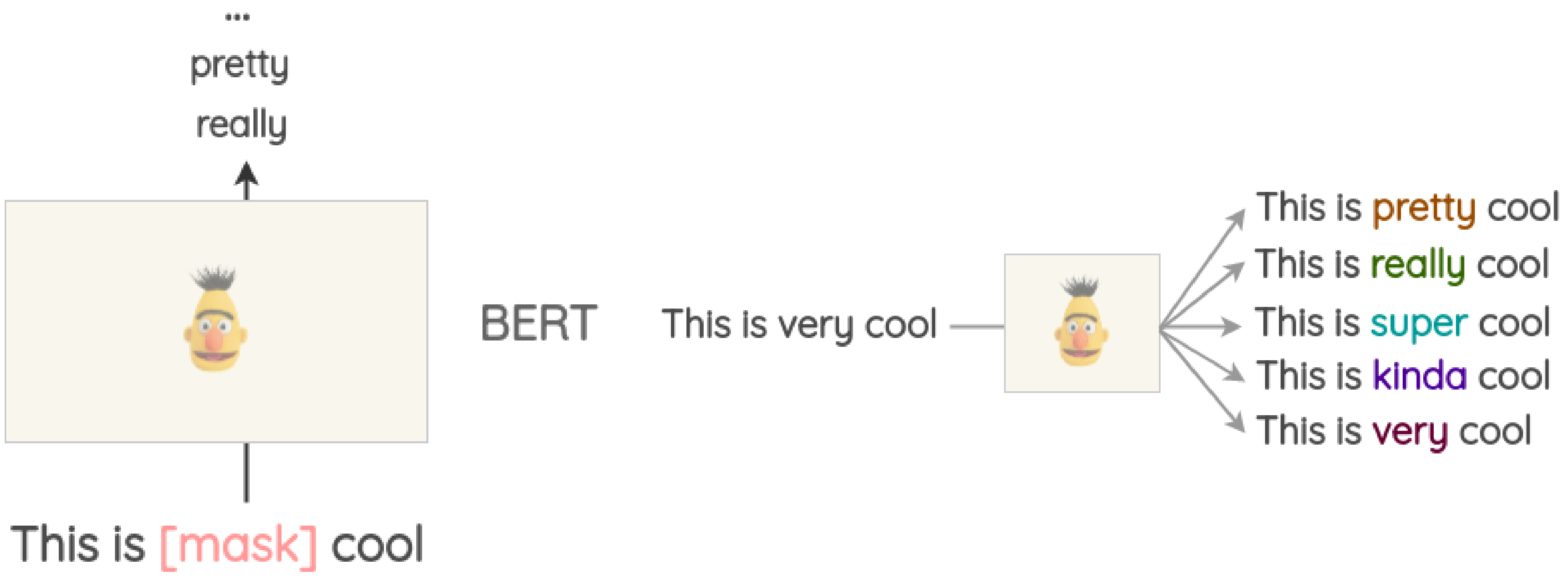
- **Lexical Substitution**
  - **Masked LM**

This is very cool

[Courtesy: Amit Chaudhary https://amitness.com/2020/05/data-augmentation-for-nlp/]

Data augmentation for text: Examples

- **Lexical Substitution**
  - **Masked LM**



[Courtesy: Amit Chaudhary https://amitness.com/2020/05/data-augmentation-for-nlp/]

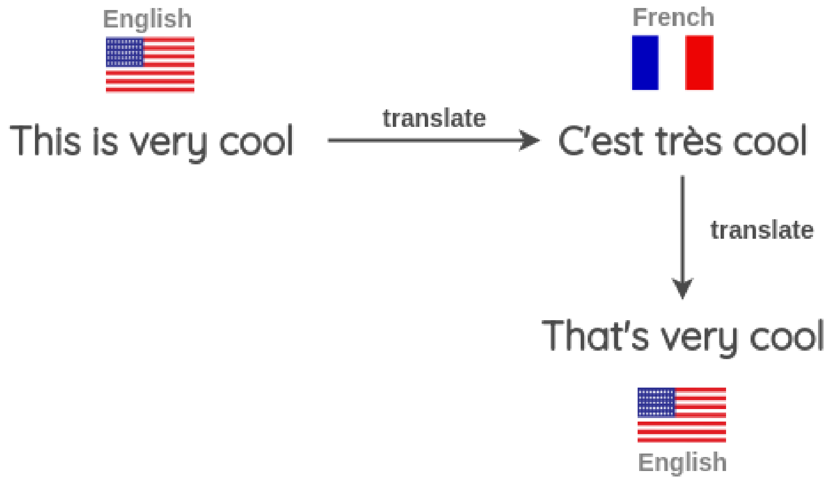
Data augmentation for text: Examples

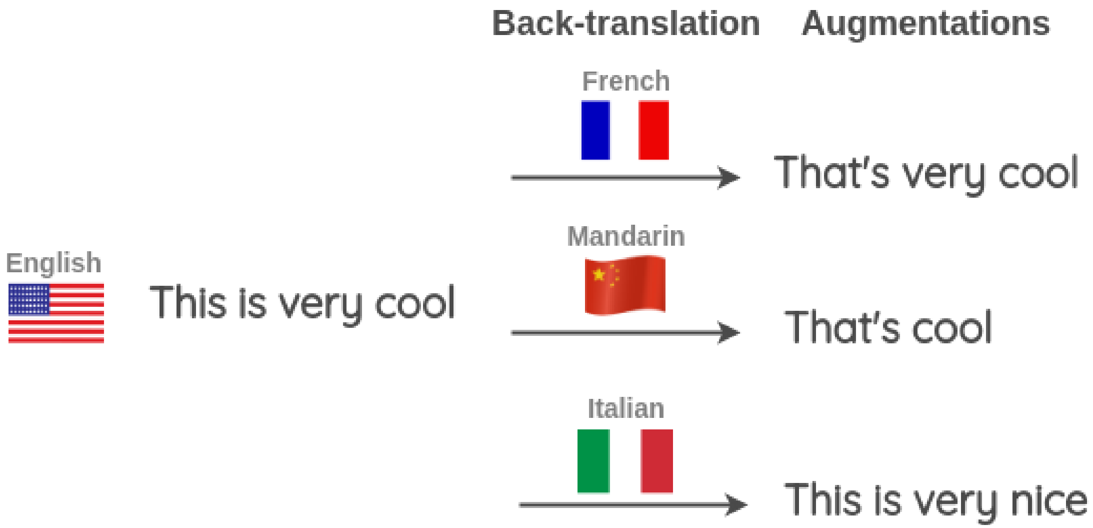
- **Paraphrasing**
  - **Back Translation**

[Courtesy: Amit Chaudhary https://amitness.com/2020/05/data-augmentation-for-nlp/]

Data augmentation for text: Examples

- **Paraphrasing**
  - **Back Translation**

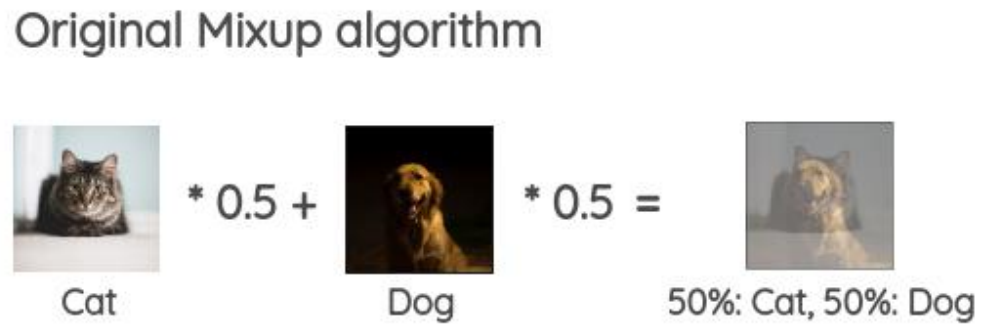




[Courtesy: Amit Chaudhary https://amitness.com/2020/05/data-augmentation-for-nlp/]

Data augmentation for text: Examples

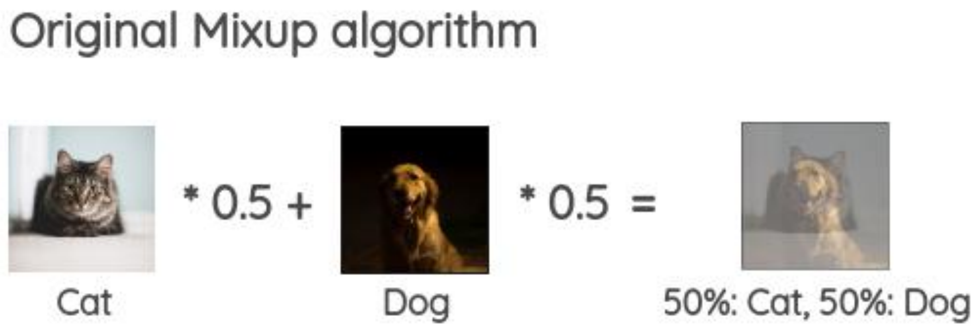
- Lexical Substitution
  - Thesaurus-based substitution
  - Word-embedding substitution
  - Masked LM
  - TF-IDF based word replacement
- Paraphrasing
  - Back Translation
- MixUp

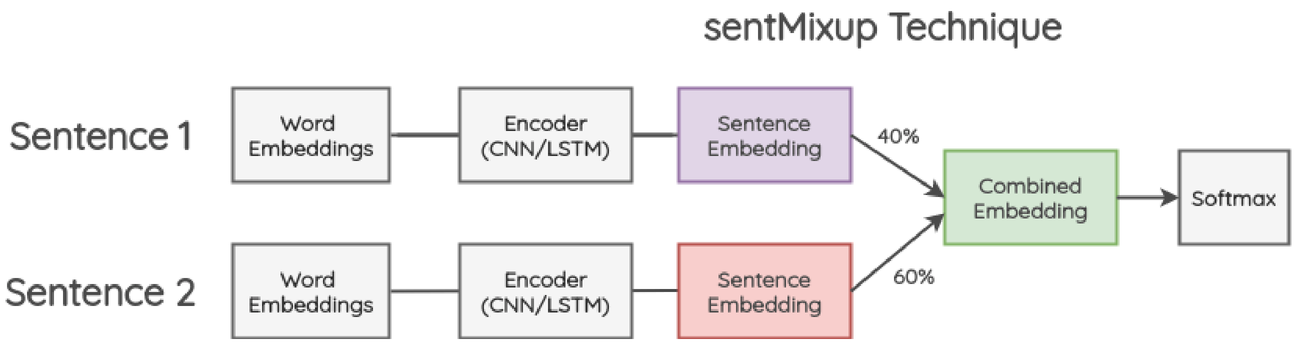


[Courtesy: Amit Chaudhary https://amitness.com/2020/05/data-augmentation-for-nlp/]

Data augmentation for text: Examples

- **Lexical Substitution**
  - **Thesaurus-based substitution**
  - **Word-embedding substitution**
  - **Masked LM**
  - **TF-IDF based word replacement**
- **Paraphrasing**
  - **Back Translation**
- **MixUp**





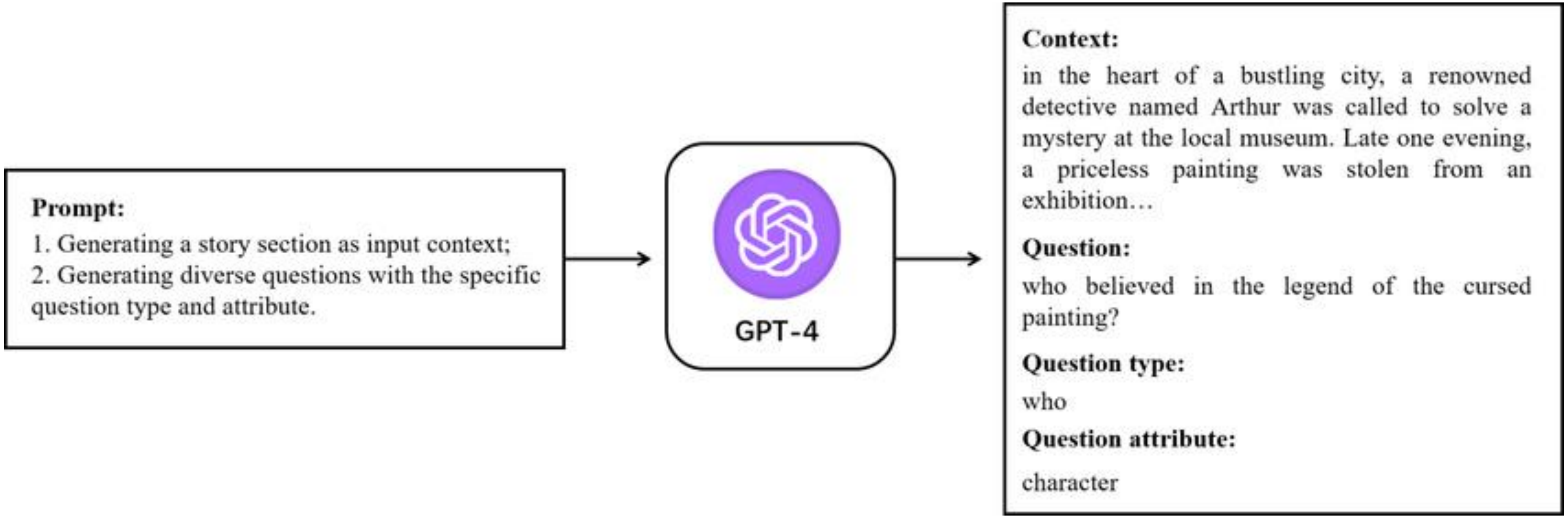
[Courtesy: Amit Chaudhary https://amitness.com/2020/05/data-augmentation-for-nlp/]

Data augmentation for text: Examples

- **Lexical Substitution**
  - **Thesaurus-based substitution**
  - **Word-embedding substitution**
  - **Masked LM**
  - **TF-IDF based word replacement**
- **Paraphrasing**
  - **Back Translation**
- **MixUp**
- **Generative Models**
  - Use pretrained LM to generate new data

[Courtesy: Amit Chaudhary https://amitness.com/2020/05/data-augmentation-for-nlp/]

Data augmentation for text: Examples



- **Generative Models**
  - Use pretrained LM to generate new data

[Courtesy: Amit Chaudhary https://amitness.com/2020/05/data-augmentation-for-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

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

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 performance on the validation set

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

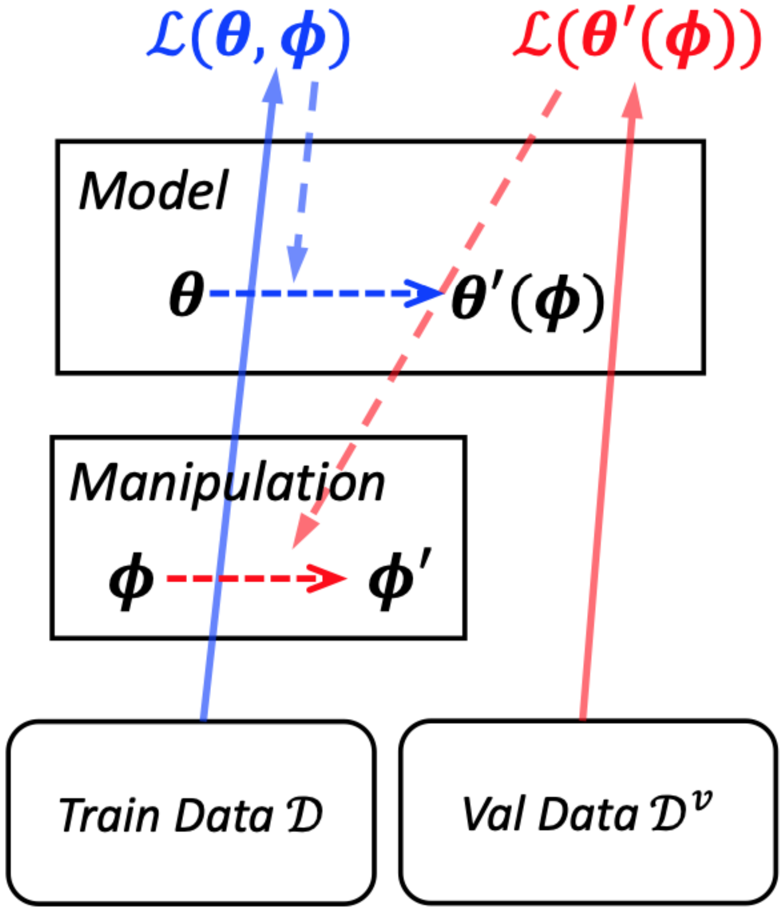
  - θ' is a function of ϕ, i.e., $\theta' = \theta'(\phi)$

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

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

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

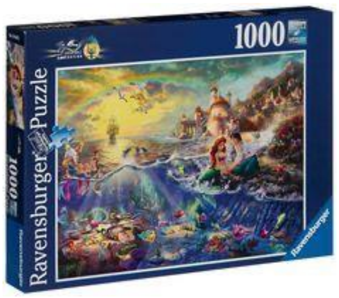
  - θ' is a function of ϕ, i.e., $\theta' = \theta'(\phi)$

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

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

Curriculum learning

Not my first Jigsaw puzzle



Credit: Weinshall, “ON THE POWER OF CURRICULUM LEARNING IN TRAINING DEEP NETWORKS“

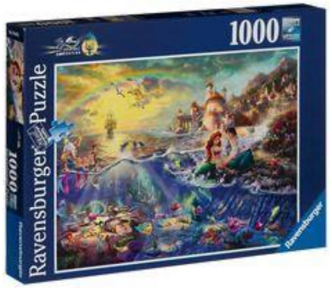
Curriculum learning

MY FIRST JIGSAW PUZZLE

Credit: Weinshall, “ON THE POWER OF CURRICULUM LEARNING IN TRAINING DEEP NETWORKS“

Curriculum learning

LEARNING COGNITIVE TASKS (CURRICULUM):



Credit: Weinshall, “ON THE POWER OF CURRICULUM LEARNING IN TRAINING DEEP NETWORKS“

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”

Credit: Weinshall, “ON THE POWER OF CURRICULUM LEARNING IN TRAINING DEEP NETWORKS“

Key Takeaways

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

How to get more labeled training data?

- **Traditional Supervision:** Have subject matter experts (SMEs) hand-label more training data
  - *Too expensive!*
  - **Active Learning:** Estimate which points are most valuable to solicit labels for
- **Semi-supervised Learning:** Use structural assumptions to automatically leverage unlabeled data
- ***Weak* Supervision:** Get lower-quality labels more efficiently and/or at a higher abstraction level
  - *Get cheaper, lower-quality labels from non-experts*
  - ***Get higher-level supervision over unlabeled data from SMEs***
    - *Heuristics*
    - *Distant Supervision*
    - *Constraints*
    - *Expected distributions*
    - *Invariances*
  - *Use one or more (noisy / biased) pre-trained models to provide supervision*
- **Transfer Learning:** Use models already trained on a different task

Credit: https://dawn.cs.stanford.edu/2017/07/16/weak-supervision/

Example (I): labeling with heuristics

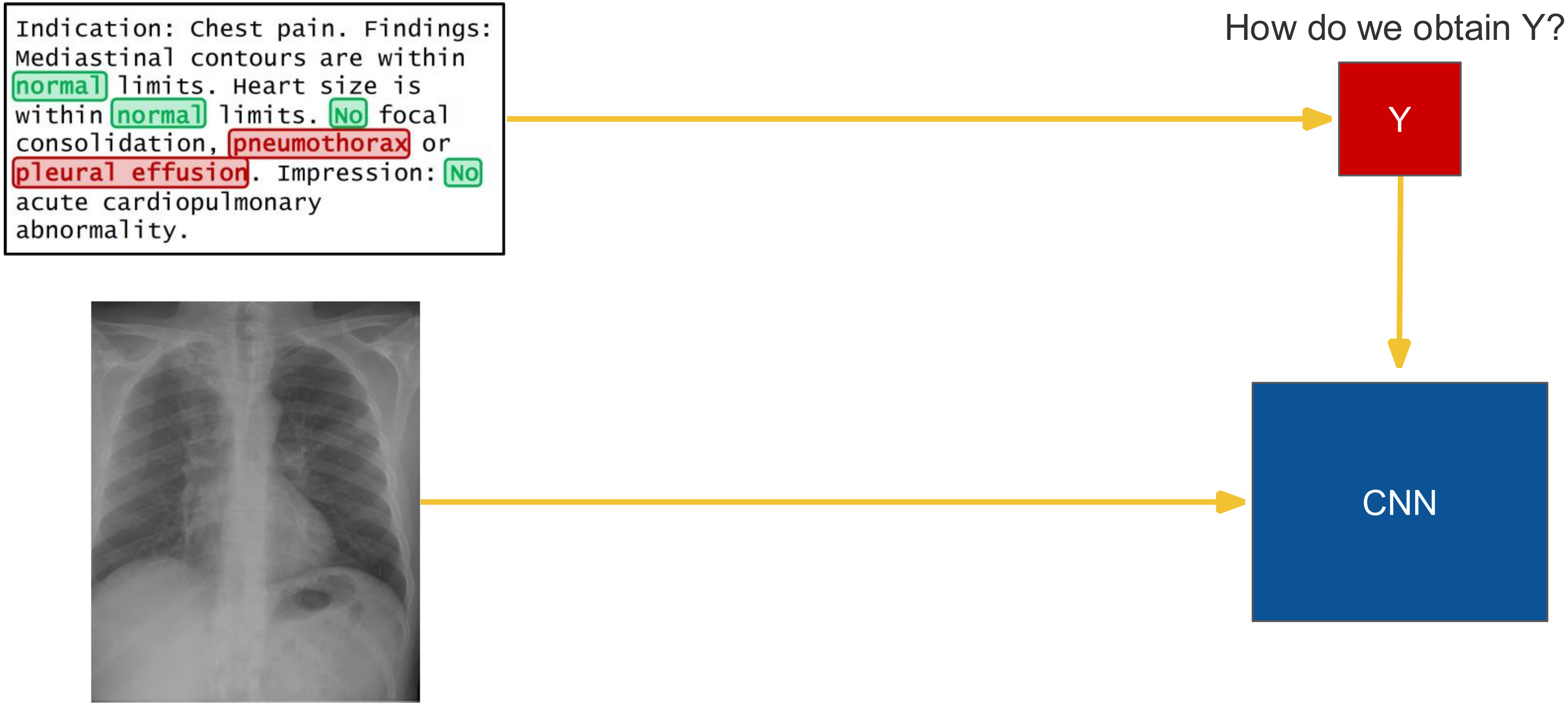
Task: Build a chest x-ray classifier
(normal/abnormal)

Indication: Chest pain. Findings: Mediastinal contours are within normal limits. Heart size is within normal limits. No focal consolidation, pneumothorax or pleural effusion. Impression: No acute cardiopulmonary abnormality.

Can you use the accompanying medical report (text modality) to label the x-ray (image modality)?

[Credit: http://cs231n.stanford.edu/slides/2018/cs231n_2018_ds07.pdf]

Example (I): labeling with heuristics



[Credit: http://cs231n.stanford.edu/slides/2018/cs231n_2018_ds07.pdf]

Example (I): labeling with heuristics

Indication: Chest pain. Findings: Mediastinal contours are within normal limits. Heart size is within normal limits. No focal consolidation, pneumothorax or pleural effusion. Impression: No acute cardiopulmonary abnormality.

Normal Report

```python
def LF_pneumothorax(c):
  if re.search(r'pneumo.*', c.report.text):
      return "ABNORMAL"

def LF_pleural_effusion(c):
  if "pleural effusion" in c.report.text:
      return "ABNORMAL"

def LF_normal_report(c, thresh=2):
    if len(NORMAL_TERMS.intersection(c.
    report.words)) > thresh:
      return "NORMAL"
```

LFs
(labeling functions)

Source: Khandwala et. al 2017, Cross Modal Data Programming for Medical Images

[Credit: http://cs231n.stanford.edu/slides/2018/cs231n_2018_ds07.pdf]

Example (II): Labeling with knowledge bases

Task: relation extraction from text

- Hypothesis: If two entities belong to a certain relation, any sentence containing those two entities is likely to express that relation
- Key idea: use a *knowledge base* of relations to get lots of *noisy* training examples

Adapted from https://courses.cs.washington.edu/courses/cse517/13wi/slides/cse517wi13-RelationExtractionII.pdf

Example (II): Labeling with knowledge bases

Frequent Freebase relations

Relation name	Size	Example
/people/person/nationality	281,107	John Dugard, South Africa
/location/location/contains	253,223	Belgium, Nijlen
/people/person/profession	208,888	Dusa McDuff, Mathematician
/people/person/place_of_birth	105,799	Edwin Hubble, Marshfield
/dining/restaurant/cuisine	86,213	MacAyo's Mexican Kitchen, Mexican
/business/business_chain/location	66,529	Apple Inc., Apple Inc., South Park, NC
/biology/organism_classification_rank	42,806	Scorpaeniformes, Order
/film/film/genre	40,658	Where the Sidewalk Ends, Film noir
/film/film/language	31,103	Enter the Phoenix, Cantonese
/biology/organism_higher_classification	30,052	Calopteryx, Calopterygidae
/film/film/country	27,217	Turtle Diary, United States
/film/writer/film	23,856	Irving Shulman, Rebel Without a Cause
/film/director/film	23,539	Michael Mann, Collateral
/film/producer/film	22,079	Diane Eskenazi, Aladdin
/people/deceased_person/place_of_death	18,814	John W. Kern, Asheville
/music/artist/origin	18,619	The Octopus Project, Austin
/people/person/religion	17,582	Joseph Chartrand, Catholicism
/book/author/works_written	17,278	Paul Auster, Travels in the Scriptorium
/soccer/football_position/players	17,244	Midfielder, Chen Tao
/people/deceased_person/cause_of_death	16,709	Richard Daintree, Tuberculosis
/book/book/genre	16,431	Pony Soldiers, Science fiction
/film/film/music	14,070	Stavisky, Stephen Sondheim
/business/company/industry	13,805	ATS Medical, Health care

Adapted from https://courses.cs.washington.edu/courses/cse517/13wi/slides/cse517wi13-RelationExtractionII.pdf

Example (II): Labeling with knowledge bases

Corpus text

Bill Gates founded Microsoft in 1975.
Bill Gates, founder of Microsoft, …
Bill Gates attended Harvard from…
Google was founded by Larry Page …

Training data

Freebase

Founder: (Bill Gates, Microsoft)
Founder: (Larry Page, Google)
CollegeAttended: (Bill Gates, Harvard)

Credit: https://courses.cs.washington.edu/courses/cse517/13wi/slides/cse517wi13-RelationExtractionII.pdf

Example (II): Labeling with knowledge bases

Corpus text

<u>Bill Gates</u> founded <u>Microsoft</u> in 1975.
Bill Gates, founder of Microsoft, …
Bill Gates attended Harvard from…
Google was founded by Larry Page …

Training data

(Bill Gates, Microsoft)
Label: Founder
Feature: X founded Y

Freebase

Founder: (<u>Bill Gates</u>, <u>Microsoft</u>)
Founder: (Larry Page, Google)
CollegeAttended: (Bill Gates, Harvard)

Credit: https://courses.cs.washington.edu/courses/cse517/13wi/slides/cse517wi13-RelationExtractionII.pdf

Example (II): Labeling with knowledge bases

Corpus text

Bill Gates founded Microsoft in 1975.
Bill Gates, founder of Microsoft, …
Bill Gates attended Harvard from…
Google was founded by Larry Page …

Training data

(Bill Gates, Microsoft)
Label: Founder
Feature: X founded Y
Feature: X, founder of Y

Freebase

Founder: (Bill Gates, Microsoft)
Founder: (Larry Page, Google)
CollegeAttended: (Bill Gates, Harvard)

Credit: https://courses.cs.washington.edu/courses/cse517/13wi/slides/cse517wi13-RelationExtractionII.pdf

Example (II): Labeling with knowledge bases

Corpus text

Bill Gates founded Microsoft in 1975.
Bill Gates, founder of Microsoft, …
Bill Gates attended Harvard from…
Google was founded by Larry Page …

Freebase

Founder: (Bill Gates, Microsoft)
Founder: (Larry Page, Google)
CollegeAttended: (Bill Gates, Harvard)

Training data

(Bill Gates, Microsoft)
Label: Founder
Feature: X founded Y
Feature: X, founder of Y

(Bill Gates, Harvard)
Label: CollegeAttended
Feature: X attended Y

Credit: https://courses.cs.washington.edu/courses/cse517/13wi/slides/cse517wi13-RelationExtractionII.pdf

Example (II): Labeling with knowledge bases

Corpus text

Bill Gates founded Microsoft in 1975.
Bill Gates, founder of Microsoft, …
Bill Gates attended Harvard from…
Google was founded by Larry Page …

Freebase

Founder: (Bill Gates, Microsoft)
Founder: (Larry Page, Google)
CollegeAttended: (Bill Gates, Harvard)

Training data

(Bill Gates, Microsoft)
Label: Founder
Feature: X founded Y
Feature: X, founder of Y

(Bill Gates, Harvard)
Label: CollegeAttended
Feature: X attended Y

(Larry Page, Google)
Label: Founder
Feature: Y was founded by X

Credit: https://courses.cs.washington.edu/courses/cse517/13wi/slides/cse517wi13-RelationExtractionII.pdf

Example (II): Labeling with knowledge bases

Negative training data

Can't train a classifier with only positive data!
Need negative training data too!

Solution?
Sample 1% of unrelated pairs of entities.

Corpus text

Larry Page took a swipe at Microsoft...
...after Harvard invited Larry Page to...
Google is Bill Gates' worst fear ...

Training data

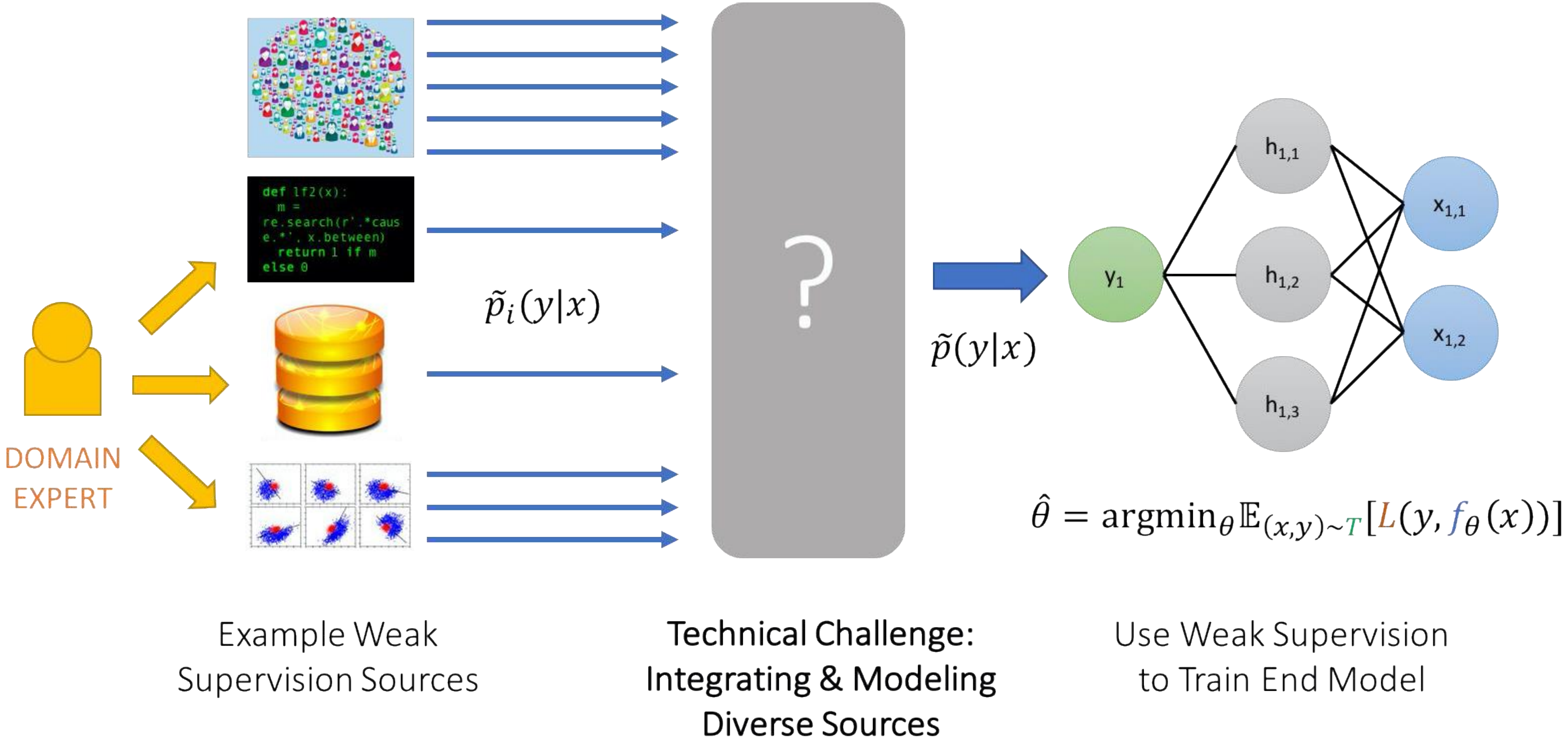
(Larry Page, Microsoft)
Label: NO_RELATION
Feature: X took a swipe at Y

(Larry Page, Harvard)
Label: NO_RELATION
Feature: Y invited X

(Bill Gates, Google)
Label: NO_RELATION
Feature: Y is X's worst fear

Credit: https://courses.cs.washington.edu/courses/cse517/13wi/slides/cse517wi13-RelationExtractionII.pdf

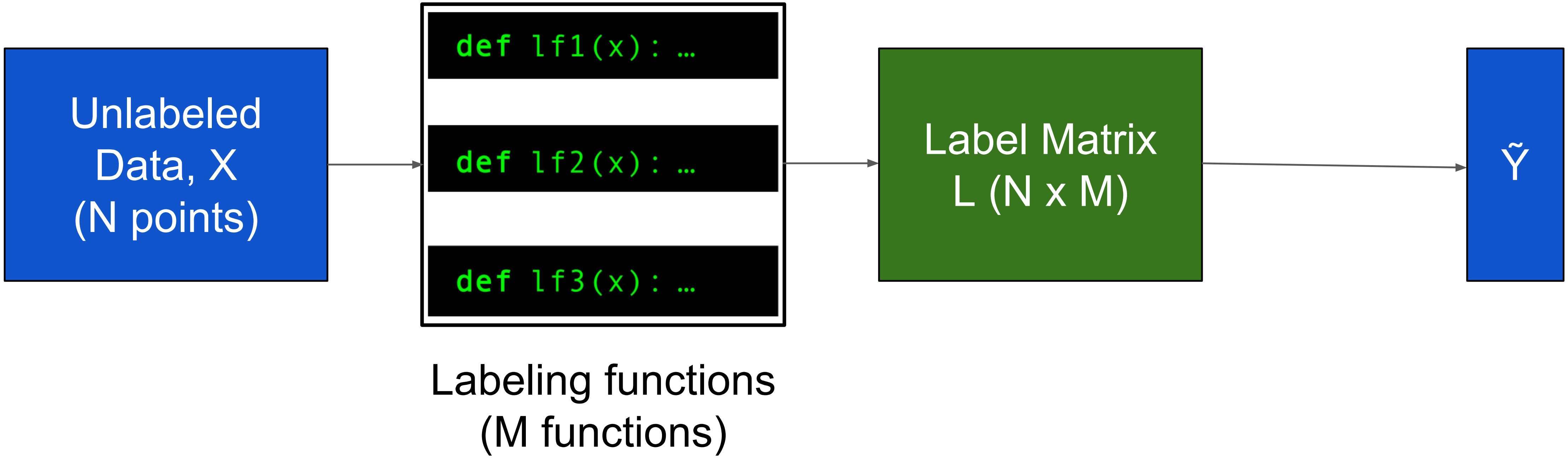
Integrating multiple noisy labels



Source: A. Ratner et. al https://dawn.cs.stanford.edu/2017/07/16/weak-supervision/

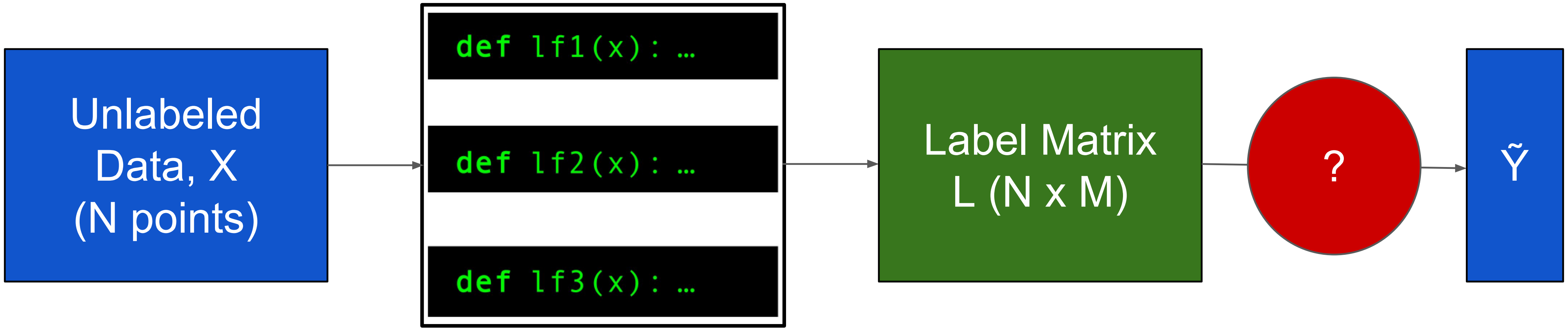
[Credit: http://cs231n.stanford.edu/slides/2018/cs231n_2018_ds07.pdf]

Integrating multiple noisy labels



[Credit: http://cs231n.stanford.edu/slides/2018/cs231n_2018_ds07.pdf]

Integrating multiple noisy labels



Labeling functions
(M functions)

[Credit: http://cs231n.stanford.edu/slides/2018/cs231n_2018_ds07.pdf]

Integrating multiple noisy labels

How do we obtain probabilistic labels, $\tilde{\mathbf{Y}}$, from the label matrix, $\mathbf{L}$?

Approach 1 - Majority Vote

Take the majority vote of the labelling functions (LFs).

Let's say $\mathbf{L}$ = [[0, 1, 0, 1, 0]; [1, 1, 1, 1, 0]].

$$\tilde{\mathbf{Y}} = [0, 1]$$

[Credit: http://cs231n.stanford.edu/slides/2018/cs231n_2018_ds07.pdf]

Integrating multiple noisy labels

How do we obtain probabilistic labels, $\tilde{\mathbf{Y}}$, from the label matrix, $\mathbf{L}$?

Approach 1 - Majority Vote

Majority vote fails:

Indication: Chest pain. Findings: Mediastinal contours are within normal limits. Heart size is within normal limits. No focal consolidation, pneumothorax or pleural effusion. Impression: No acute cardiopulmonary abnormality.

Normal Report

```
def LF_pneumothorax(c):
  if re.search(r'pneumo.*', c.report.text):
      return "ABNORMAL"

def LF_pleural_effusion(c):
  if "pleural effusion" in c.report.text:
      return "ABNORMAL"

def LF_normal_report(c, thresh=2):
    if len(NORMAL_TERMS.intersection(c.
    report.words)) > thresh:
      return "NORMAL"
```

LFs

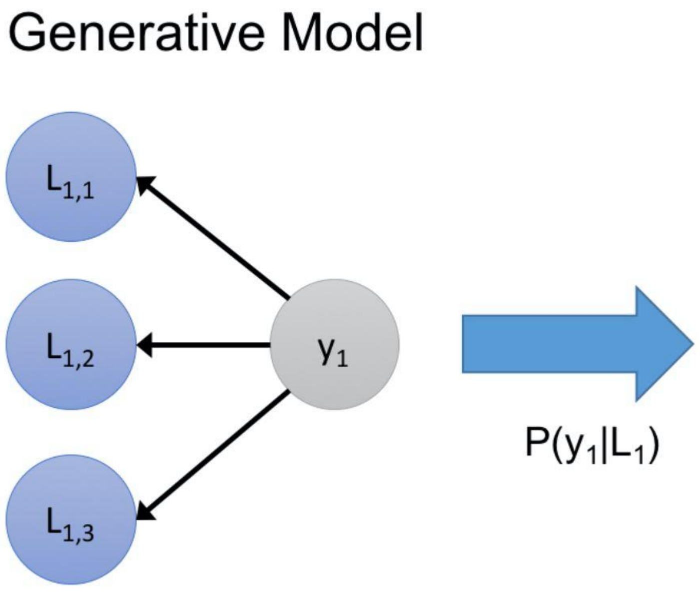
[Credit: http://cs231n.stanford.edu/slides/2018/cs231n_2018_ds07.pdf]

Integrating multiple noisy labels

How do we obtain probabilistic labels, $\tilde{\mathbf{Y}}$, from the label matrix, **L**?

Approach 2

Train a generative model over **P(L, Y)** where **Y** are the **(unknown)** true labels



[Credit: http://cs231n.stanford.edu/slides/2018/cs231n_2018_ds07.pdf]

Summary: Weak/distant supervision

- Noisy labels from heuristics, knowledge bases, constraints, …
- Integrating multiple noisy labels
  - Majority vote
  - Generative modeling
  - …

- Not all information/experiences can easily be converted into labels
  - “Every part of speech sequence should have a verb“
  - “In a sentence with word ‘but’, the sentiment of text after ‘but’ dominates”
  - “Every image patch that is recognized as a bicycle should have at least one patch that is recognized as a wheel”
  - I have a “discriminator” model that can tell me whether a model-generated image is good or not
- Need a more flexible framework to incorporate all forms of experience

Questions?