CSCE 636 Neural Networks (Deep Learning)

Lecture 18: Transfer Learning

Anxiao (Andrew) Jiang

Based on the interesting lecture of Prof. Hung-yi Lee "Transfer Learning"
https://www.youtube.com/watch?v=qD6iD4TFsdQ&list=PLJV_el3uVTsPy9oCRY30oBPNLCo89yu49&index=28

Transfer Learning

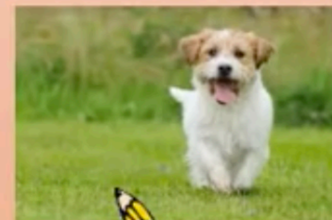
Transfer Learning

http://weebly110810.weebly.com/396403913129399.html

http://www.sucaitianxia.com/png/cartoon/200811/4261.html

Dog/Cat Classifier

cat



dog

Data ***<u>not directly related to</u>*** the task considered http://weebly110810.weebly.com/396403913129399.html

http://www.sucaitianxia.com/png/cartoon/200811/4261.html

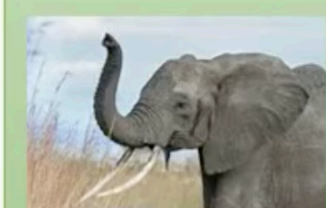
Transfer Learning

Dog/Cat Classifier

cat

dog

Data ***not directly related to*** the task considered



elephant

tiger

Similar domain, different task

Transfer Learning

http://weebly110810.weebly.com/396403913129399.html

http://www.sucaitianxia.com/png/cartoon/200811/4261.html

Dog/Cat
Classifier

Data ***not directly related to*** the task considered

Similar domain, different task

Different domain, same task

Why?

http://www.bigr.nl/website/structure/main.php?page=researchlines&subpage=project&id=64

http://www.spear.com.hk/Translation-company-Directory.html

	Task Considered	Data not directly related
Speech Recognition	 Taiwanese	

Why?

http://www.bigr.nl/website/structure/main.php?page=resear
chlines&subpage=project&id=64

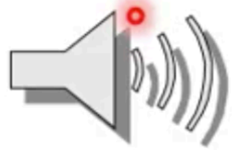
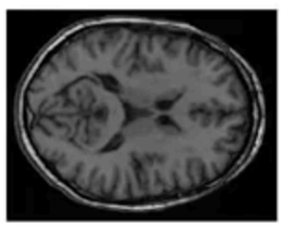
http://www.spear.com.hk/Translation-company-Directory.html

	Task Considered	Data not directly related
Speech Recognition	 Taiwanese	English Chinese ……

Why?

http://www.bigr.nl/website/structure/main.php?page=researchlines&subpage=project&id=64

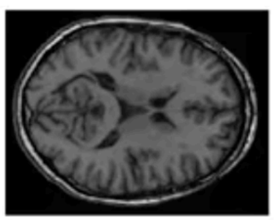
http://www.spear.com.hk/Translation-company-Directory.html

	Task Considered	Data not directly related
Speech Recognition	 Taiwanese	YouTube English Chinese ……
Image Recognition	 Medical Images	

Why?

http://www.bigr.nl/website/structure/main.php?page=researchlines&subpage=project&id=64

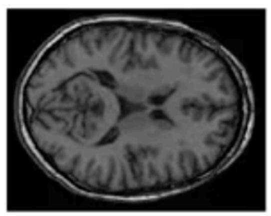
http://www.spear.com.hk/Translation-company-Directory.html

	Task Considered	Data not directly related
Speech Recognition	 Taiwanese	YouTube English Chinese ……
Image Recognition	 Medical Images	

Why?

http://www.bigr.nl/website/structure/main.php?page=researchlines&subpage=project&id=64

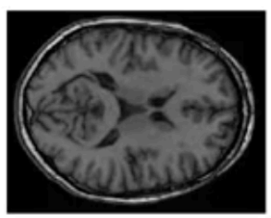
http://www.spear.com.hk/Translation-company-Directory.html

	Task Considered	Data not directly related
Speech Recognition	 Taiwanese	English Chinese ……
Image Recognition	 Medical Images	
Text Analysis	 Specific domain	

Why?

http://www.bigr.nl/website/structure/main.php?page=researchlines&subpage=project&id=64

http://www.spear.com.hk/Translation-company-Directory.html

	Task Considered		Data not directly related
Speech Recognition		Taiwanese	YouTube English Chinese ……
Image Recognition		Medical Images	
Text Analysis		Specific domain	

Transfer Learning

- Example in real life

We do it all the time.

Transfer Learning - Overview

		Source Data (not directly related to the task)	
		labelled	unlabeled
Target Data	labelled		
	unlabeled		

Warning: different terminology in different literature

Transfer Learning - Overview

		Source Data (not directly related to the task)	
		labelled	unlabeled
Target Data	labelled	Model Fine-tuning	
	unlabeled		

Warning: different terminology in different literature

Model Fine-tuning

- Task description
  - Target data: (x^t, y^t)
  - Source data: (x^s, y^s)

Model Fine-tuning

- Task description
  - Target data: (x^t, y^t) ⬅ Very little
  - Source data: (x^s, y^s) ⬅ A large amount

Model Fine-tuning

One-shot learning: only a few examples in target domain

- Task description
  - Target data: (x^t, y^t) ← Very little
  - Source data: (x^s, y^s) ← A large amount

Model Fine-tuning

One-shot learning: only a few examples in target domain

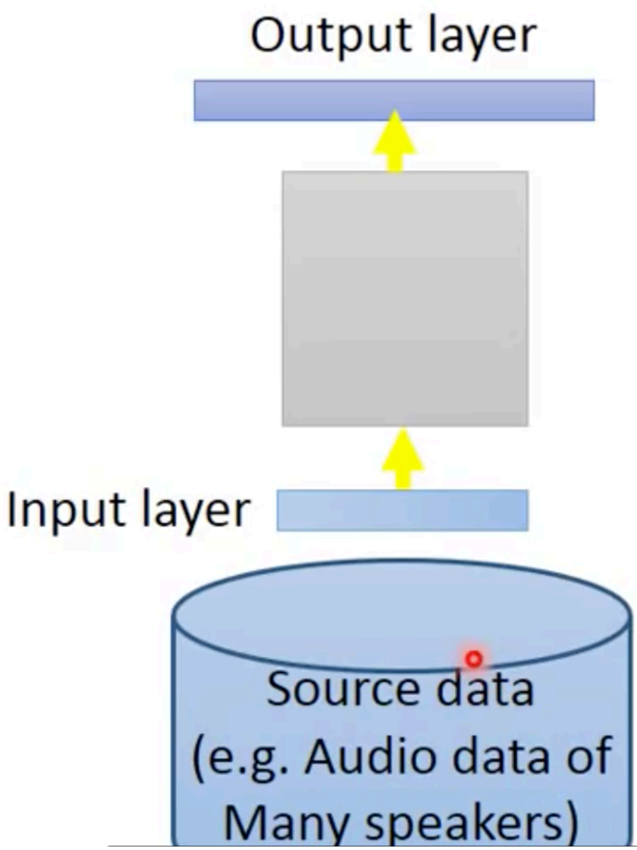
- Task description
  - Target data: (x^t, y^t) ⬅ Very little
  - Source data: (x^s, y^s) ⬅ A large amount
- Example: (supervised) speaker adaption
  - Target data: audio data and its transcriptions of specific user
  - Source data: audio data and transcriptions from many speakers

Model Fine-tuning

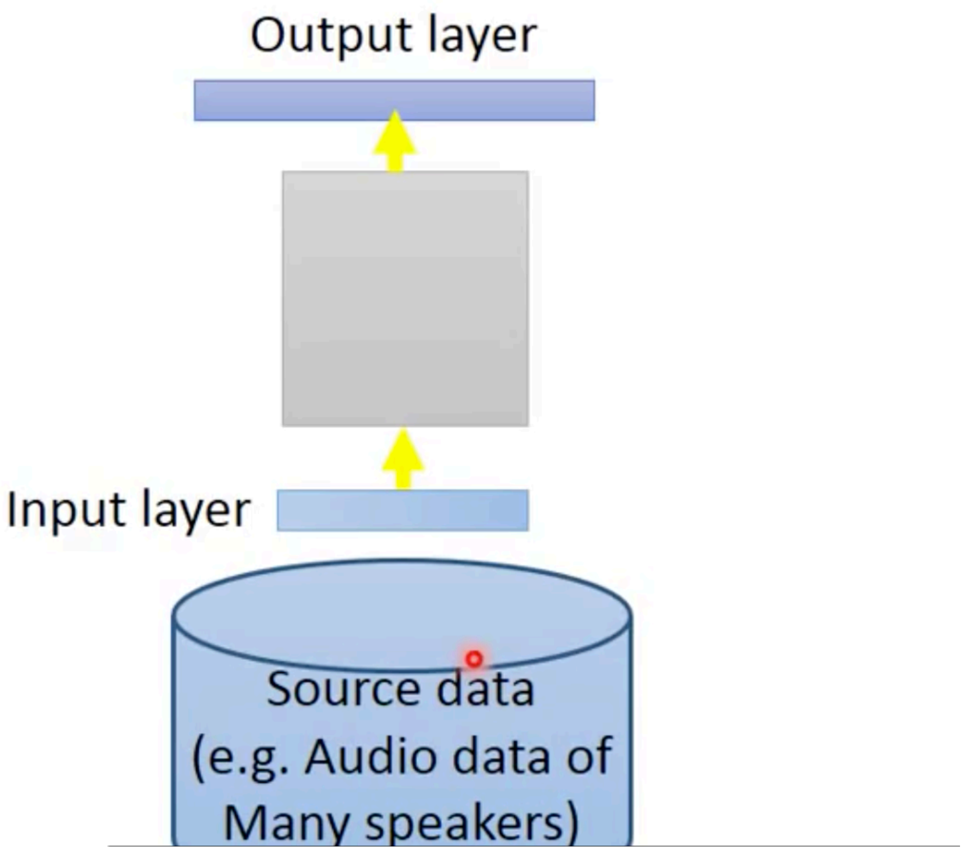
One-shot learning: only a few examples in target domain

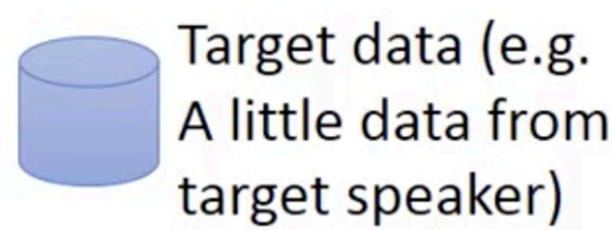
- Task description
  - Target data: (x^t, y^t) ⬅ Very little
  - Source data: (x^s, y^s) ⬅ A large amount
- Example: (supervised) speaker adaption
  - Target data: audio data and its transcriptions of specific user
  - Source data: audio data and transcriptions from many speakers
- Idea: training a model by source data, then fine-tune the model by target data
  - Challenge: only limited target data, so be careful about overfitting

Conservative Training

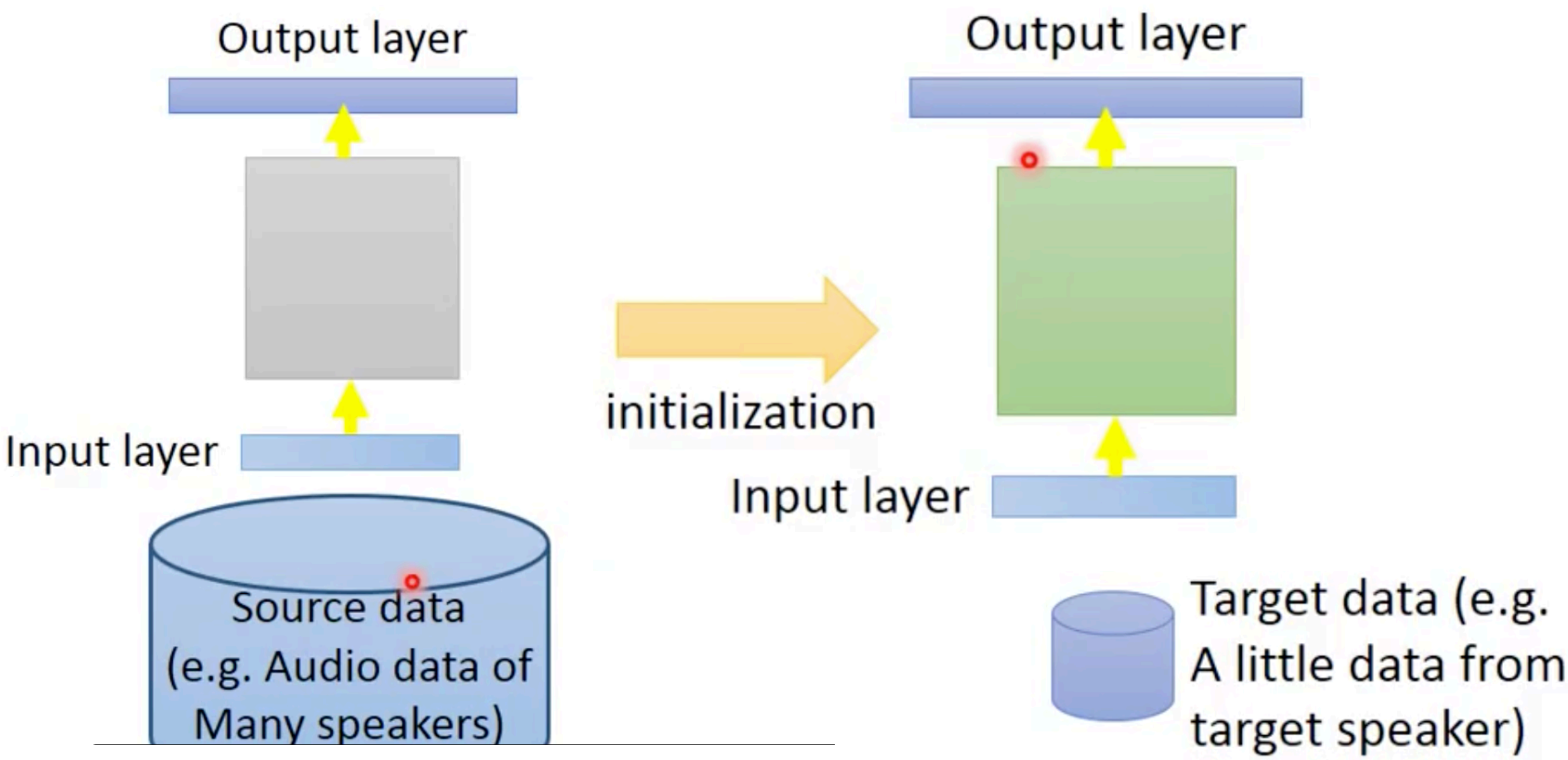


Conservative Training

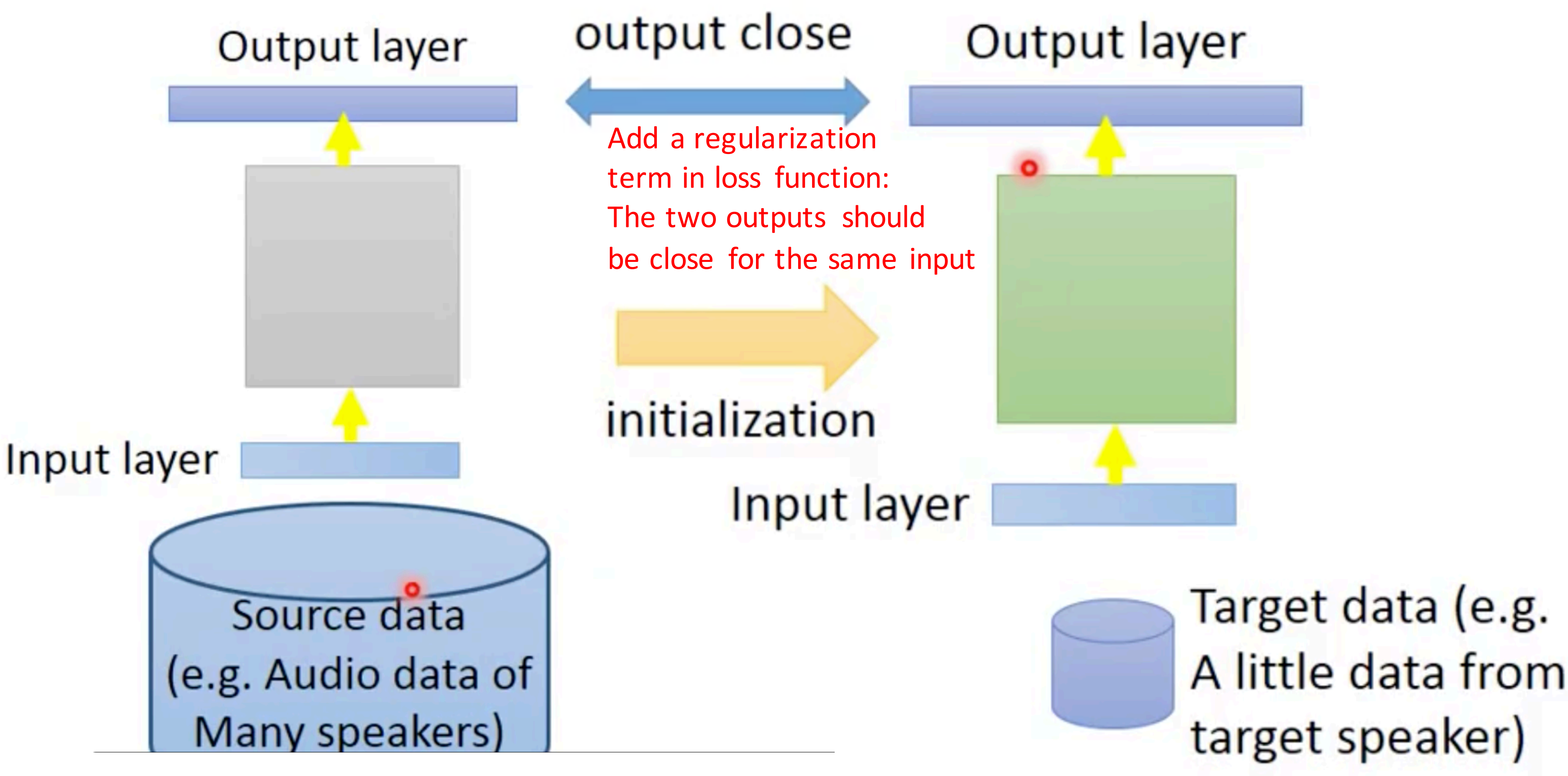




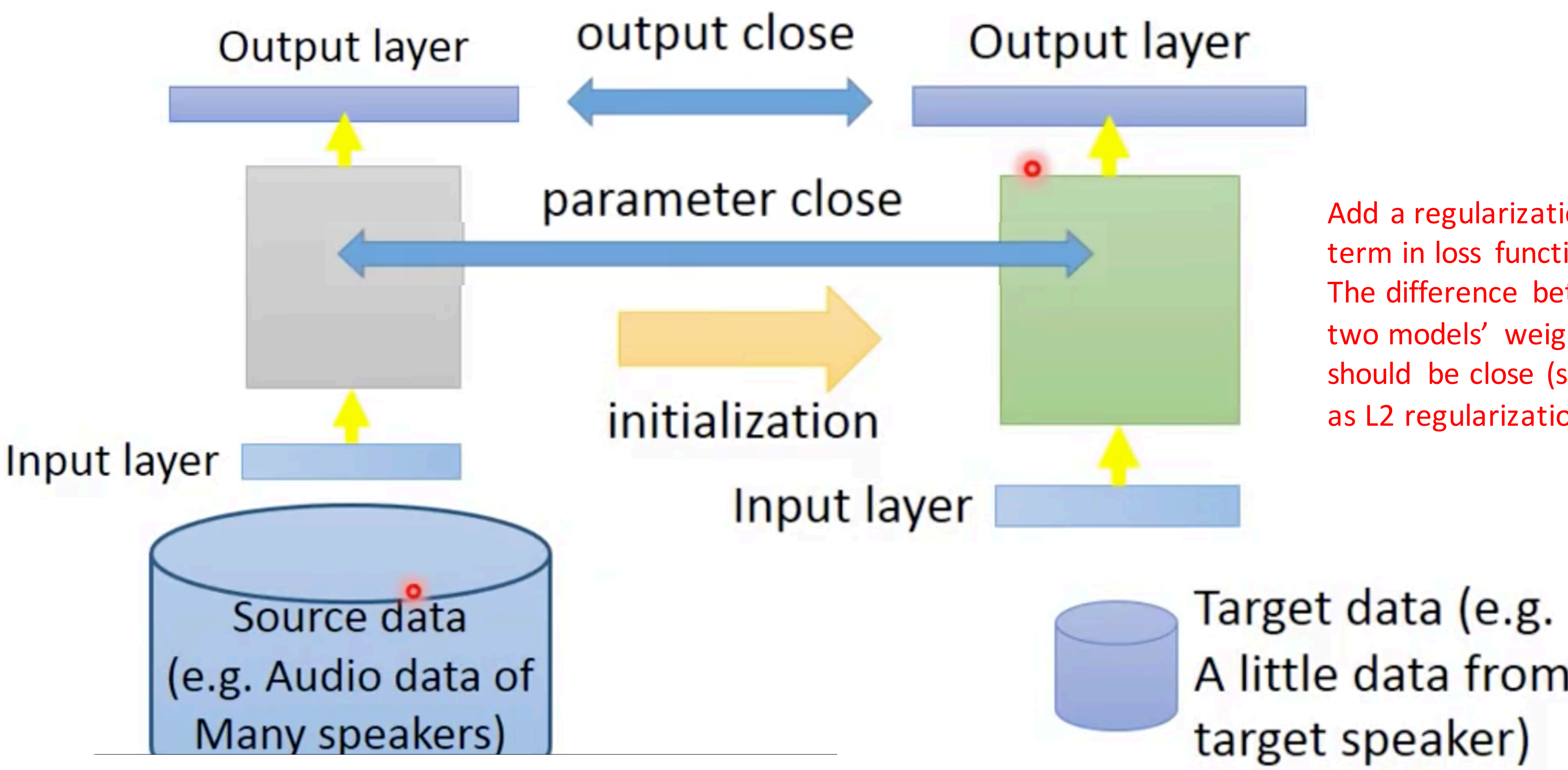
Conservative Training



Conservative Training

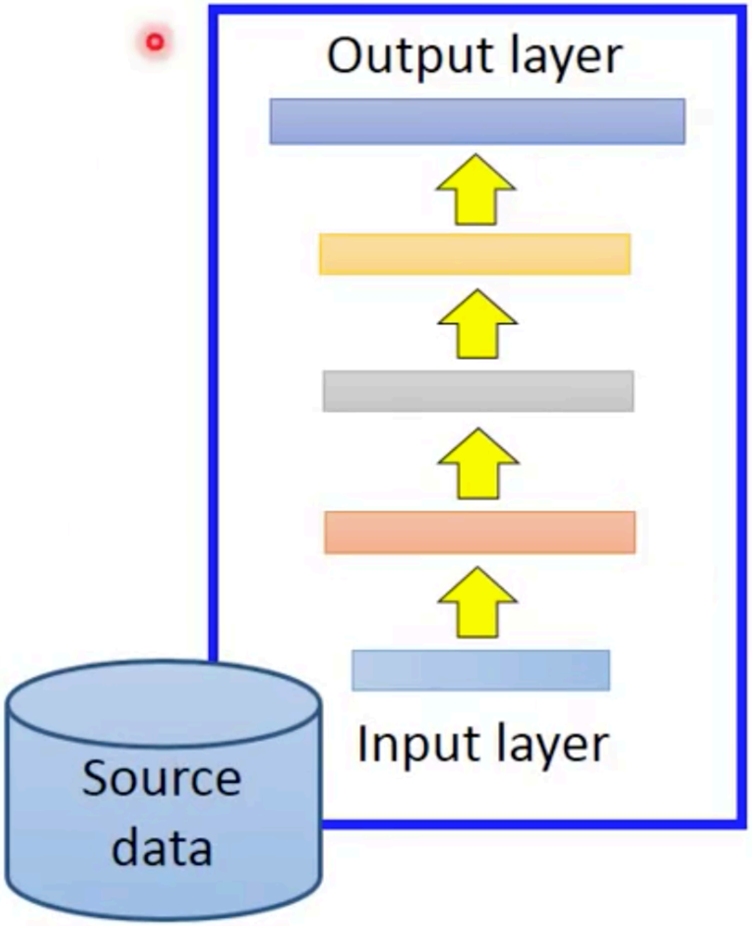


Conservative Training

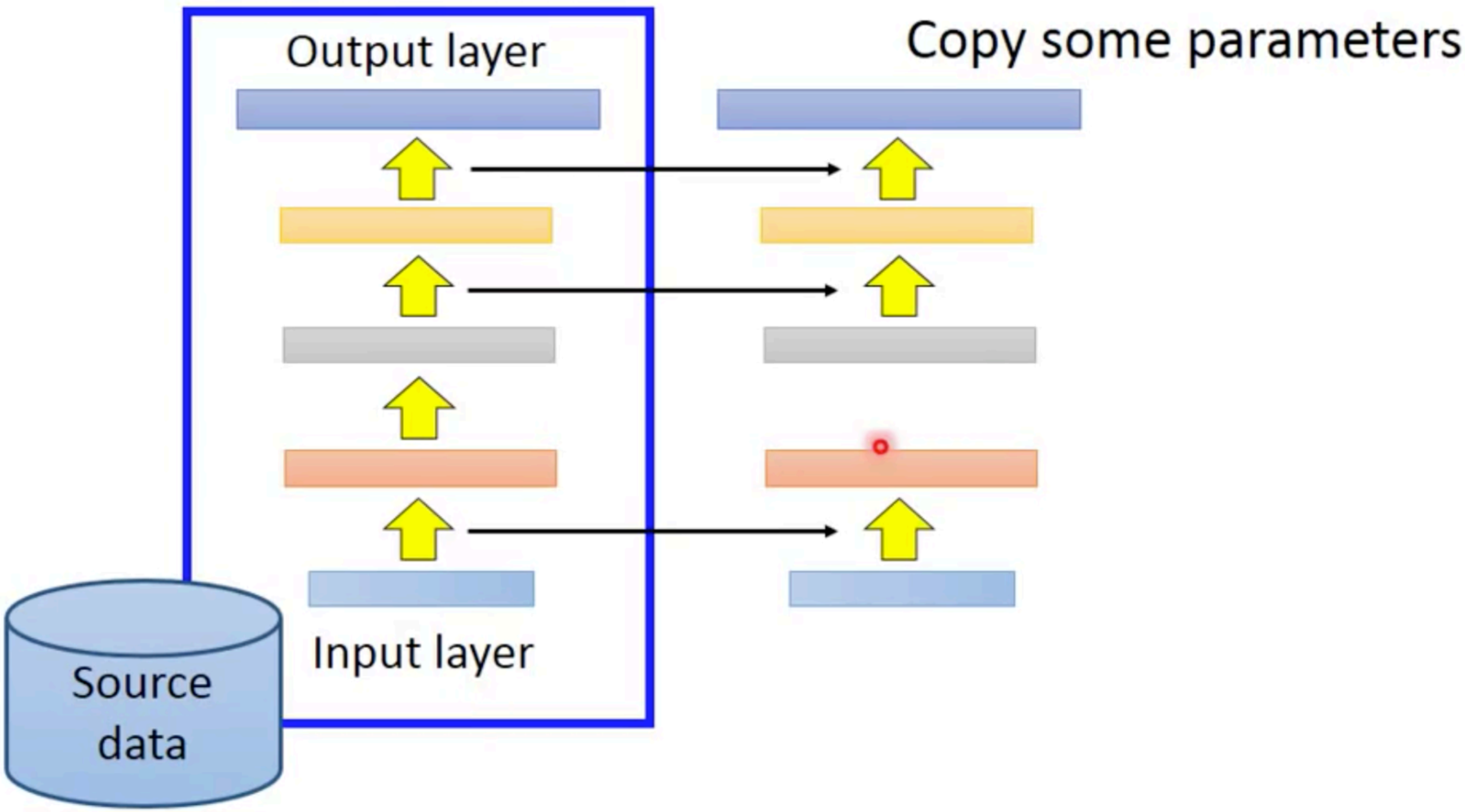


Add a regularization term in loss function: The difference between two models' weights should be close (such as L2 regularization)

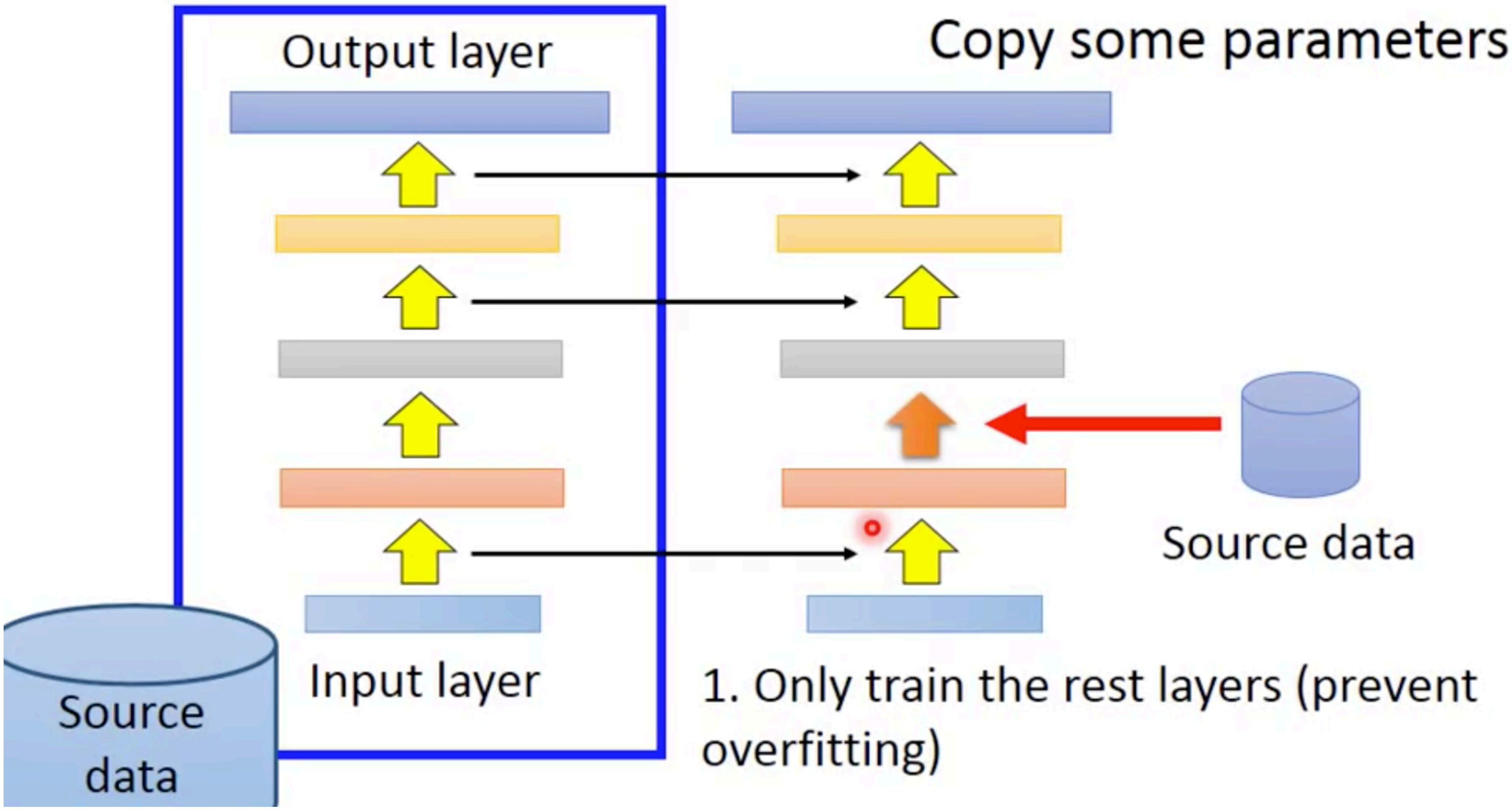
Layer Transfer



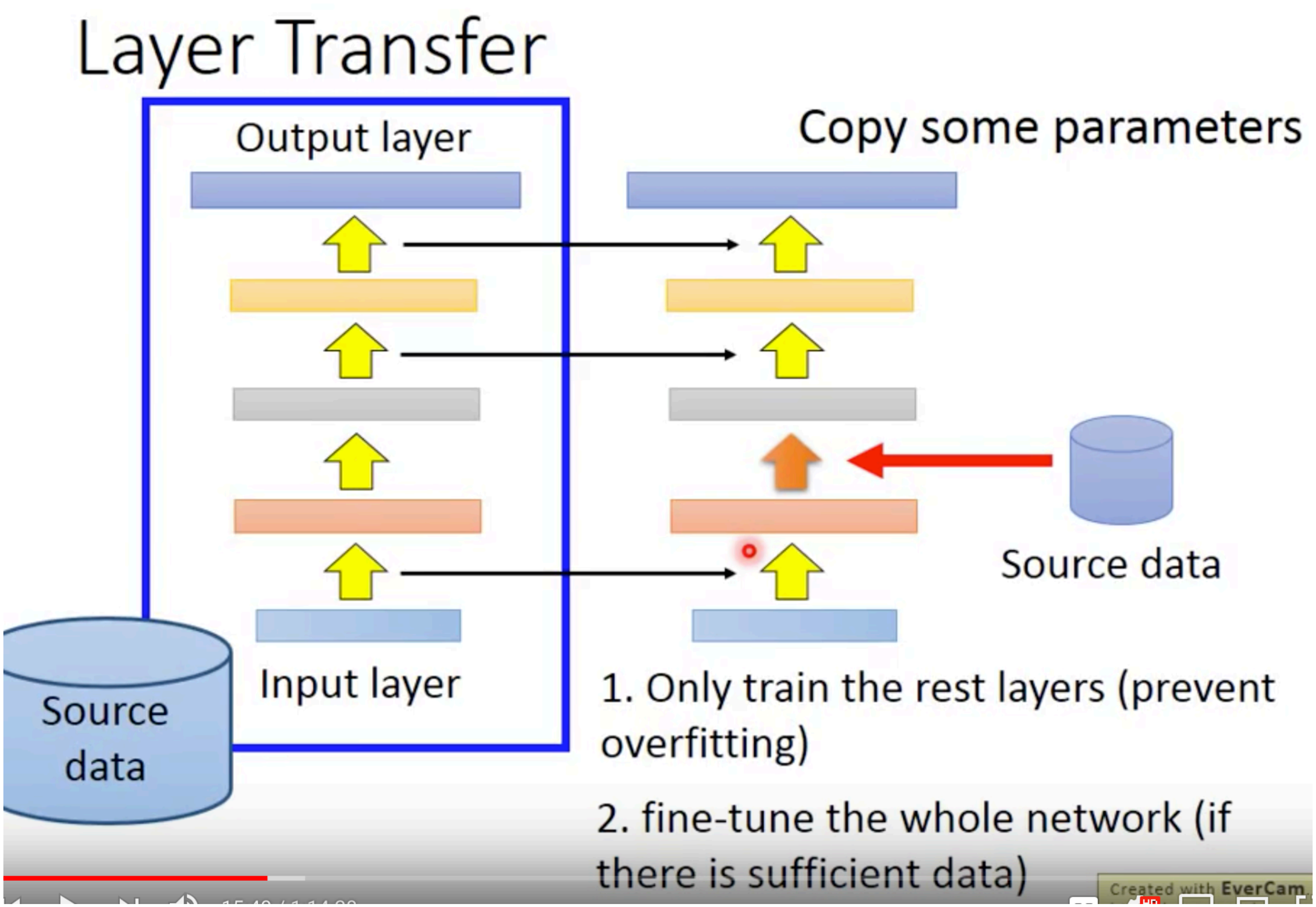
Layer Transfer



Layer Transfer



1. Only train the rest layers (prevent overfitting)


Layer Transfer
Output layer
Copy some parameters
Source data
Input layer
Source
data
1. Only train the rest layers (prevent overfitting)
2. fine-tune the whole network (if there is sufficient data)
Created with EverCam

Layer Transfer

- Which layer can be transferred (copied)?

It depends on the application (task).

Layer Transfer

- Which layer can be transferred (copied)?
  - Speech: usually copy the last few layers

Because in speech recognition, the last few layers are about the meaning/words of the sentences, which are the same for different speakers; but the first few layers are (more) about the different voices of speakers.

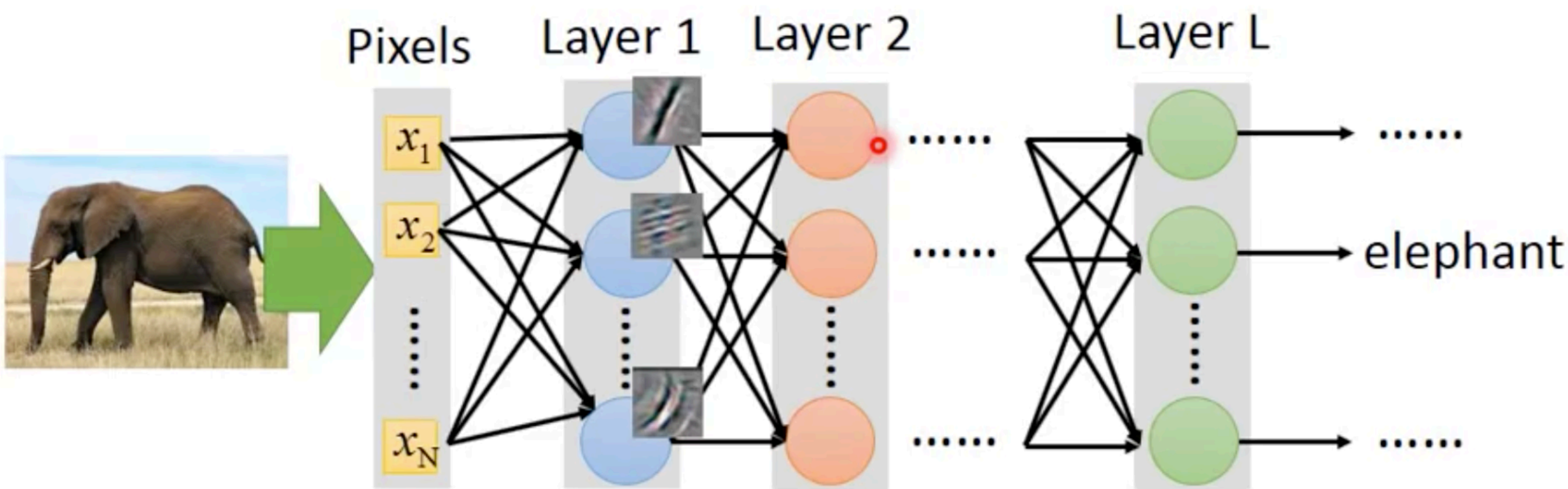
Layer Transfer

- Which layer can be transferred (copied)?
  - Speech: usually copy the last few layers
  - Image: usually copy the first few layers

Because the first few layers are about small features
(such as lines, corners, etc.), which exists in all types of images;
But the last few layers are about large features
(such as faces, wheels, etc.), which only apply to specific type of images
and tasks.

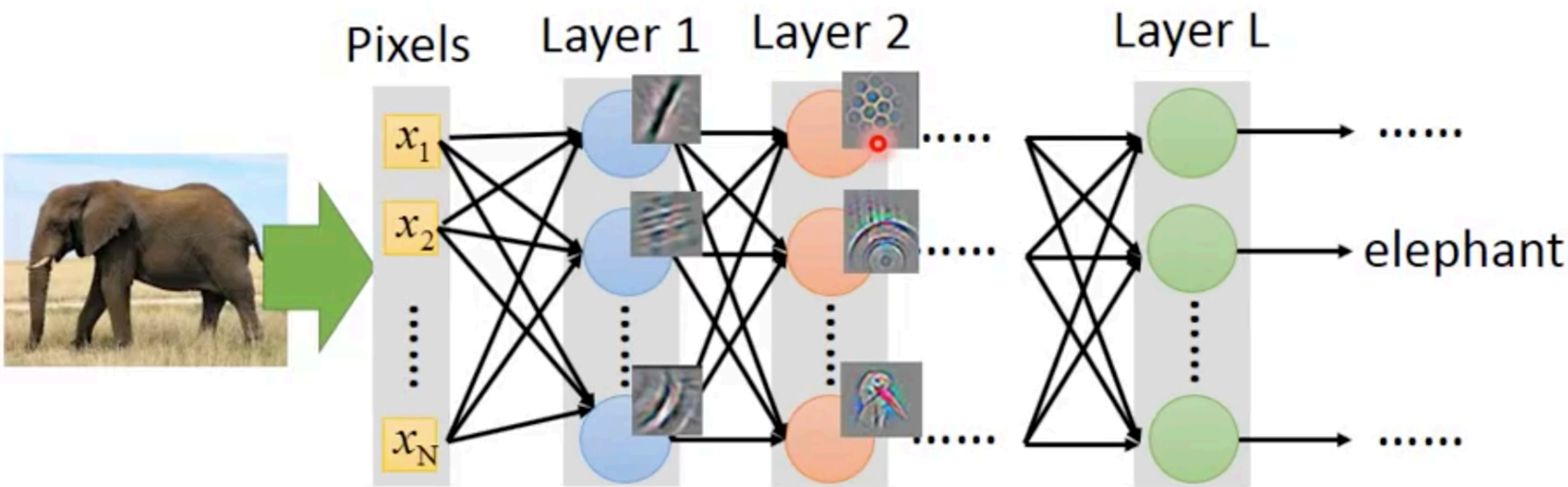
Layer Transfer

- Which layer can be transferred (copied)?
  - Speech: usually copy the last few layers
  - Image: usually copy the first few layers



Layer Transfer

- Which layer can be transferred (copied)?
  - Speech: usually copy the last few layers
  - Image: usually copy the first few layers



Transfer Learning - Overview

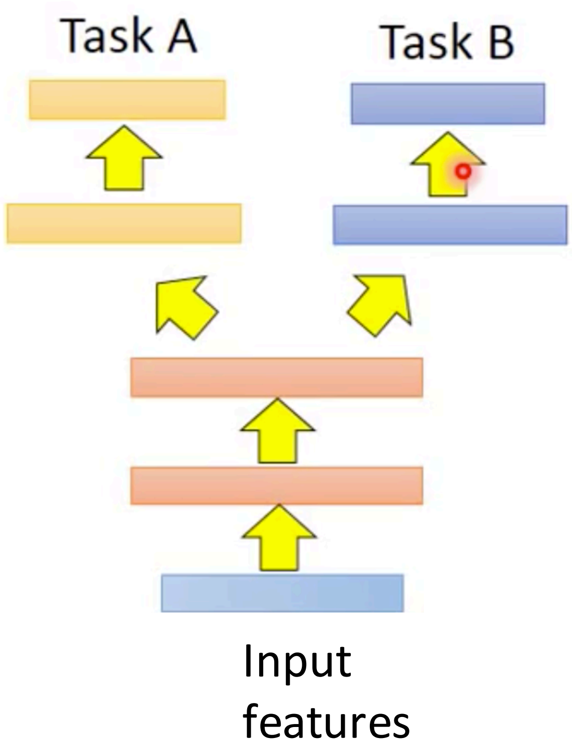
		Source Data (not directly related to the task)	
		labelled	unlabeled
Target Data	labelled	Fine-tuning Multitask Learning	
	unlabeled		

Multitask Learning

- The multi-layer structure makes NN suitable for multitask learning

Multitask Learning

- The multi-layer structure makes NN suitable for multitask learning



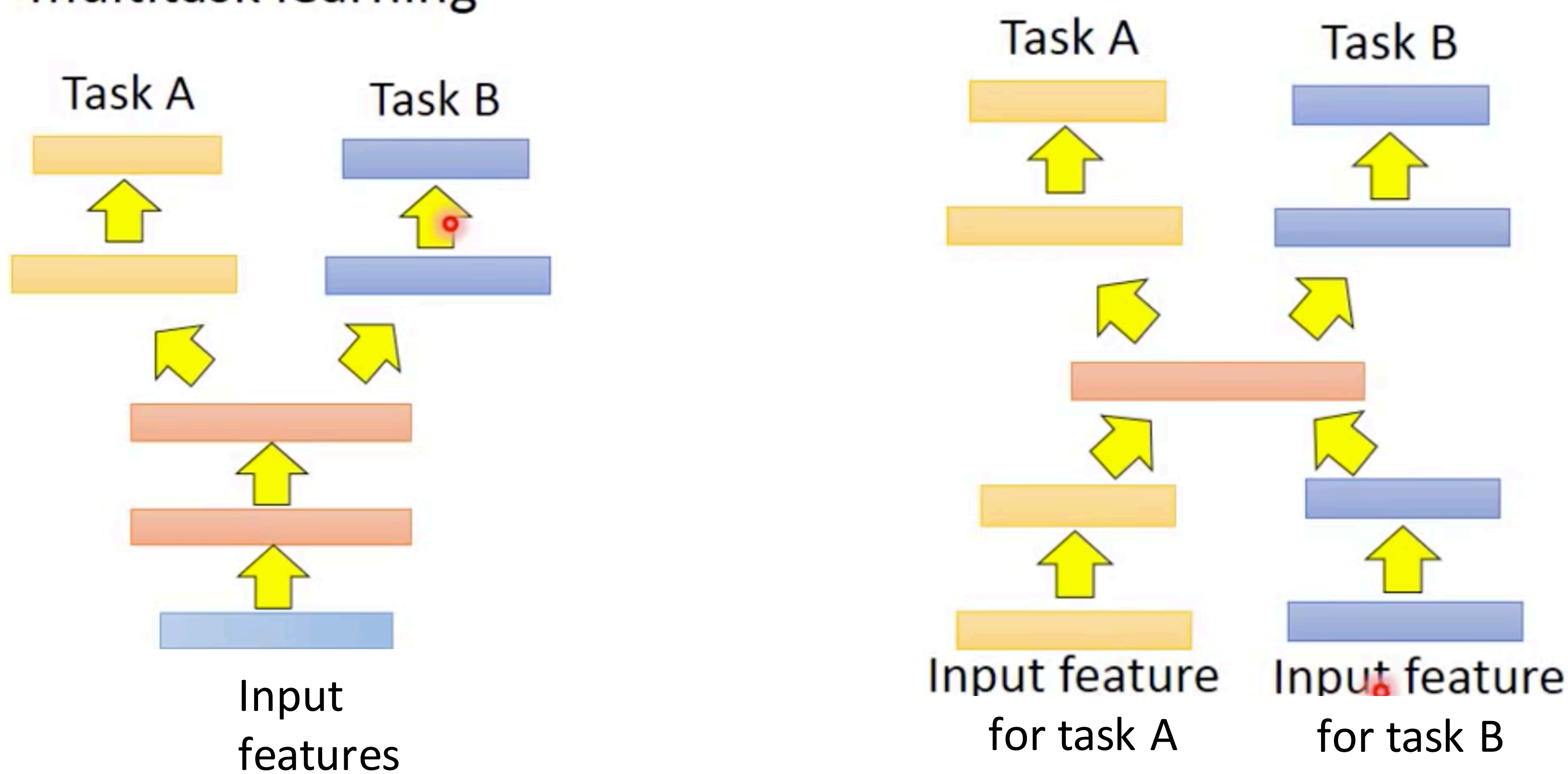
Example:

task A: classify ImageNet images

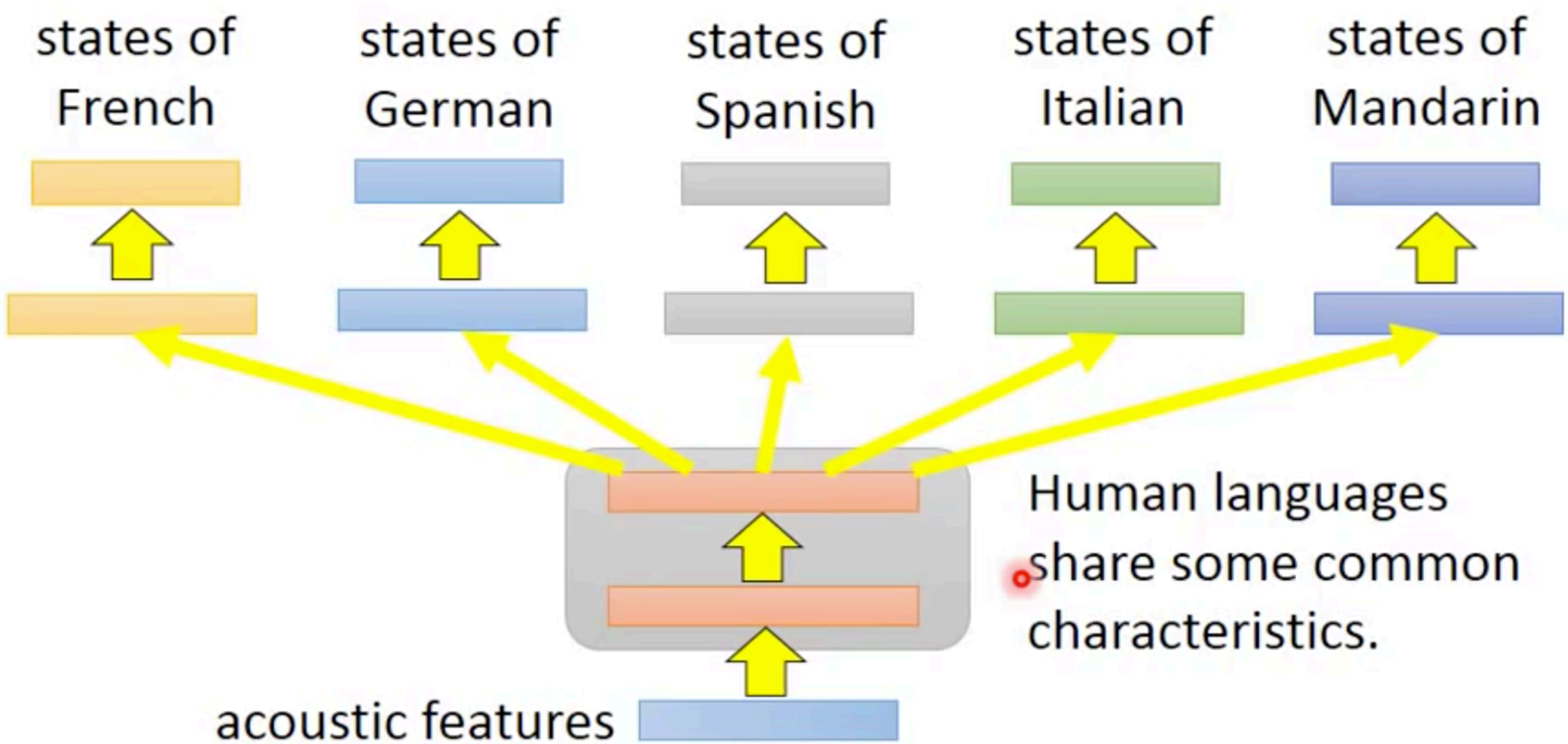
task B: classify medical images

Multitask Learning

- The multi-layer structure makes NN suitable for multitask learning



Multitask Learning
- Multilingual Speech Recognition



Human languages share some common characteristics.

Progressive Neural Networks

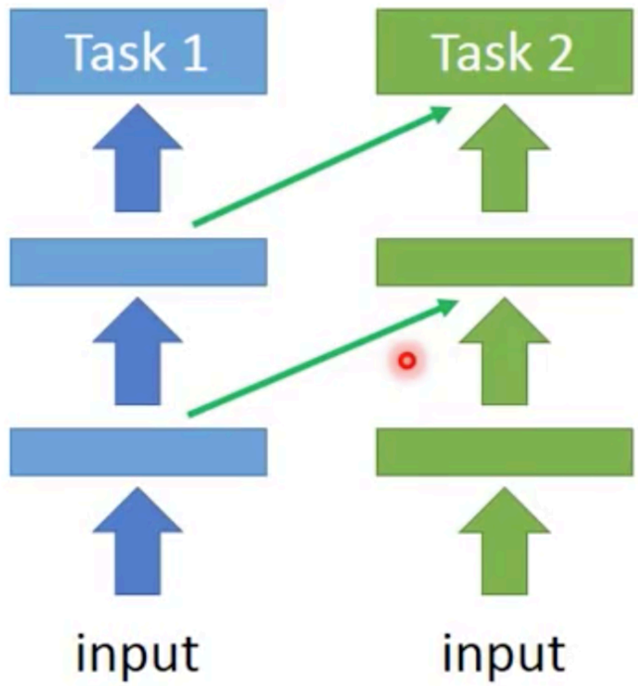
If the two tasks are actually different and cannot share common layers, transfer learning may degrade the performance for both tasks.

Too much “trial-and-error” can be a waste of time.

Progressive neural networks is an approach for such “uncertain” cases.

Progressive Neural Networks

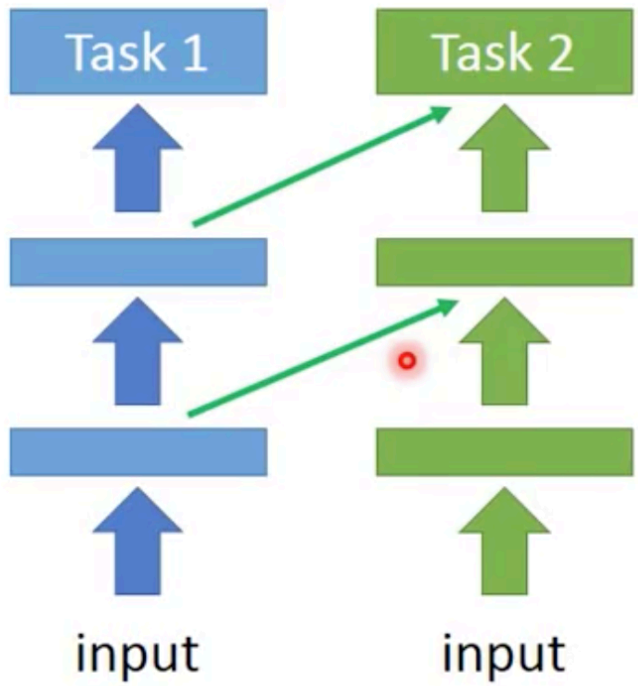
Frist, train network for task 1, then fix its weights.



Andrei A. Rusu, Neil C. Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, Raia Hadsell, "Progressive Neural Networks", arXiv preprint 2016

Progressive Neural Networks

Frist, train network
for task 1, then fix
its weights.



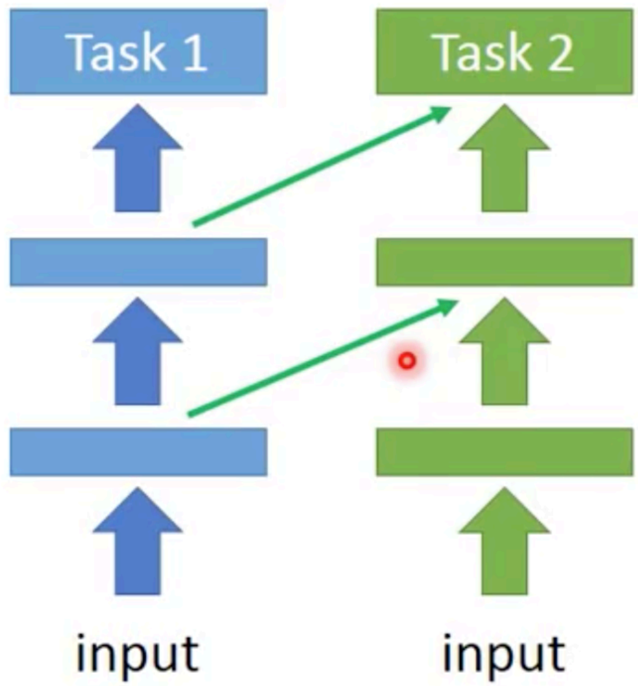
Next, train network
for task 2, and connect
cach layer of the first network
to the second network.

Andrei A. Rusu, Neil C. Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, Raia Hadsell, "Progressive Neural Networks", arXiv preprint 2016

Progressive Neural Networks

The training for task 2 will not degrade the performance for task 1.

Frist, train network for task 1, then fix its weights.



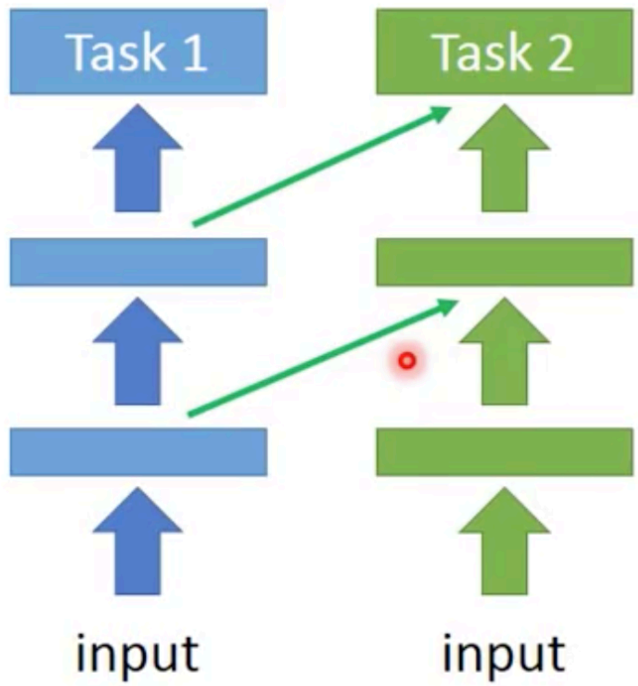
Next, train network for task 2, and connect cach layer of the first network to the second network.

Andrei A. Rusu, Neil C. Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, Raia Hadsell, "Progressive Neural Networks", arXiv preprint 2016

Progressive Neural Networks

When training the network for task 2, the weights from task 1 can be trained (changed) to be 0, thus not degrading the performance for task 2.

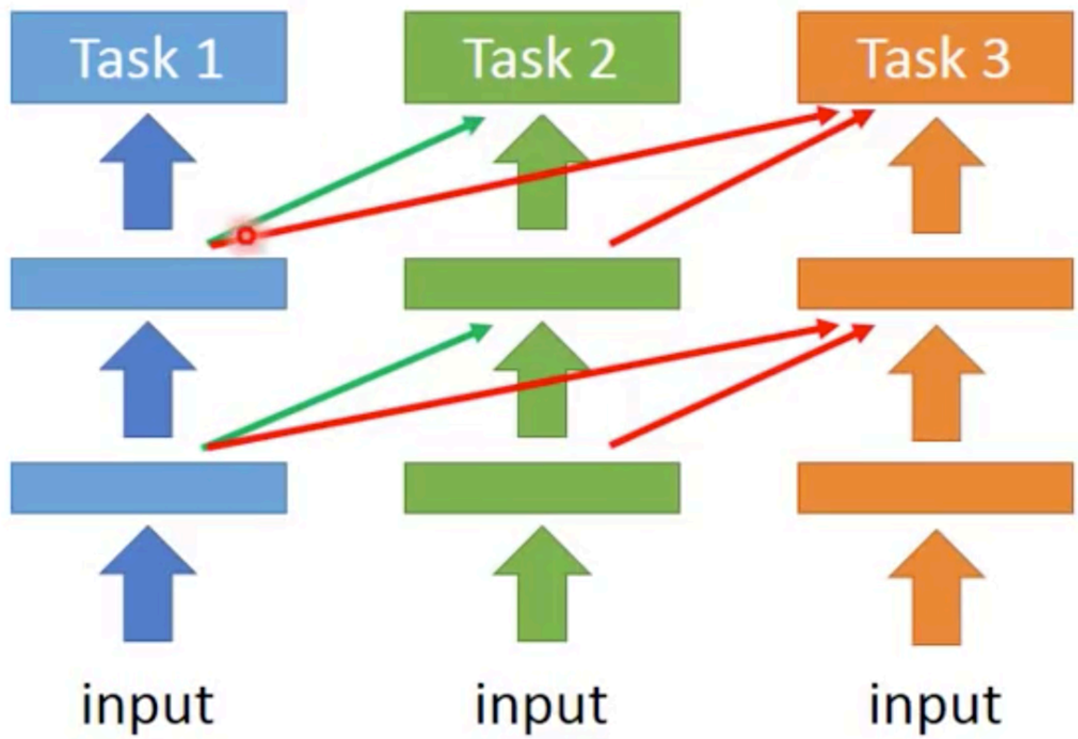
Frist, train network for task 1, then fix its weights.



Next, train network for task 2, and connect cach layer of the first network to the second network.

Andrei A. Rusu, Neil C. Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, Raia Hadsell, "Progressive Neural Networks", arXiv preprint 2016

Progressive Neural Networks



Transfer Learning - Overview

		Source Data (not directly related to the task)	
		labelled	unlabeled
Target Data	labelled	Fine-tuning Multitask Learning	
	unlabeled	Domain-adversarial training	

Task description

- Source data: (x^s, y^s)
- Target data: (x^t)

Task description

- Source data: (x^s, y^s)
- Target data: (x^t)

Task description

- Source data: (x^s, y^s) → Training data
- Target data: (x^t) → Testing data

MNIST

SOURCE — with label

TARGET — without label

MNIST-M

Task description

- Source data: (x^s, y^s) $\rightarrow$ Training data
- Target data: (x^t) $\rightarrow$ Testing data

mismatch

MNIST

SOURCE

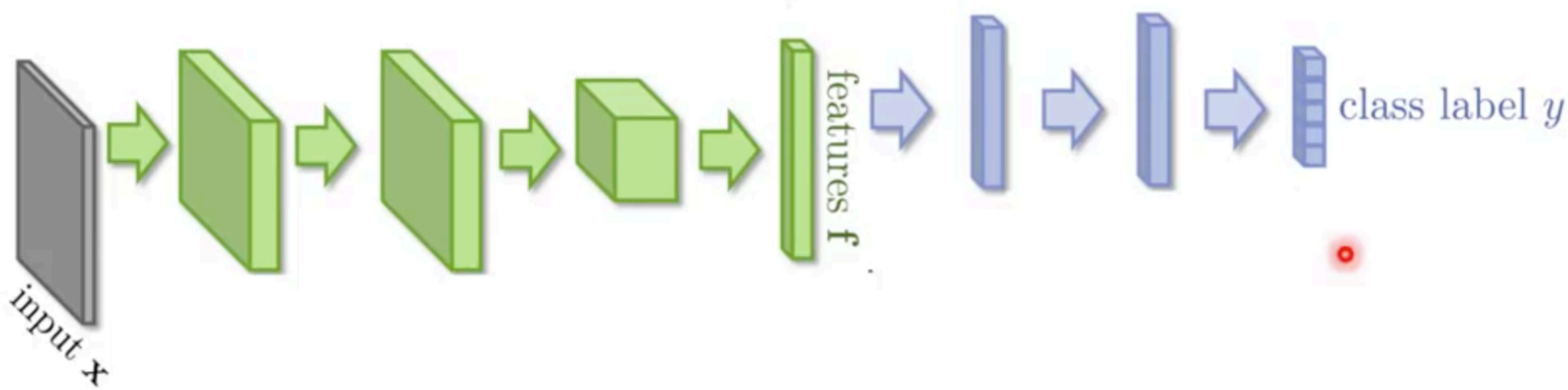
with label

TARGET

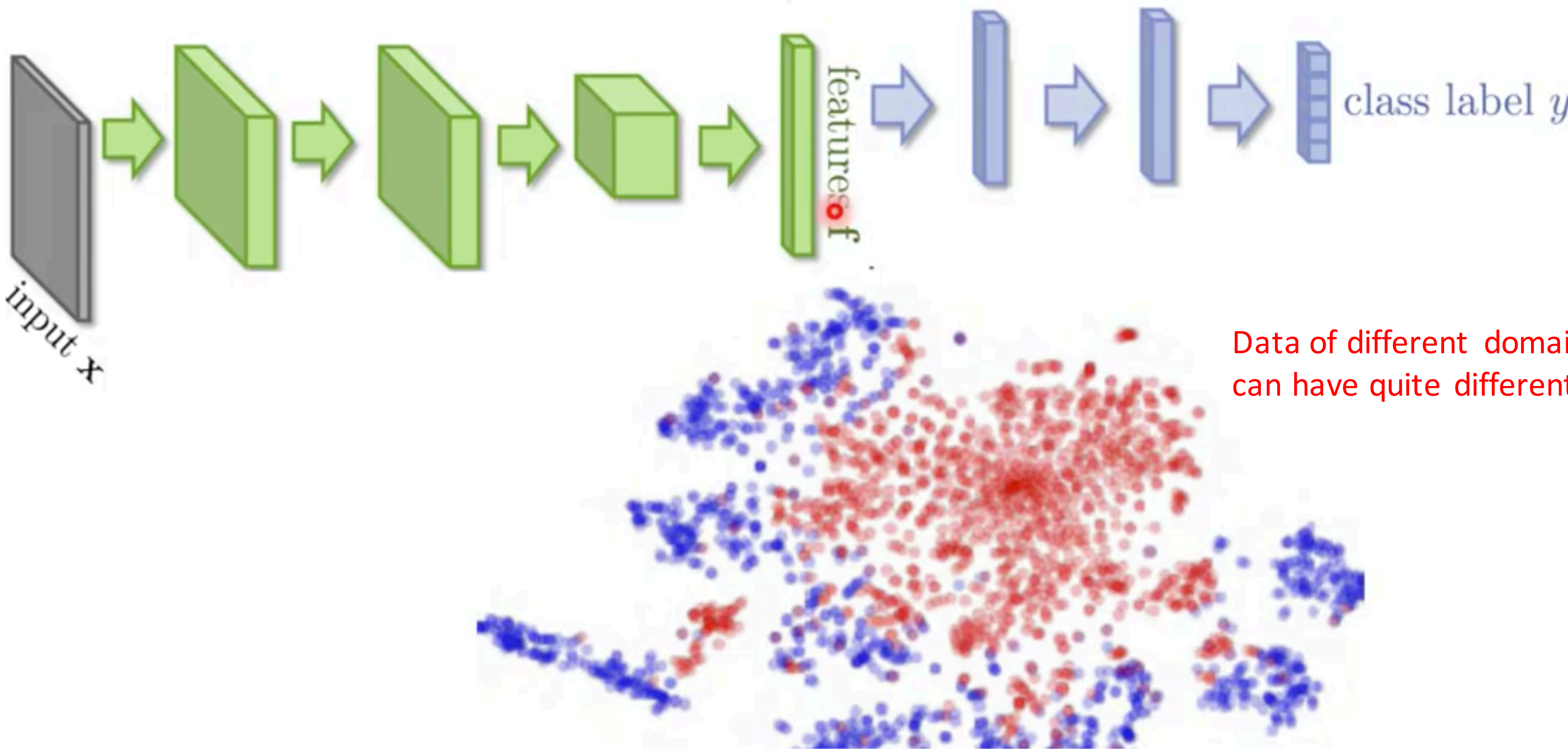
without label

MNIST-M

Domain-adversarial training

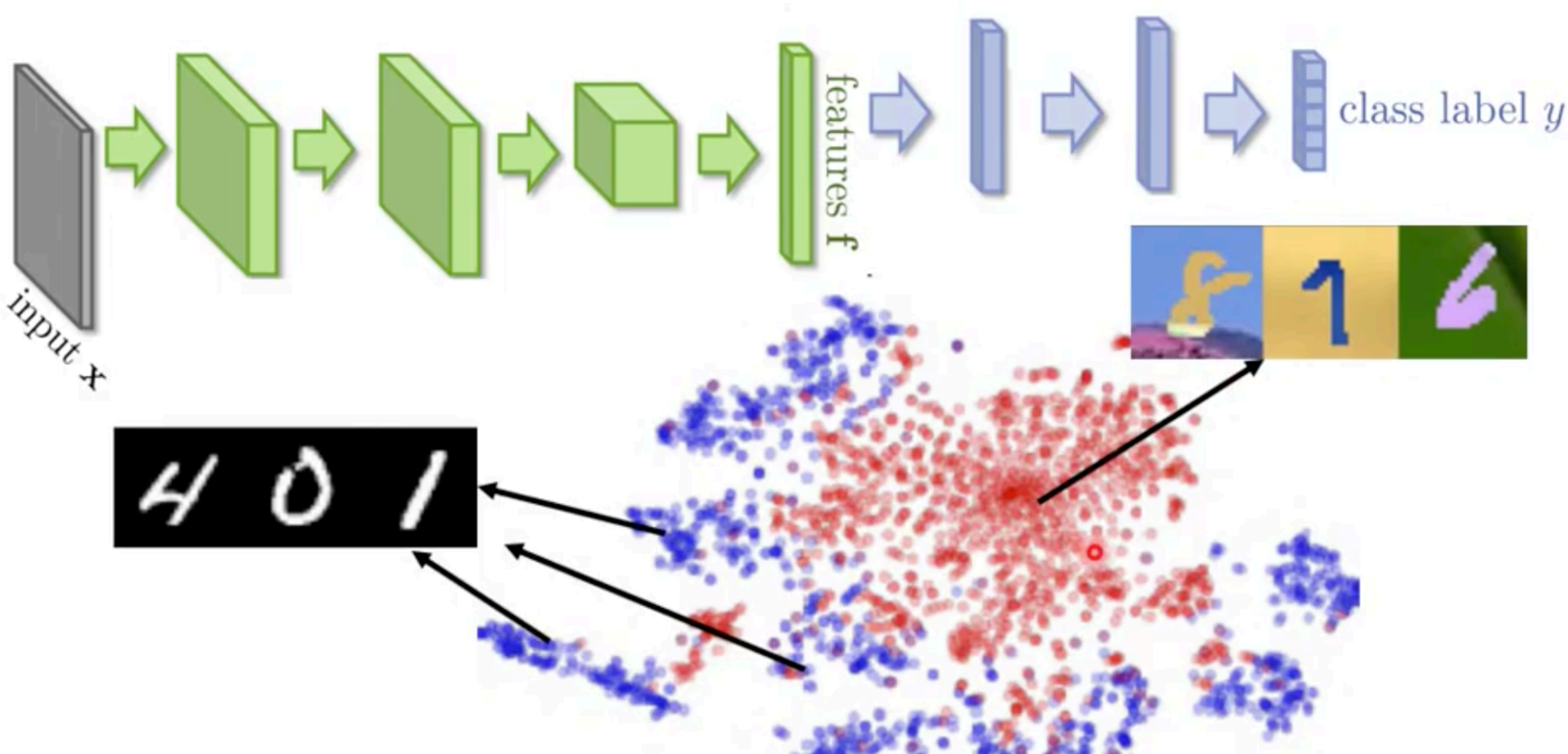


Domain-adversarial training



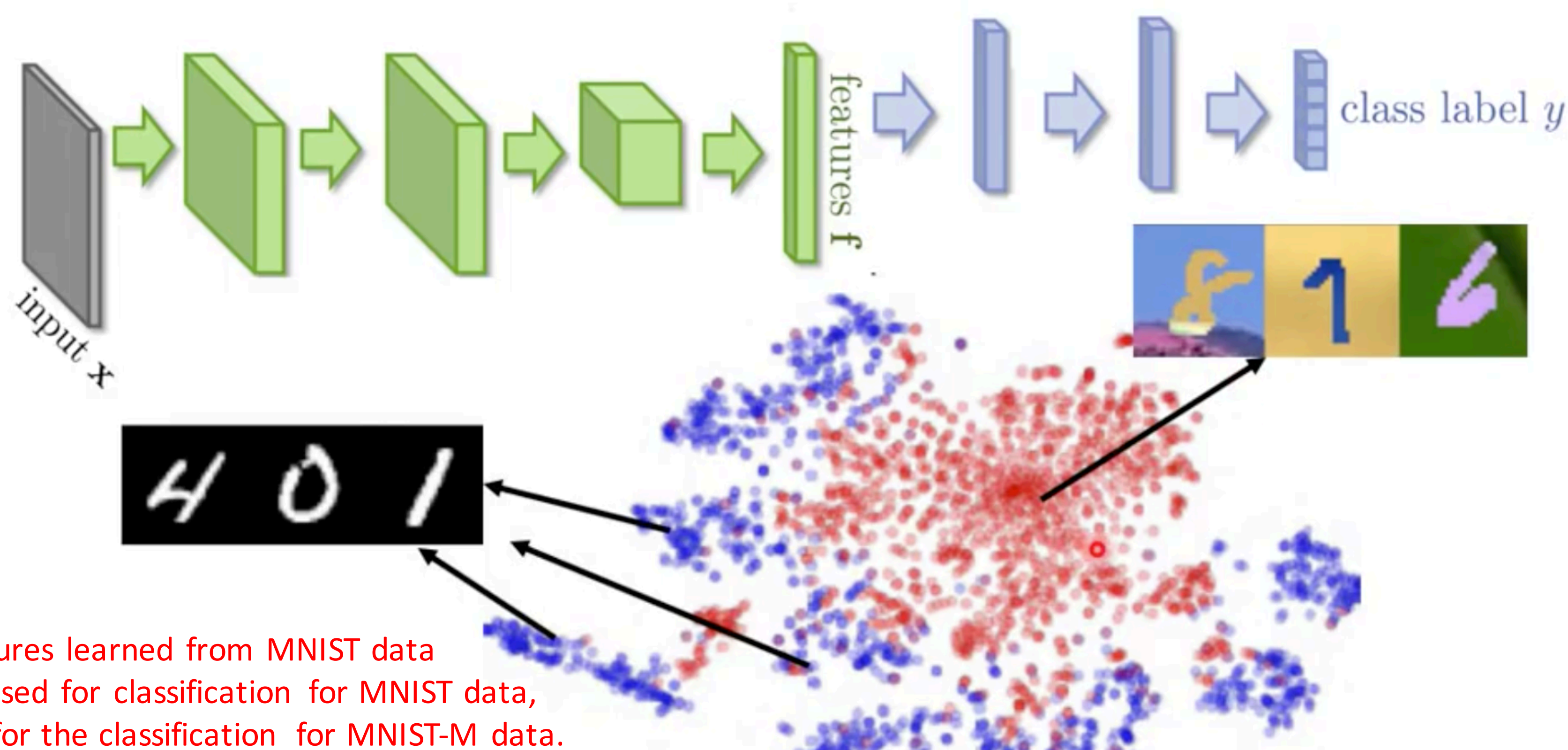
Data of different domains
can have quite different features

Domain-adversarial training



這邊有把 t-SNE 降維以後的結果

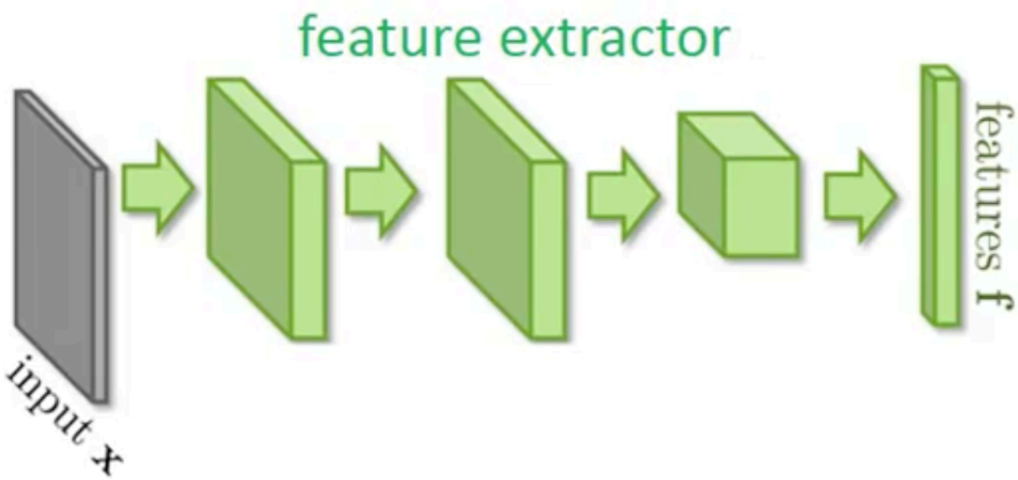
Domain-adversarial training



The features learned from MNIST data
Can be used for classification for MNIST data,
But not for the classification for MNIST-M data.

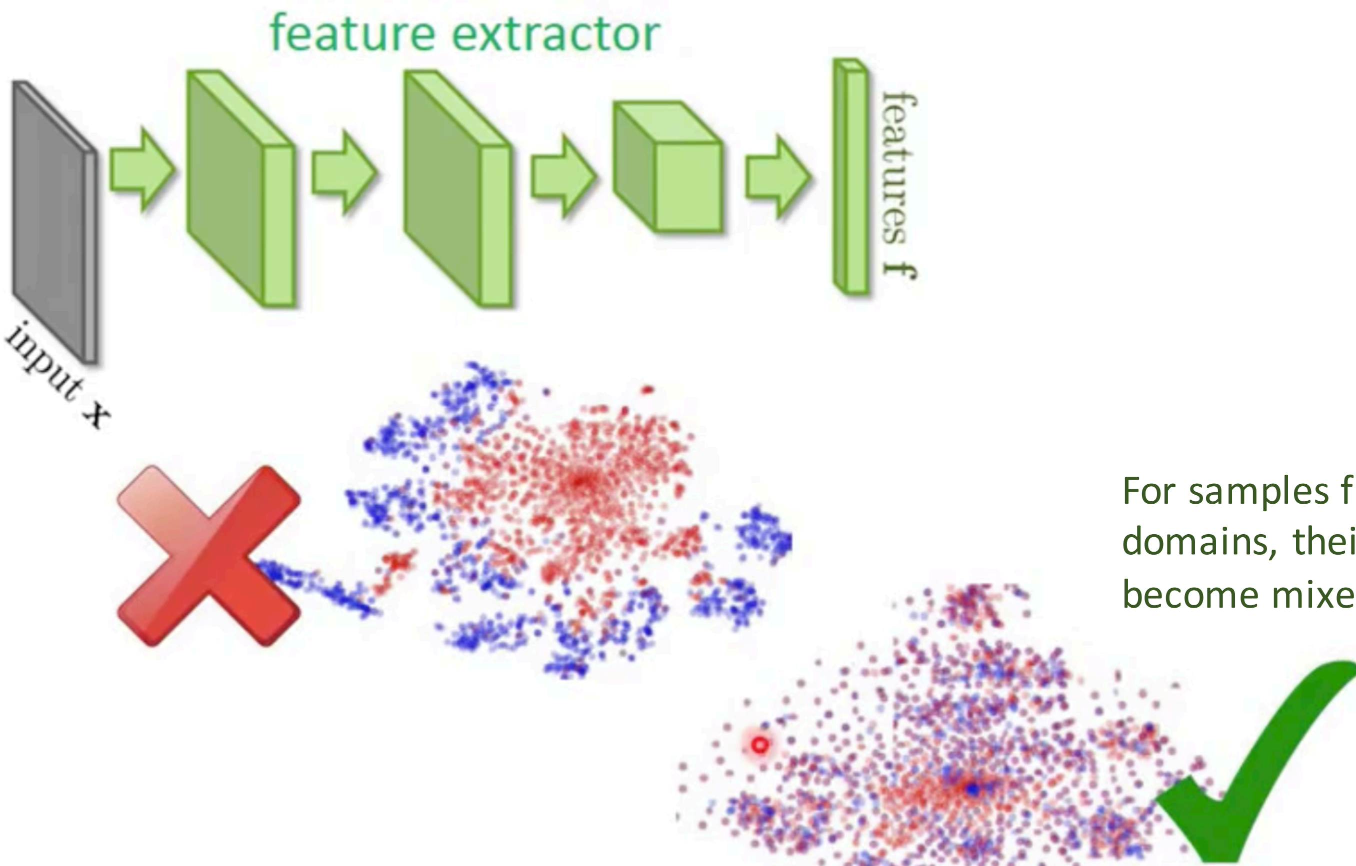
這邊有把 t-SNE 降維以後的結果

Domain-adversarial training



Idea: Remove domain-specific information from the extracted features

Domain-adversarial training

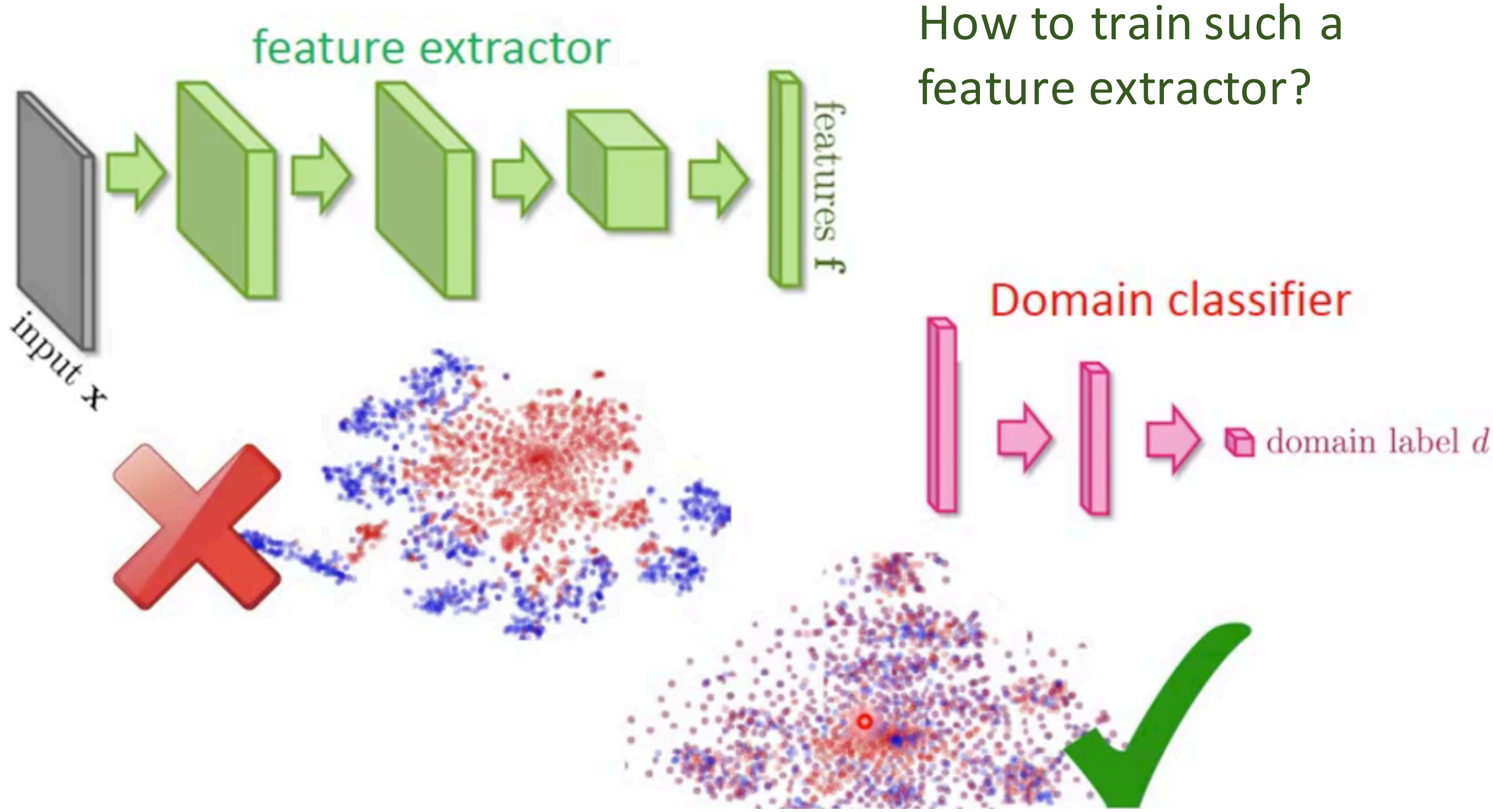


For samples from different domains, their features should become mixed together.

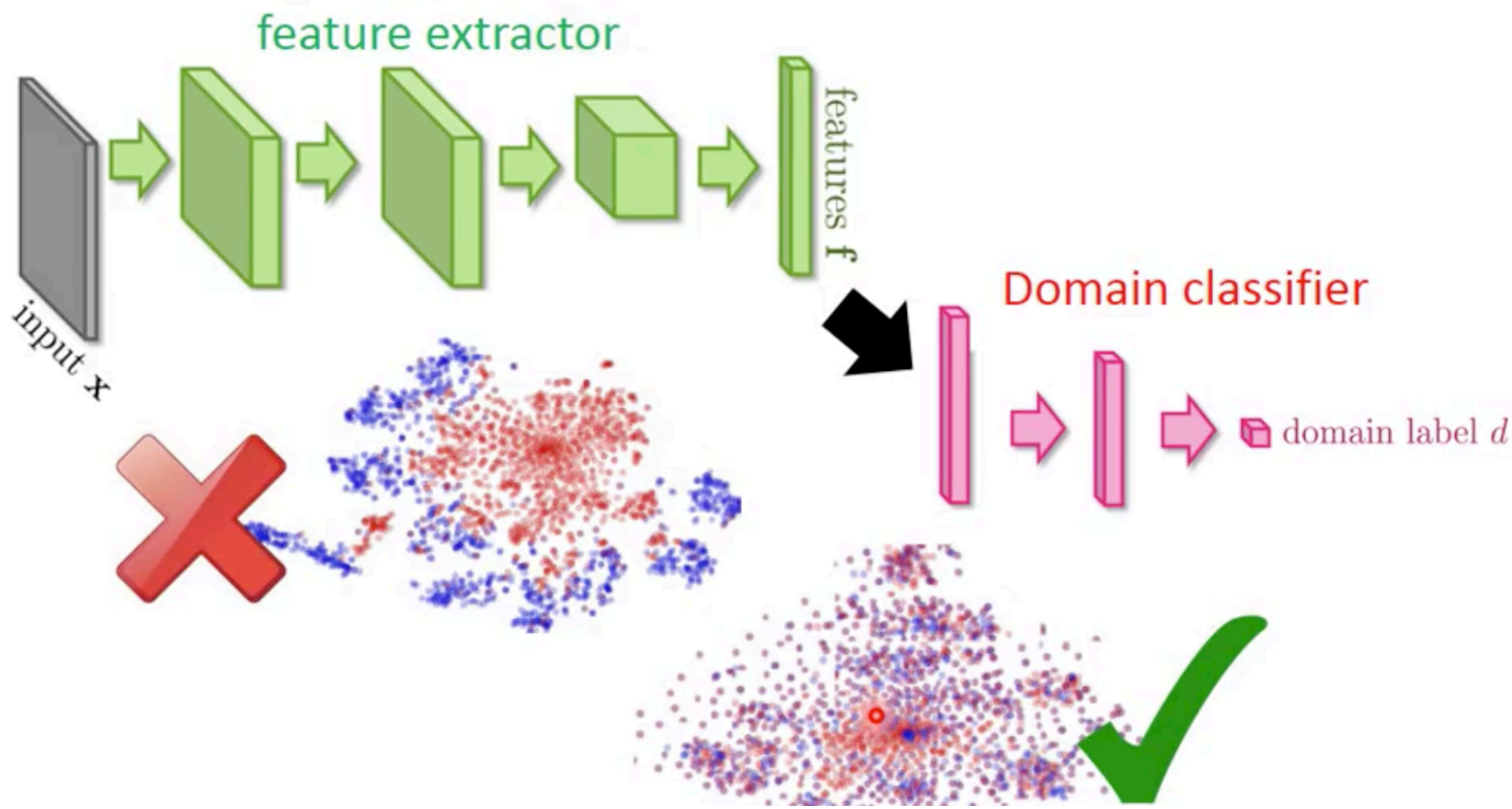
而是不同的 domain 都應該被混合在一起

Domain-adversarial training

How to train such a feature extractor?



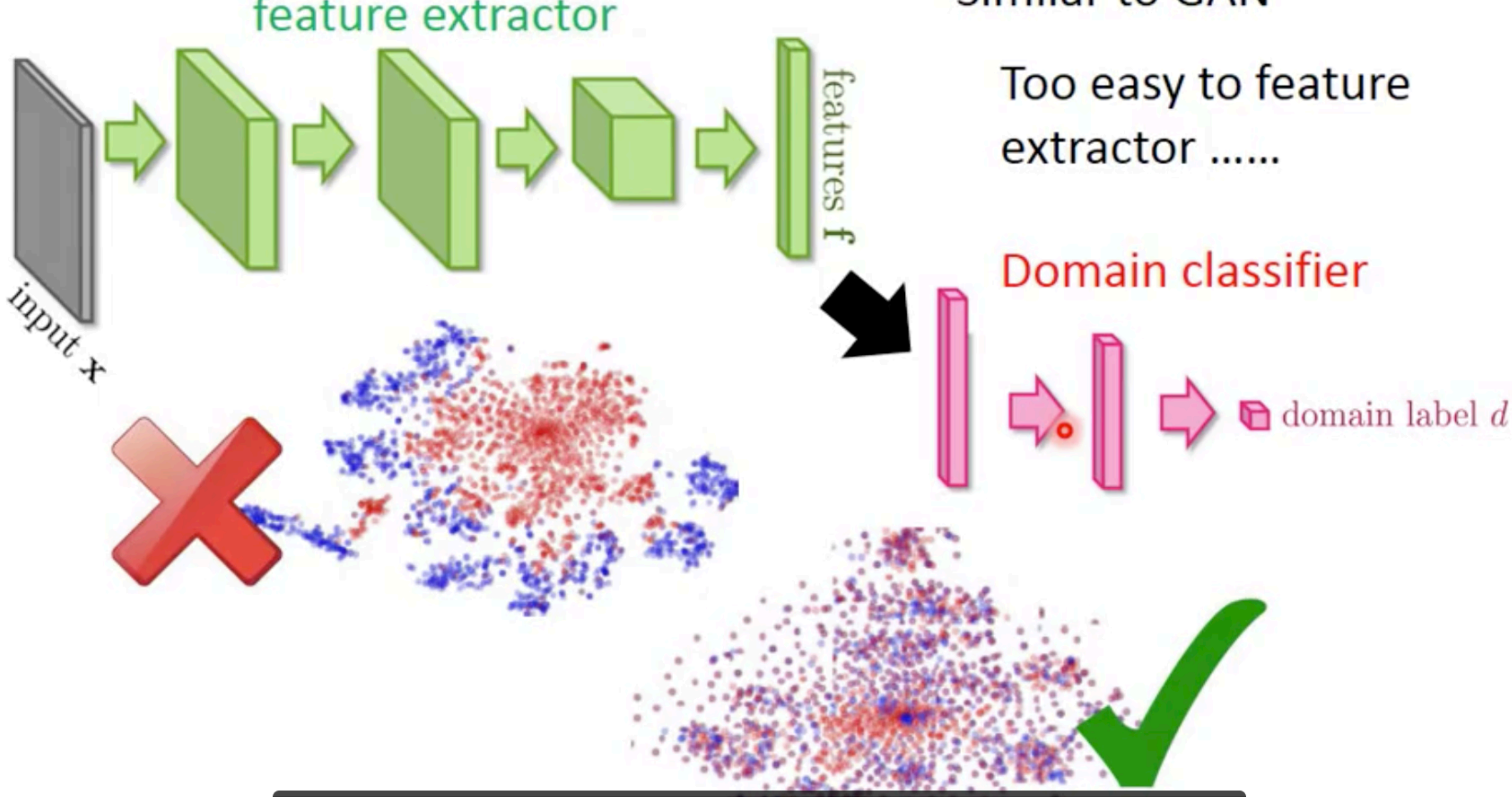
Domain-adversarial training



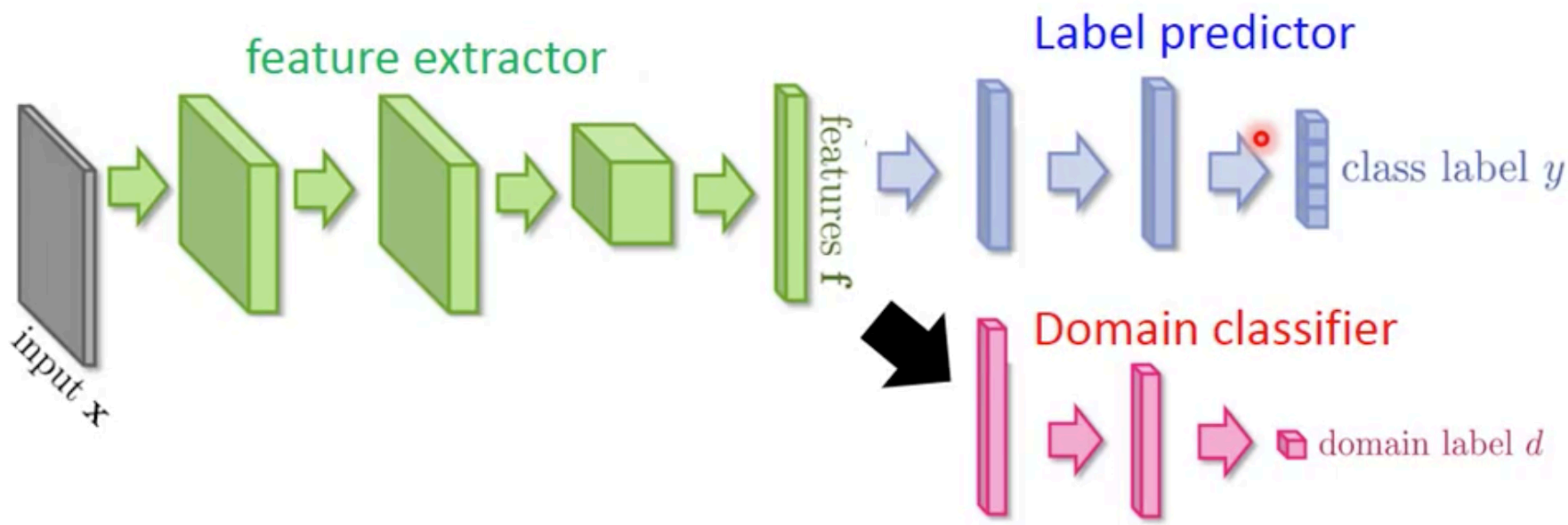
Domain-adversarial training

Similar to GAN

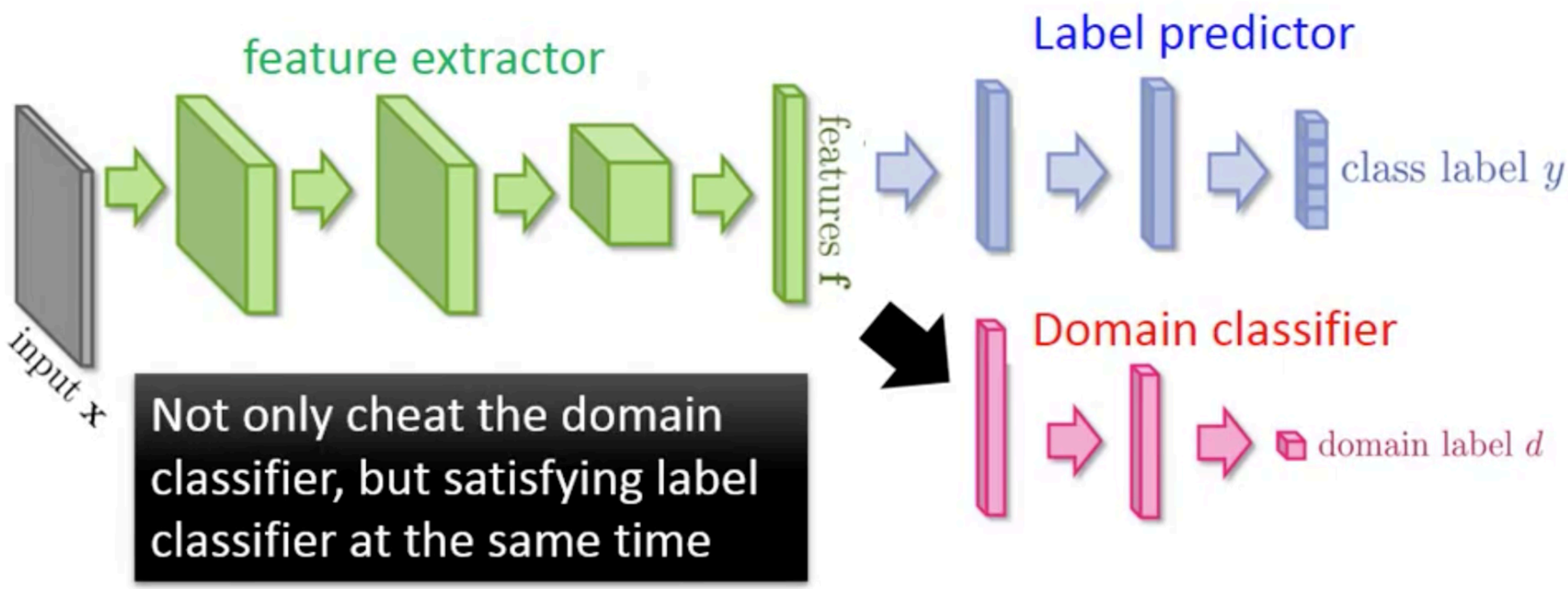
Too easy to feature extractor ......



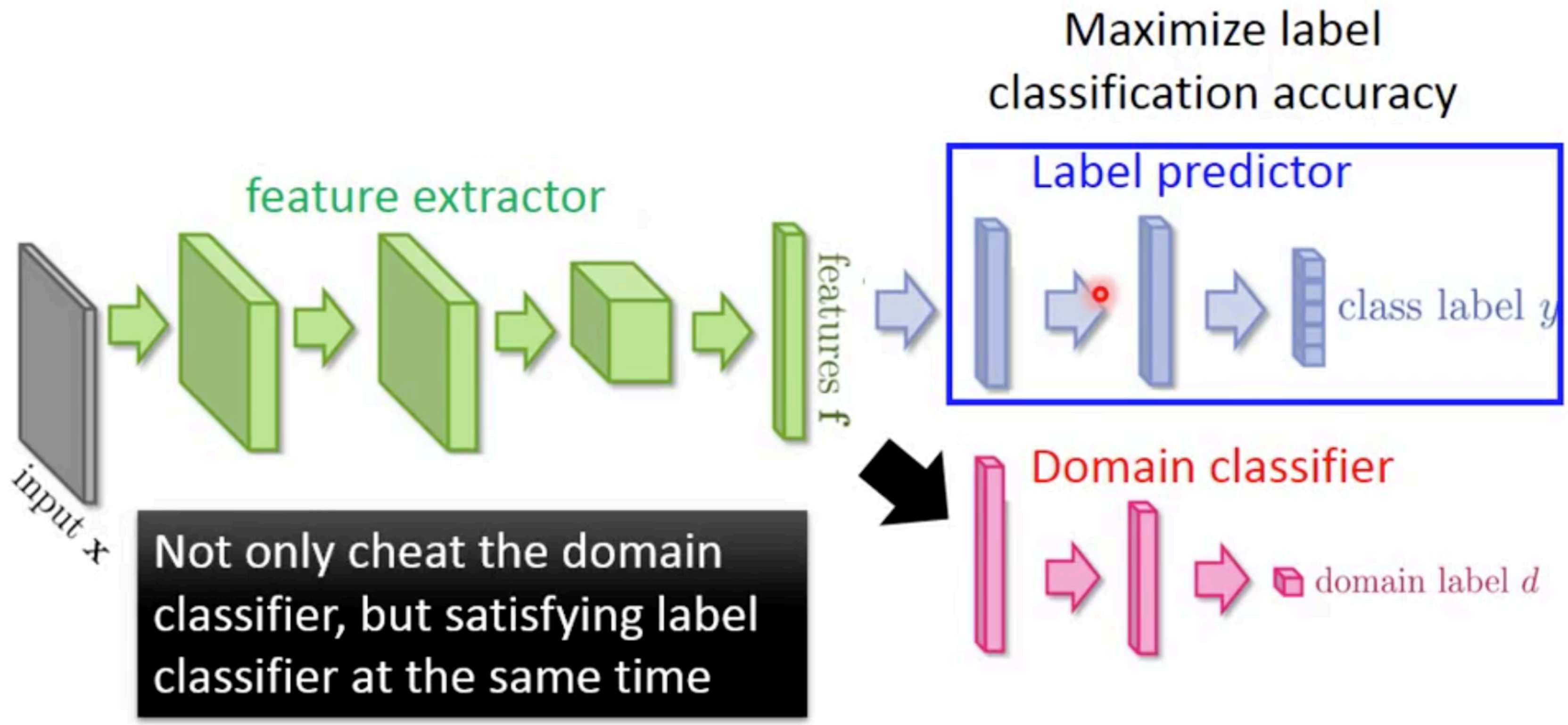
Domain-adversarial training



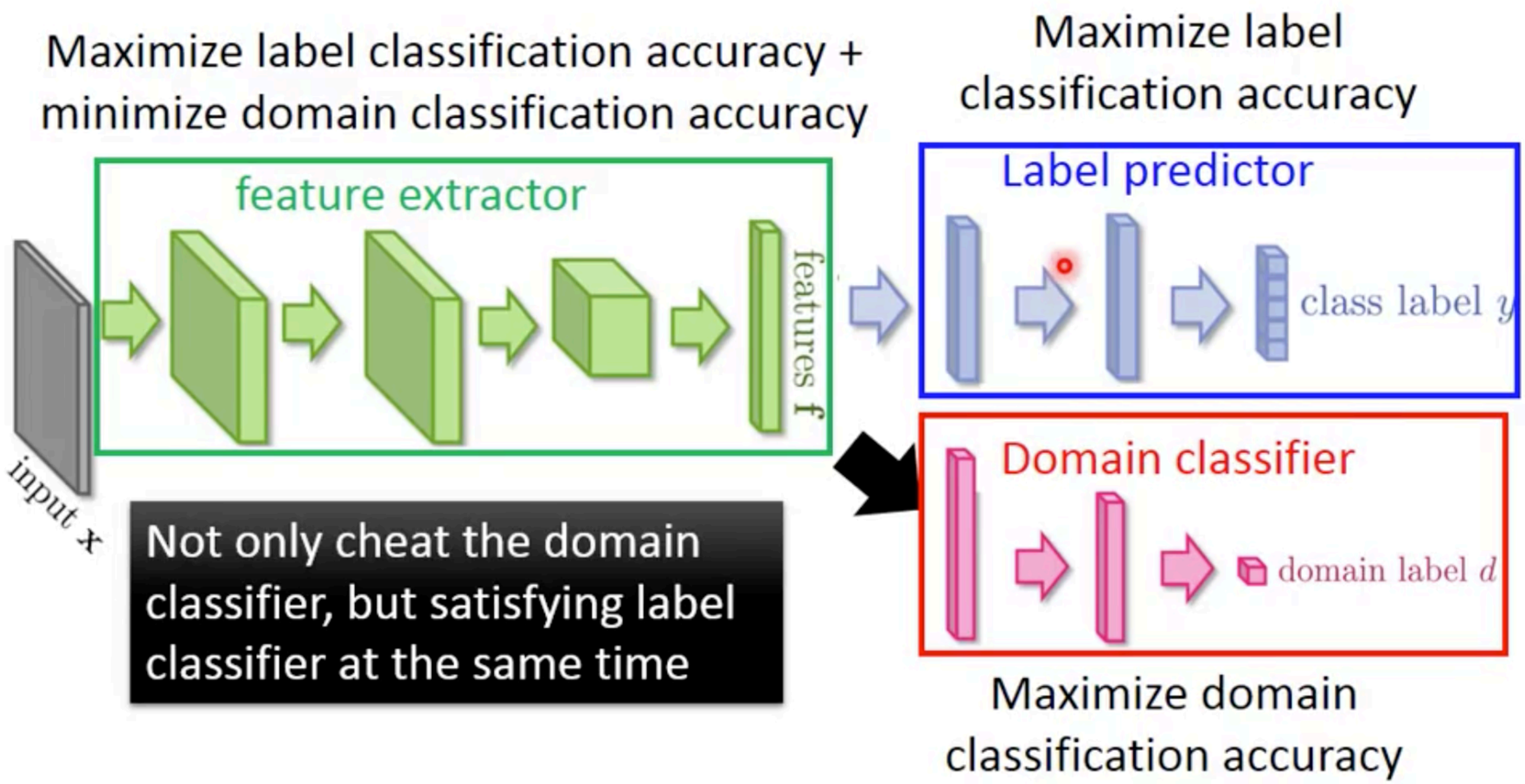
Domain-adversarial training



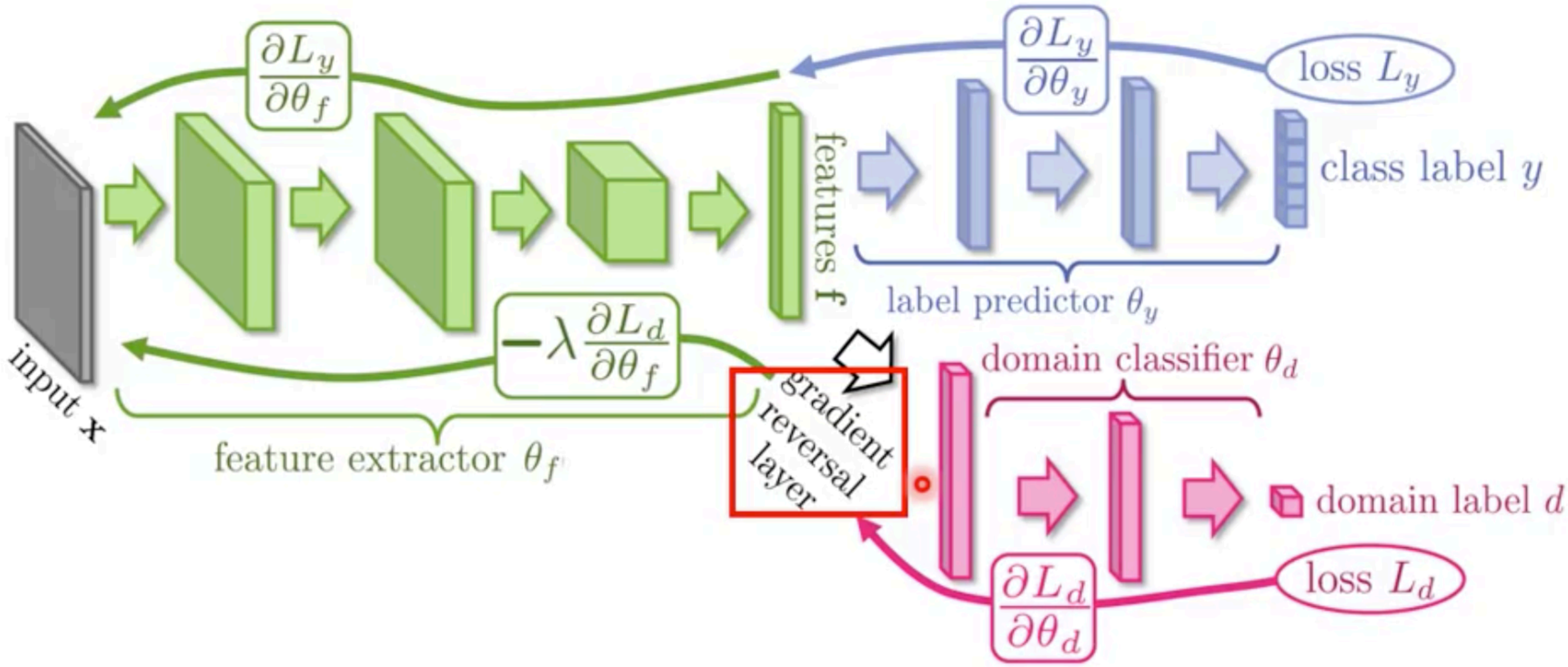
Domain-adversarial training



Domain-adversarial training



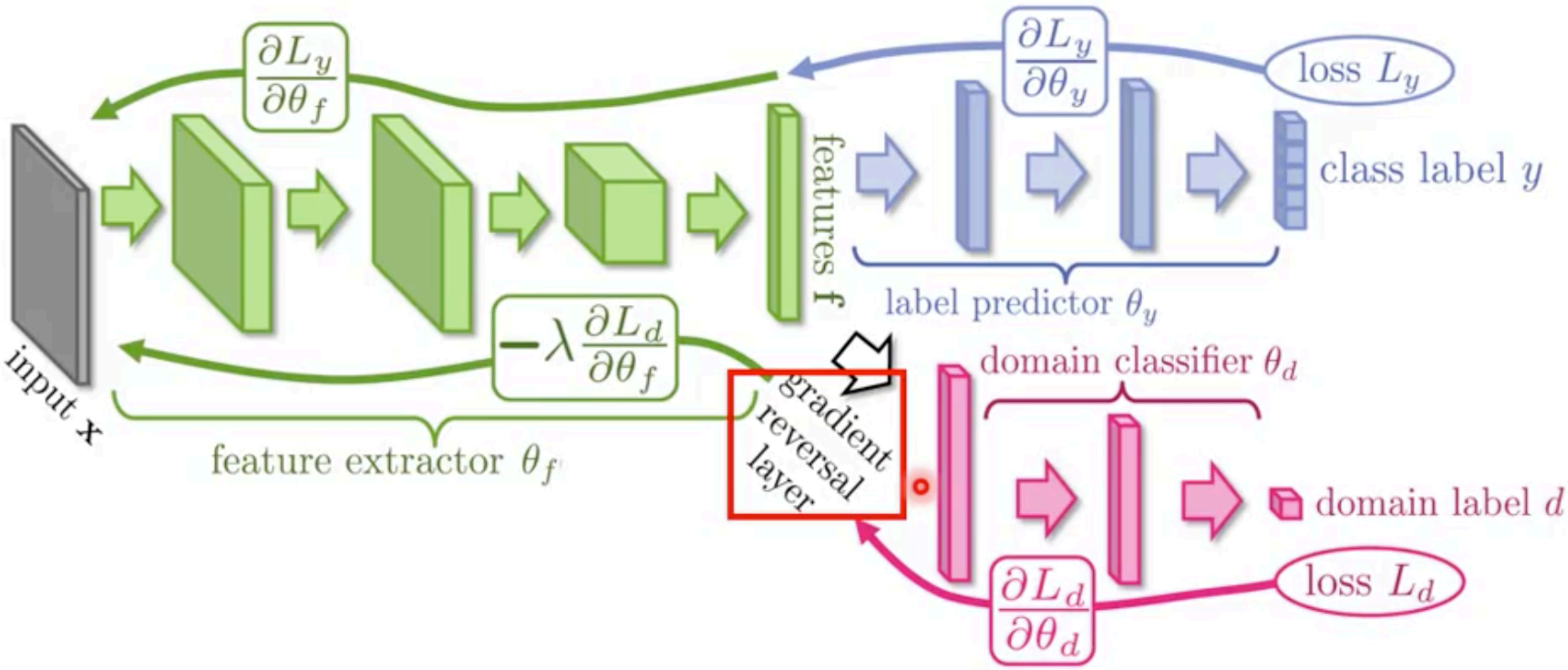
Domain-adversarial training



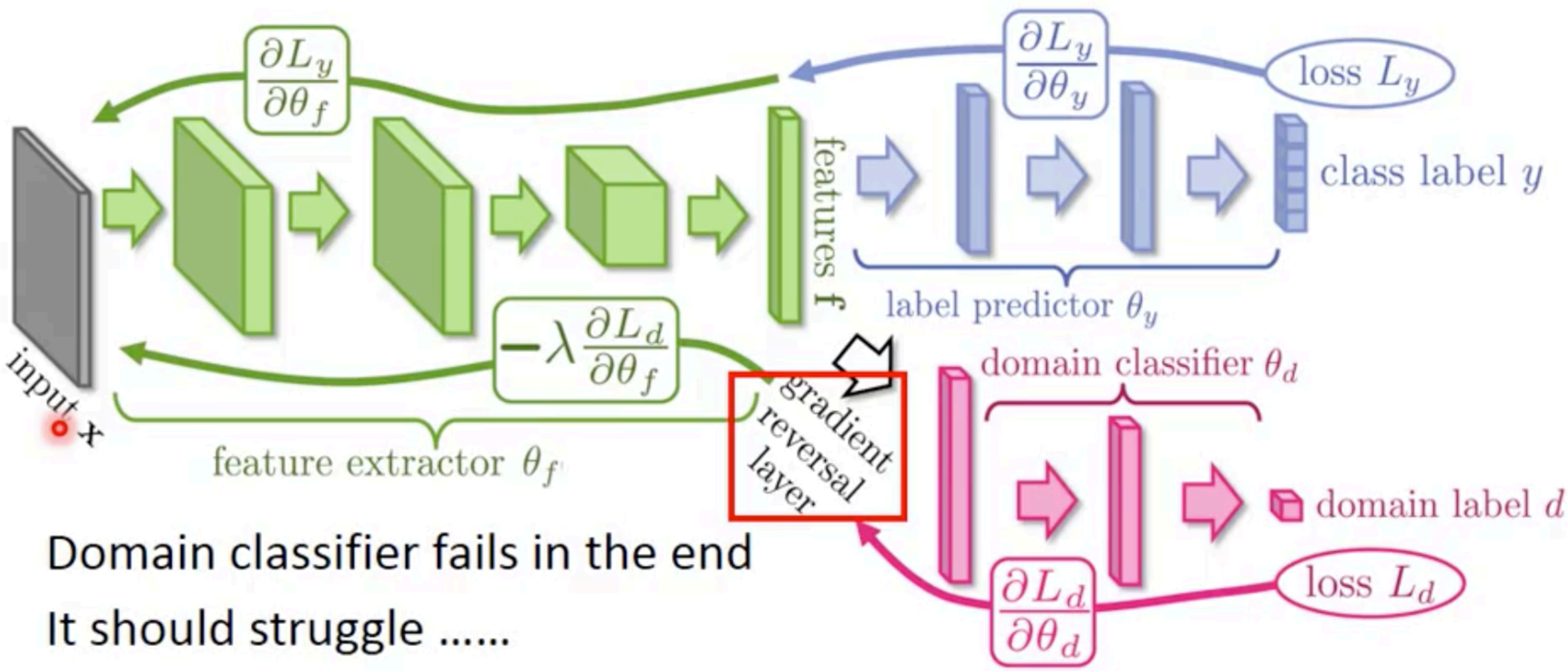
Yaroslav Ganin, Victor Lempitsky, Unsupervised Domain Adaptation by Backpropagation, ICML, 2015

Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, Domain-Adversarial Training of Neural Networks, JMLR, 2016

Domain-adversarial training



Domain-adversarial training



Domain classifier fails in the end

It should struggle ……

Domain-adversarial training

METHOD (SOURCE / TARGET)	MNIST / MNIST-M	SYN NUMBERS / SVHN	SVHN / MNIST	SYN SIGNS / GTSRB
SOURCE ONLY	.5749	.8665	.5919	.7400
SA (FERNANDO ET AL., 2013)	.6078 (7.9%)	.8672 (1.3%)	.6157 (5.9%)	.7635 (9.1%)
PROPOSED APPROACH	**.8149** (57.9%)	**.9048** (66.1%)	**.7107** (29.3%)	**.8866** (56.7%)
TRAIN ON TARGET	.9891	.9244	.9951	.9987

Yaroslav Ganin, Victor Lempitsky, Unsupervised Domain Adaptation by Backpropagation, ICML, 2015

Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, Domain-Adversarial Training of Neural Networks, JMLR, 2016

Transfer Learning - Overview

		Source Data (not directly related to the task)	
		labelled	unlabeled
Target Data	labelled	Fine-tuning Multitask Learning	
	unlabeled	Domain-adversarial training Zero-shot learning	

Zero-shot Learning

http://evchk.wikia.com/wiki/%E8%8D%89%E6%B3%A5%E9%A6%AC

- Source data: (x^s, y^s) → Training data
- Target data: (x^t) → Testing data

Zero-shot Learning

http://evchk.wikia.com/wiki/%E8%8D%89%E6%B3%A5%E9%A6%AC

- Source data: (x^s, y^s) → Training data
- Target data: (x^t) → Testing data

Different tasks

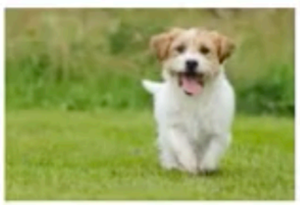
Zero-shot Learning

http://evchk.wikia.com/wiki/%E8%8D%89%E6%B3%A5%E9%A6%AC

- Source data: (x^s, y^s) → Training data
- Target data: (x^t) → Testing data

Different tasks

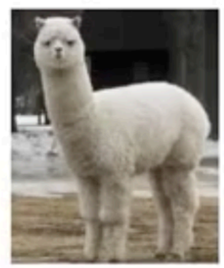
x^s:

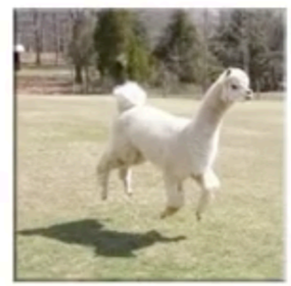


......

y^s: cat dog ......

x^t:





(true lable: llama)

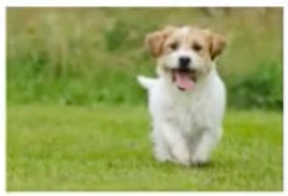
But it is not in the source data.
How can we recognize it?

Zero-shot Learning

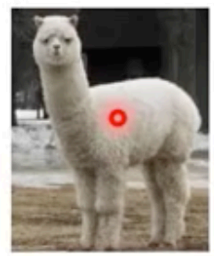
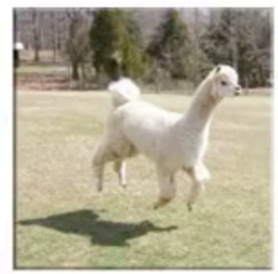
http://evchk.wikia.com/wiki/%E8%8D%89%E6%B3%A5%E9%A6%AC

- Source data: (x^s, y^s) → Training data
- Target data: (x^t) → Testing data

Different tasks

x^s:   ......

y^s: cat dog ......

x^t :  

In speech recognition, we can not have all possible words in the source (training) data.

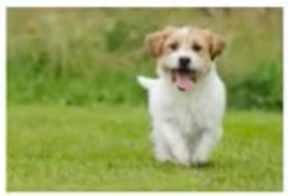
Ho to solve this problem in the speech recognition task?

Zero-shot Learning

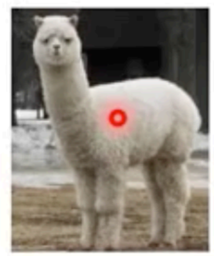
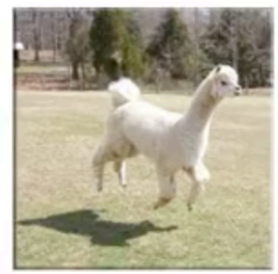
http://evchk.wikia.com/wiki/%E8%8D%89%E6%B3%A5%E9%A6%AC

- Source data: (x^s, y^s) → Training data
- Target data: (x^t) → Testing data

Different tasks

x^s:   ……

y^s: cat dog ……

x^t:  

In speech recognition, we can not have all possible words in the source (training) data.

Ho to solve this problem in the speech recognition task? Idea: recognize phoneme.

Zero-shot Learning

- Representing each class by its attributes

Zero-shot Learning

- Representing each class by its attributes

Database

class

	furry	4 legs	tail	…
Dog	O	O	O	
Fish	X	X	O	
Chimp	O	X	X	
…				

Zero-shot Learning

- Representing each class by its attributes

Database

class

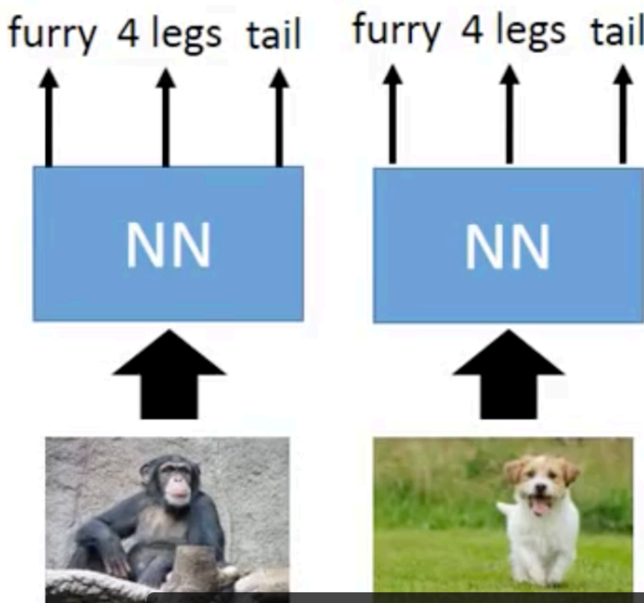
	furry	4 legs	tail	...
Dog	O	O	O	
Fish	X	X	O	
Chimp	O	X	X	
...				

Sufficient attributes for one-to-one mapping

Zero-shot Learning

- Representing each class by its attributes

Training



Database

attributes

class

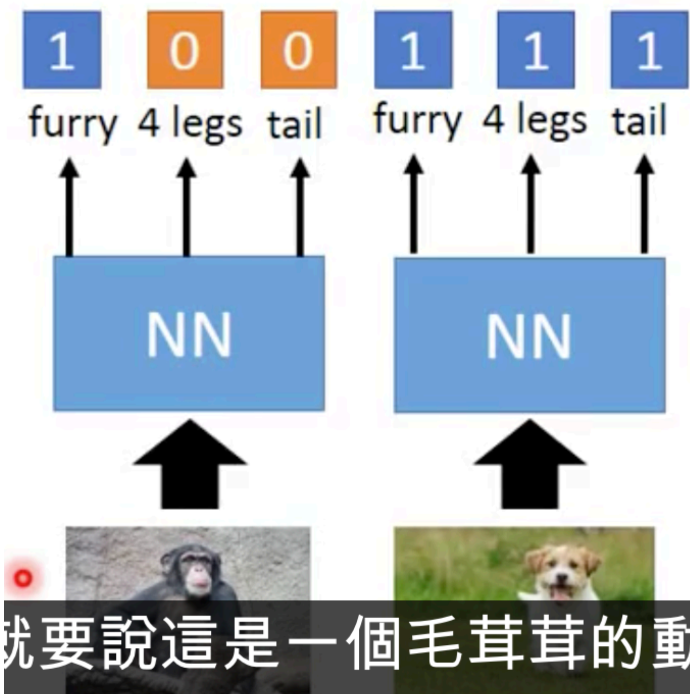
	furry	4 legs	tail	...
Dog	O	O	O	
Fish	X	X	O	
Chimp	O	X	X	
...				

sufficient attributes for one to one mapping

Zero-shot Learning

- Representing each class by its attributes

Training



Database

attributes

class

	furry	4 legs	tail	...
Dog	O	O	O	
Fish	X	X	O	
Chimp	O	X	X	
...				

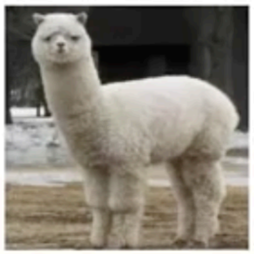
就要說這是一個毛茸茸的動物、沒有四隻腳的動物、沒有尾巴的
動物

Zero-shot Learning

- Representing each class by its attributes

Testing

It is OK even if
we have never
seen this animal.



attributes

class

	furry	4 legs	tail	...
Dog	O	O	O	
Fish	X	X	O	
Chimp	O	X	X	
...				

sufficient attributes for one to one mapping

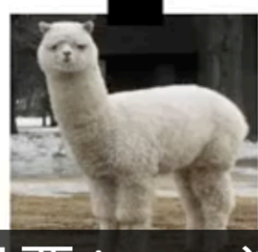
Zero-shot Learning

- Representing each class by its attributes

Testing

Attributes

NN



attributes

class

	furry	4 legs	tail	...
Dog	O	O	O	
Fish	X	X	O	
Chimp	O	X	X	
...				

sufficient attributes for one to one mapping

只要 input 這張 image，他有甚麼樣的 attribute

Zero-shot Learning

- Representing each class by its attributes

Testing

Find the class with the most similar attributes

Attributes

NN

class

attributes

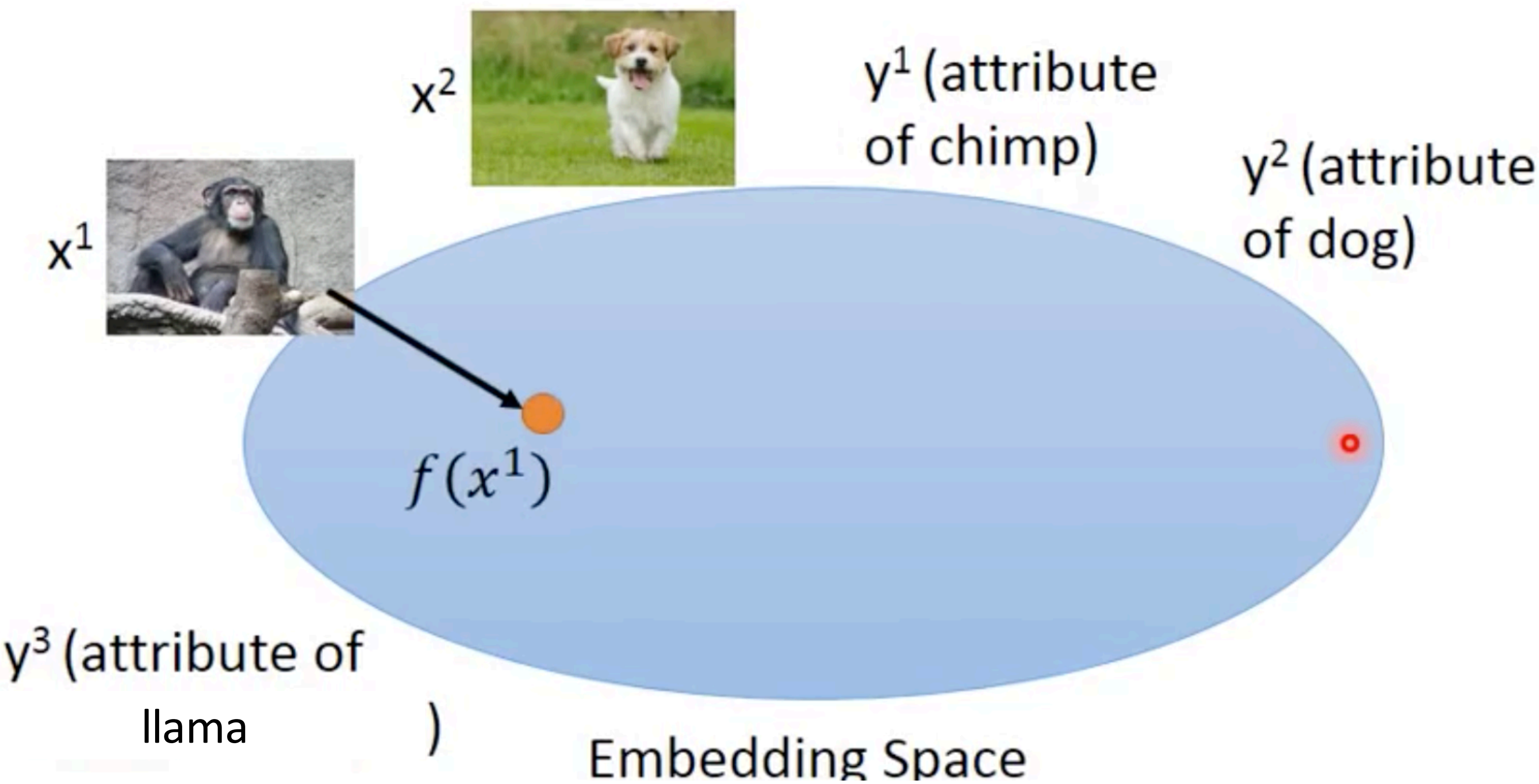
	furry	4 legs	tail	...
Dog	O	O	O	
Fish	X	X	O	
Chimp	O	X	X	
...				

sufficient attributes for one to one mapping

有時候可能沒有一模一樣，也沒有關係，就看誰最接近

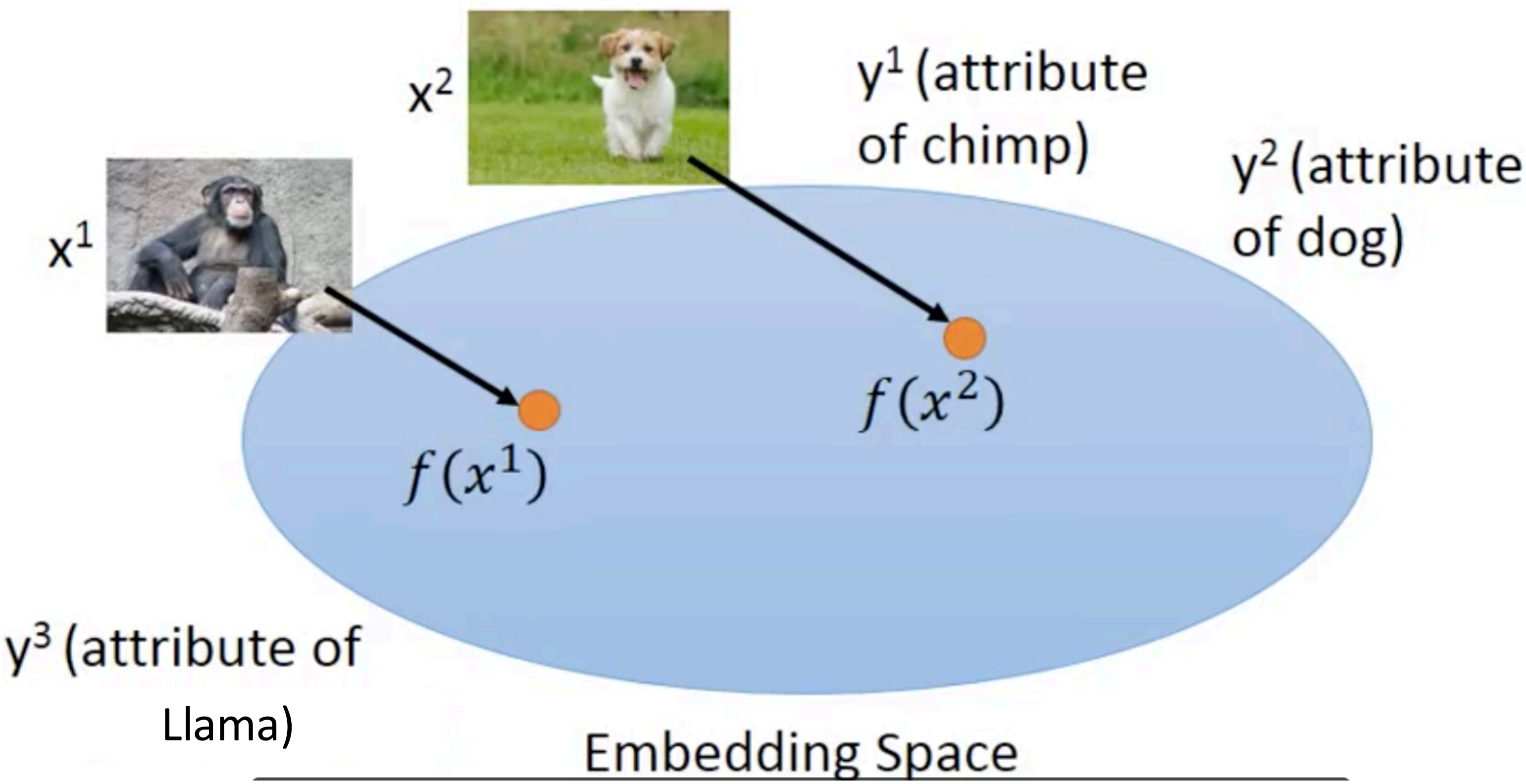
Zero-shot Learning

- Attribute embedding



Zero-shot Learning

- Attribute embedding



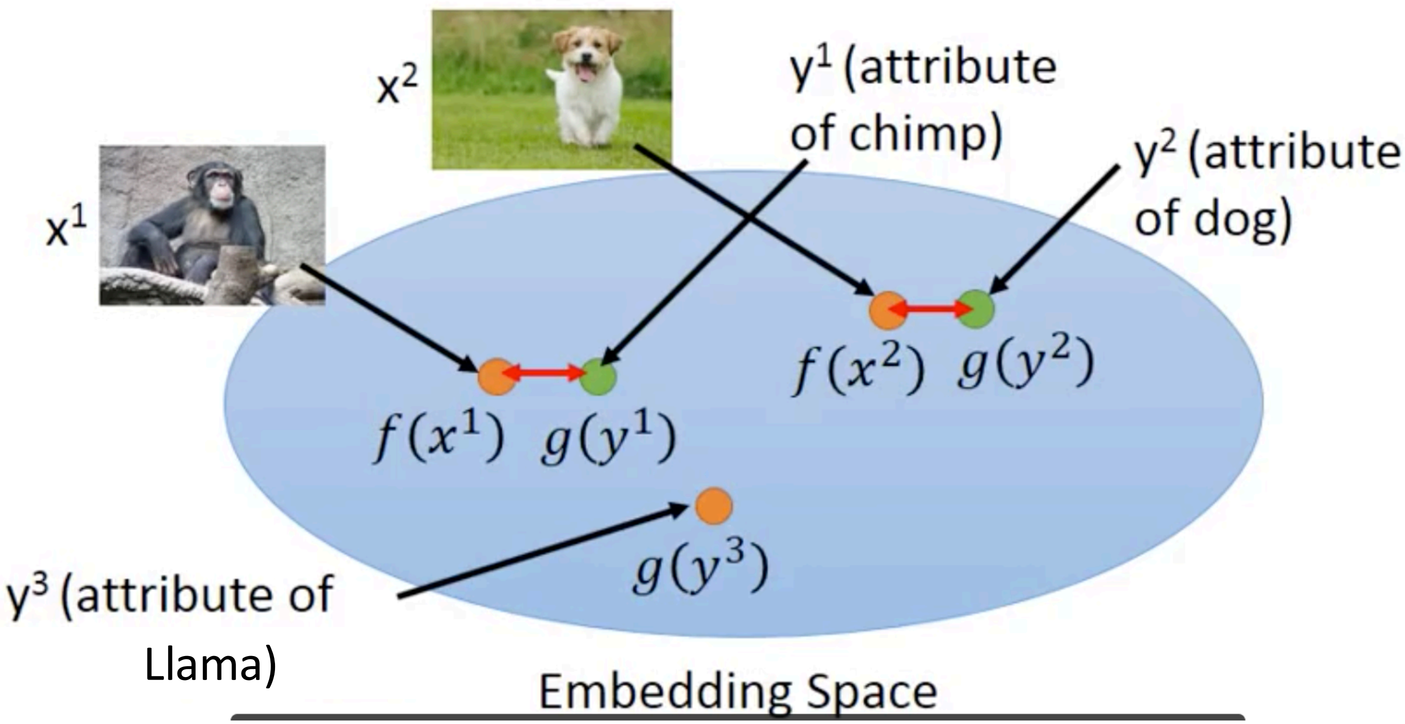
Zero-shot Learning

$f(*)$ and $g(*)$ can be NN.

Training target:

$f(x^n)$ and $g(y^n)$ as close as possible

- Attribute embedding



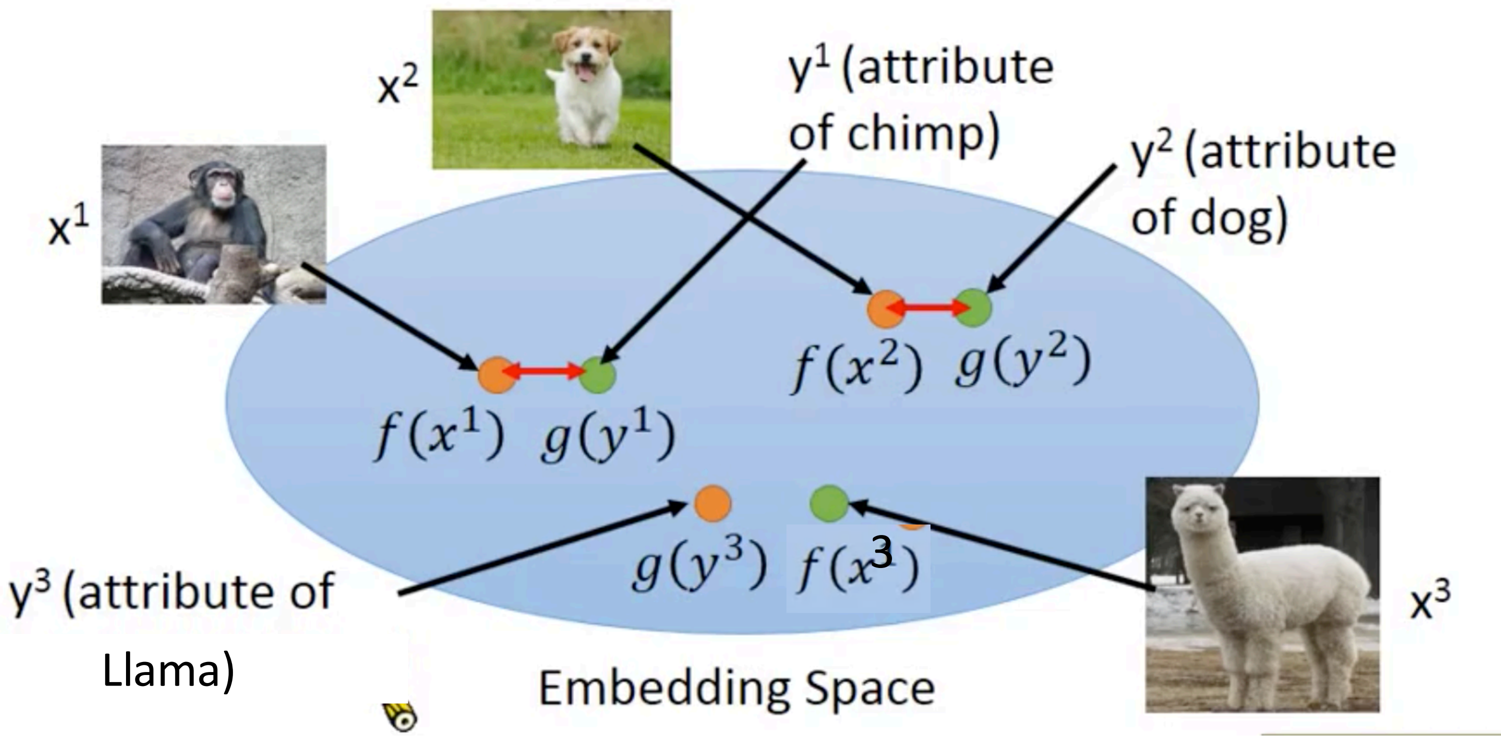
Zero-shot Learning

$f(*)$ and $g(*)$ can be NN.

Training target:

$f(x^n)$ and $g(y^n)$ as close as possible

- Attribute embedding

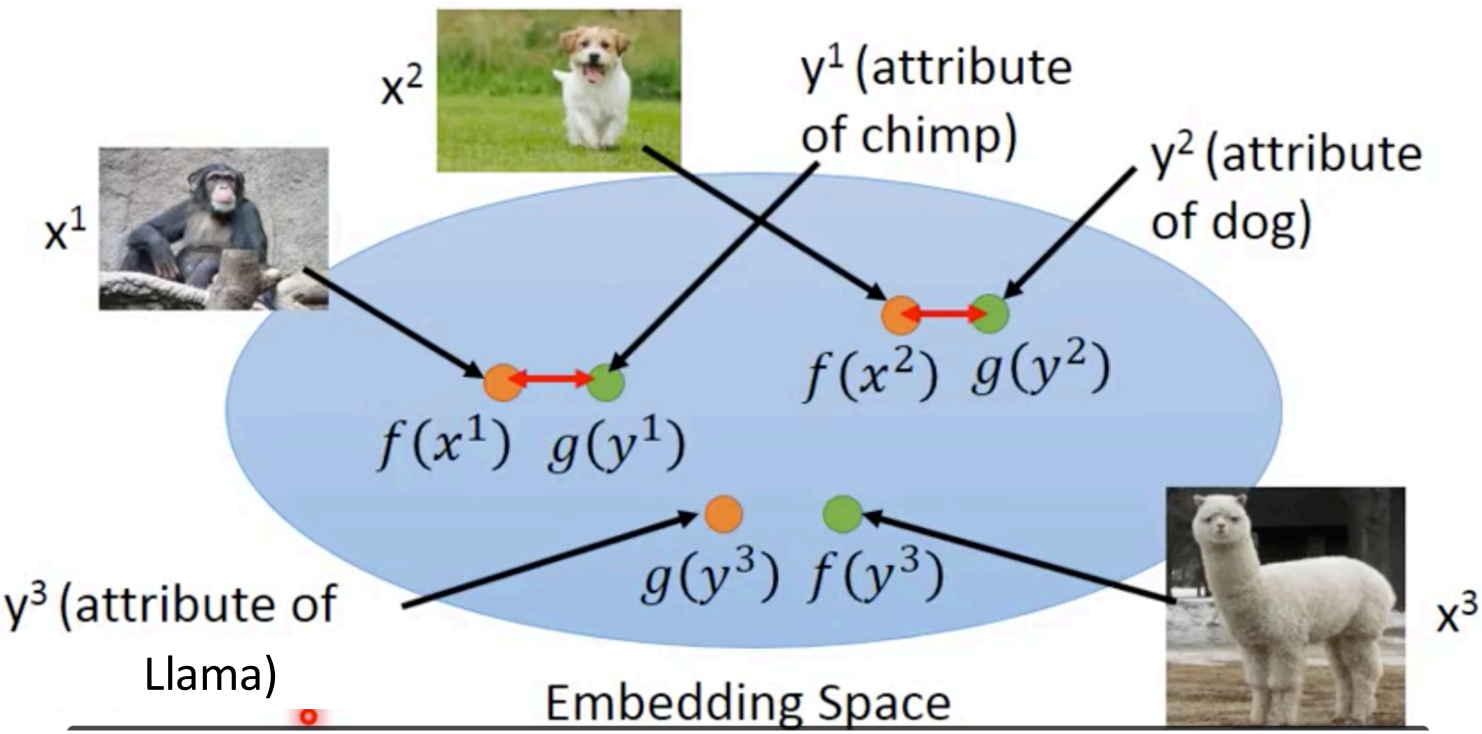


Zero-shot Learning

What if we don't have database

Use word2vec.

- Attribute embedding + word embedding



Zero-shot Learning

$$f^*, g^* = arg \min_{f,g} \sum_n \|f(x^n) - g(y^n)\|_2$$

Zero-shot Learning

$$f^*, g^* = arg \min_{f,g} \sum_n \|f(x^n) - g(y^n)\|_2$$

Problem?

The network can simply map all inputs to the same point in the feature space.

Zero-shot Learning

$$f^*, g^* = arg \min_{f,g} \sum_n \|f(x^n) - g(y^n)\|_2$$

Problem?

$$f^*, g^* = arg \min_{f,g} \sum_n max\left(0, k - f(x^n) \cdot g(y^n) + \max_{m \neq n} f(x^n) \cdot g(y^m)\right)$$

Zero-shot Learning

$$f^*, g^* = arg \min_{f,g} \sum_n \|f(x^n) - g(y^n)\|_2$$ Problem?

$$f^*, g^* = arg \min_{f,g} \sum_n max\Big(0, k - f(x^n) \cdot g(y^n) + \max_{m \neq n} f(x^n) \cdot g(y^m)\Big)$$

Margin you defined

Zero loss:

Zero-shot Learning

$$f^*, g^* = arg \min_{f,g} \sum_n \|f(x^n) - g(y^n)\|_2$$ Problem?

$$f^*, g^* = arg \min_{f,g} \sum_n max\Big(0, \underset{\substack{\uparrow \\ \text{Margin you defined}}}{k} - f(x^n) \cdot g(y^n) + \max_{m \neq n} f(x^n) \cdot g(y^m)\Big)$$

Zero loss: $k - f(x^n) \cdot g(y^n) + \max_{m \neq n} f(x^n) \cdot g(y^m) < 0$

Zero-shot Learning

$$f^*, g^* = arg \min_{f,g} \sum_n \|f(x^n) - g(y^n)\|_2$$ Problem?

$$f^*, g^* = arg \min_{f,g} \sum_n max\left(0, k - f(x^n) \cdot g(y^n) + \max_{m \neq n} f(x^n) \cdot g(y^m)\right)$$

Margin you defined

Zero loss: $k - f(x^n) \cdot g(y^n) + \max_{m \neq n} f(x^n) \cdot g(y^m) < 0$

$$f(x^n) \cdot g(y^n) - \max_{m \neq n} f(x^n) \cdot g(y^m) > k$$

Zero-shot Learning

$$f^*, g^* = arg \min_{f,g} \sum_n \|f(x^n) - g(y^n)\|_2$$ Problem?

$$f^*, g^* = arg \min_{f,g} \sum_n max\Big(0, k - f(x^n) \cdot g(y^n) + \max_{m \neq n} f(x^n) \cdot g(y^m)\Big)$$

Margin you defined

Zero loss: $k - f(x^n) \cdot g(y^n) + \max_{m \neq n} f(x^n) \cdot g(y^m) < 0$

$$\underline{f(x^n) \cdot g(y^n)} - \underline{\max_{m \neq n} f(x^n) \cdot g(y^m)} > k$$

$f(x^n)$ and $g(y^n)$ as close

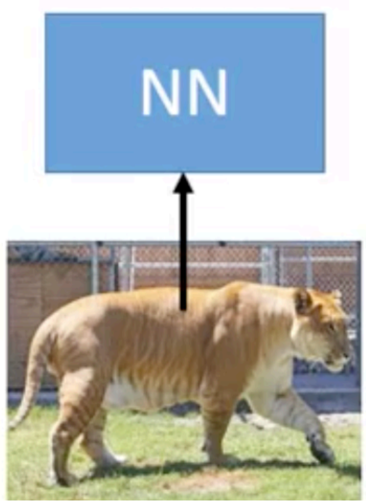
$f(x^n)$ and $g(y^m)$ not as close

Zero-shot Learning

- Convex Combination of Semantic Embedding

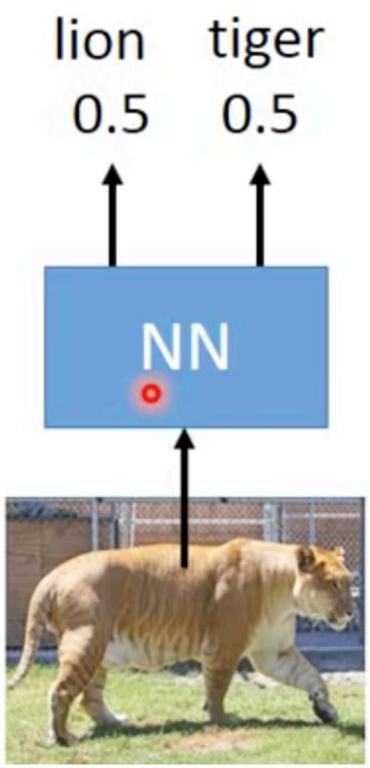
Zero-shot Learning

- Convex Combination of Semantic Embedding



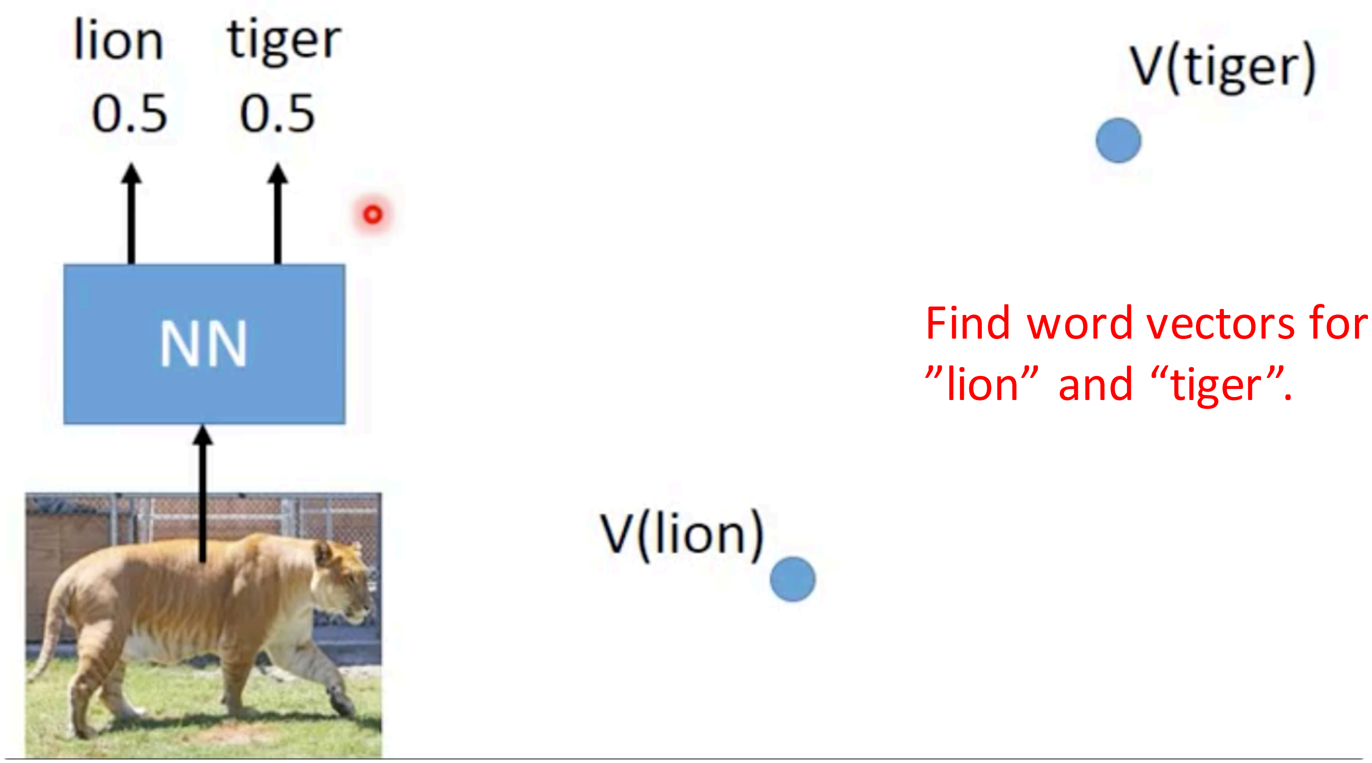
Zero-shot Learning

- Convex Combination of Semantic Embedding



Zero-shot Learning

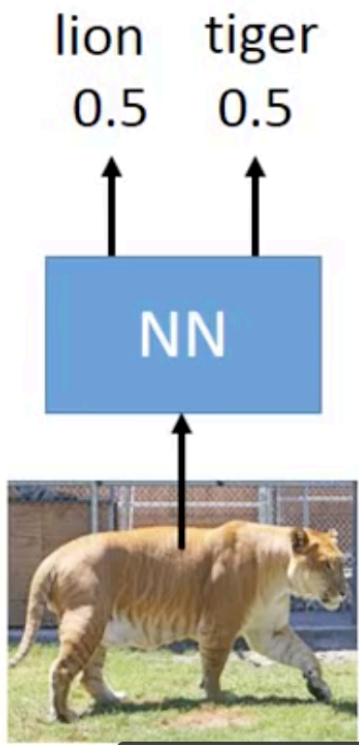
- Convex Combination of Semantic Embedding

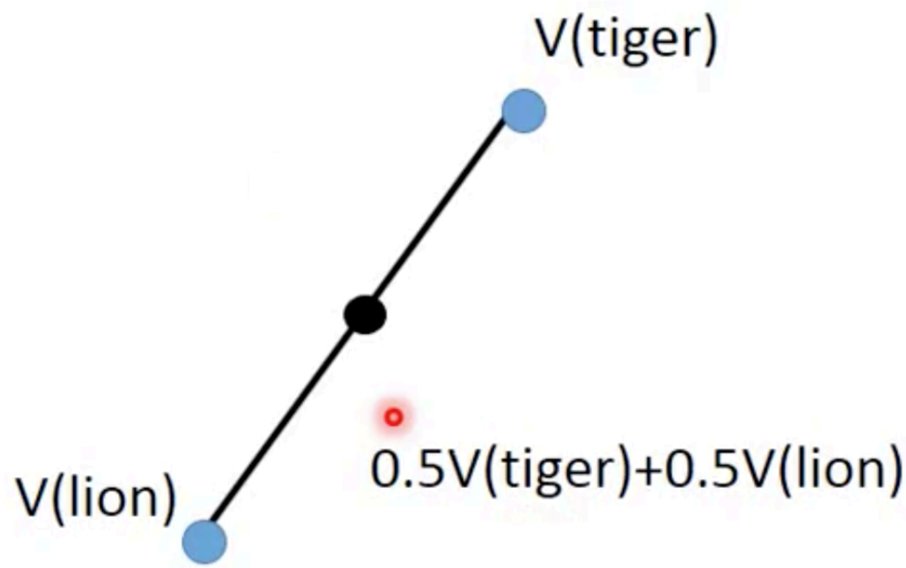


Find word vectors for
”lion” and “tiger”.

Zero-shot Learning

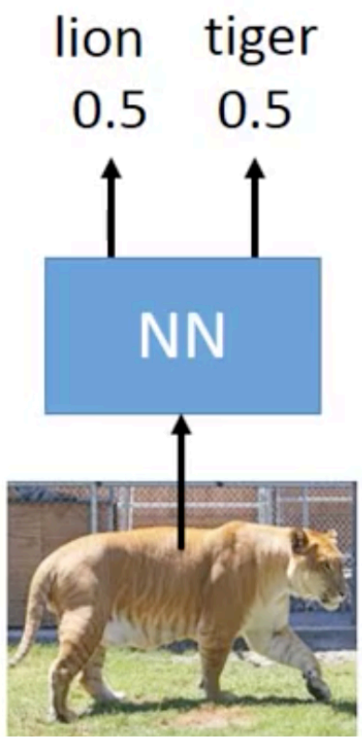
- Convex Combination of Semantic Embedding

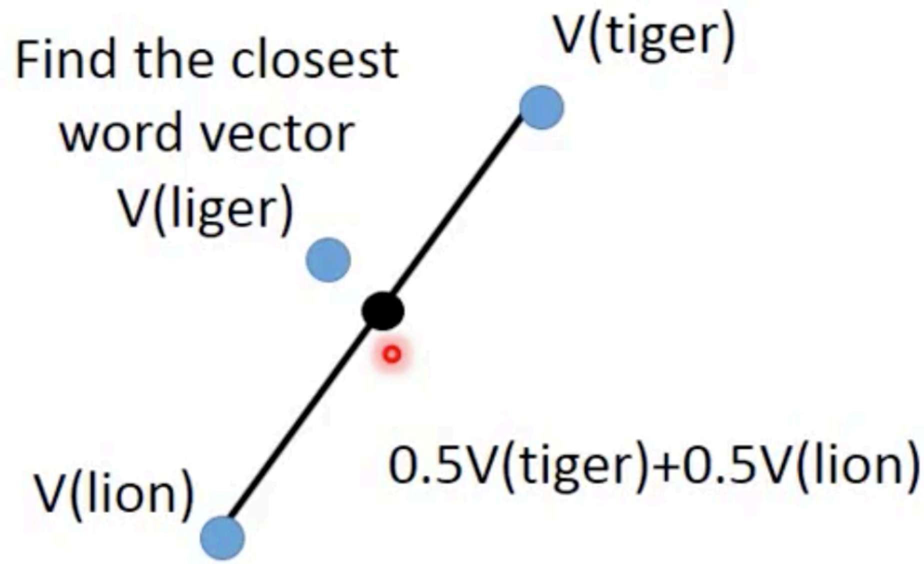




Zero-shot Learning

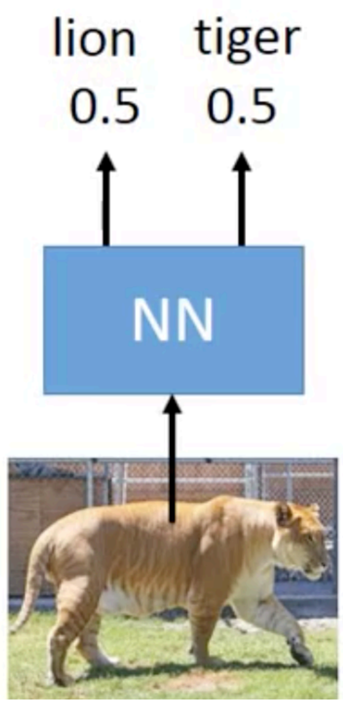
- Convex Combination of Semantic Embedding

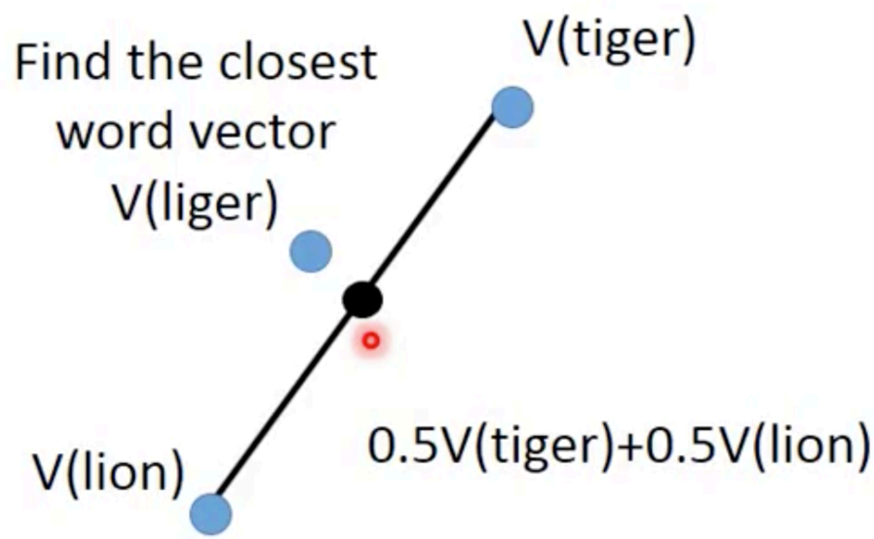




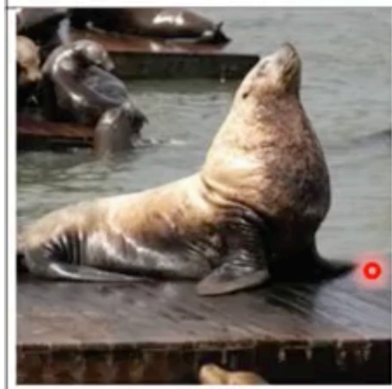
Zero-shot Learning

- Convex Combination of Semantic Embedding

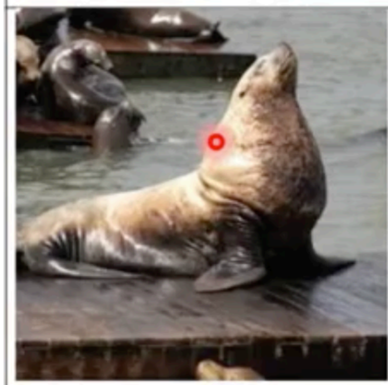


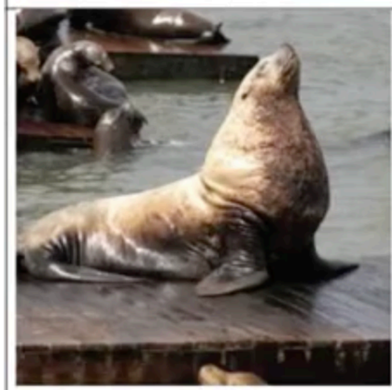


Only need off-the-shelf NN
for ImageNet and Word2Vec.

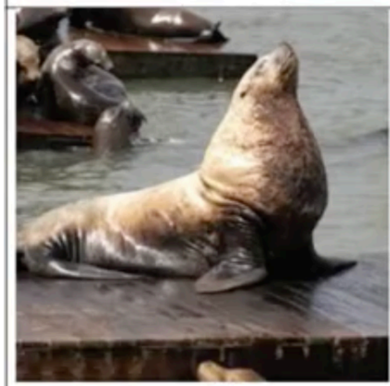
Test Image	ConvNet	DeViSE	ConSE(10)
	sea lion carpenter's plane cowboy boot loggerhead goose		

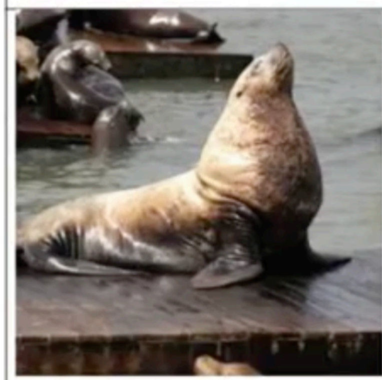
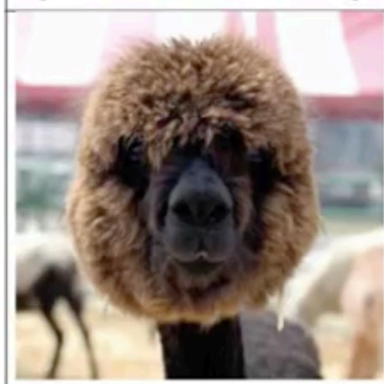
直接拿一個 CNN

Test Image	ConvNet	DeViSE	ConSE(10)
 (Stellar sea lion)	sea lion carpenter's plane cowboy boot loggerhead goose		

Test Image	ConvNet	DeViSE	ConSE(10)
 **(Stellar sea lion)**	sea lion carpenter's plane cowboy boot loggerhead goose	elephant turtle turtleneck flip-flop cart, handcart	

DSeViSE:
Project image and features to nearby points in the same space.

Test Image	ConvNet	DeViSE	ConSE(10) Convex combination for semantic embedding
 (Stellar sea lion)	sea lion carpenter's plane cowboy boot loggerhead goose	elephant turtle turtleneck flip-flop cart, handcart	California sea lion **Steller sea lion** Australian sea lion South American sea lion eared seal

Test Image	ConvNet	DeViSE	ConSE(10)
 (Stellar sea lion)	sea lion carpenter's plane cowboy boot loggerhead goose	elephant turtle turtleneck flip-flop cart, handcart	California sea lion **Steller sea lion** Australian sea lion South American sea lion eared seal
 (Lama pacos)	Tibetan mastiff titi monkey Koala llama chow-chow	kernel littoral zone carillon Cabernet Sauvignon poodle dog	domestic dog domestic cat schnauzer Belgian sheepdog domestic llama

這個 network 也沒有得到正確的結果

More about Zero-shot learning

- Mark Palatucci, Dean Pomerleau, Geoffrey E. Hinton, Tom M. Mitchell, “Zero-shot Learning with Semantic Output Codes”, NIPS 2009
- Zeynep Akata, Florent Perronnin, Zaid Harchaoui and Cordelia Schmid, “Label-Embedding for Attribute-Based Classification”, CVPR 2013
- Andrea Frome, Greg S. Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, Tomas Mikolov, “DeViSE: A Deep Visual-Semantic Embedding Model”, NIPS 2013
- Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S. Corrado, Jeffrey Dean, “Zero-Shot Learning by Convex Combination of Semantic Embeddings”, arXiv preprint 2013
- Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, Raymond Mooney, Trevor Darrell, Kate Saenko, “Captioning Images with Diverse Objects”, arXiv preprint 2016

Transfer Learning - Overview

		Source Data (not directly related to the task)	
		labelled	unlabeled
Target Data	labelled	Fine-tuning Multitask Learning	Self-taught learning Rajat Raina , Alexis Battle , Honglak Lee , Benjamin Packer , Andrew Y. Ng, Self-taught learning: transfer learning from unlabeled data, ICML, 2007
	unlabeled	Domain-adversarial training Zero-shot learning	Self-taught Clustering Wenyuan Dai, Qiang Yang, Gui-Rong Xue, Yong Yu, "Self-taught clustering", ICML 2008

所以你遇到一個狀況是

Self-taught learning

- Learning to extract better representation from the source data (unsupervised approach)
- Extracting better representation for target data

Self-taught learning

- Learning to extract better representation from the source data (unsupervised approach)
- Extracting better representation for target data

Domain	Unlabeled data	Labeled data	Classes	Raw features
Image classification	10 images of outdoor scenes	Caltech101 image classification dataset	101	Intensities in 14x14 pixel patch
Handwritten character recognition	Handwritten digits ("0"–"9")	Handwritten English characters ("a"–"z")	26	Intensities in 28x28 pixel character/digit image
Font character recognition	Handwritten English characters ("a"–"z")	Font characters ("a"/"A" – "z"/"Z")	26	Intensities in 28x28 pixel character image
Song genre classification	Song snippets from 10 genres	Song snippets from 7 *different* genres	7	Log-frequency spectrogram over 50ms time windows
Webpage classification	100,000 news articles (Reuters newswire)	Categorized webpages (from DMOZ hierarchy)	2	Bag-of-words with 500 word vocabulary
UseNet article classification	100,000 news articles (Reuters newswire)	Categorized UseNet posts (from "SRAA" dataset)	2	Bag-of-words with 377 word vocabulary