

# Transfer Learning

<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



dog



cat

Similar domain, different tasks

Different domains, same task

# Why?

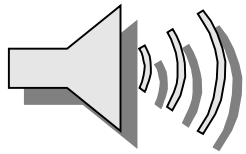
<http://www.bigr.nl/website/structure/main.php?page=researchlines&subpage=project&id=64>

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

Task Considered

Data not directly related

Speech  
Recognition

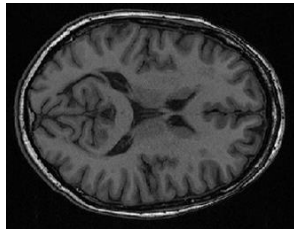


Taiwanese



English  
Chinese  
.....

Image  
Recognition



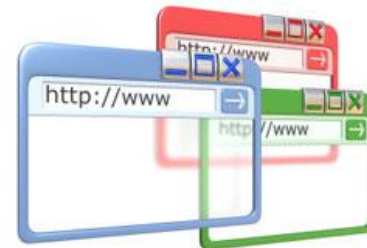
Medical  
Images



Text  
Analysis



Specific  
domain



Webpages

# Transfer Learning

- Example in real life

研究生

漫画家

研究生



漫画家

跑实验



画分镜

指导教授



责编

投稿期刊



投稿 jump

(word embedding knows that)



爆漫王



# Transfer Learning - Overview

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

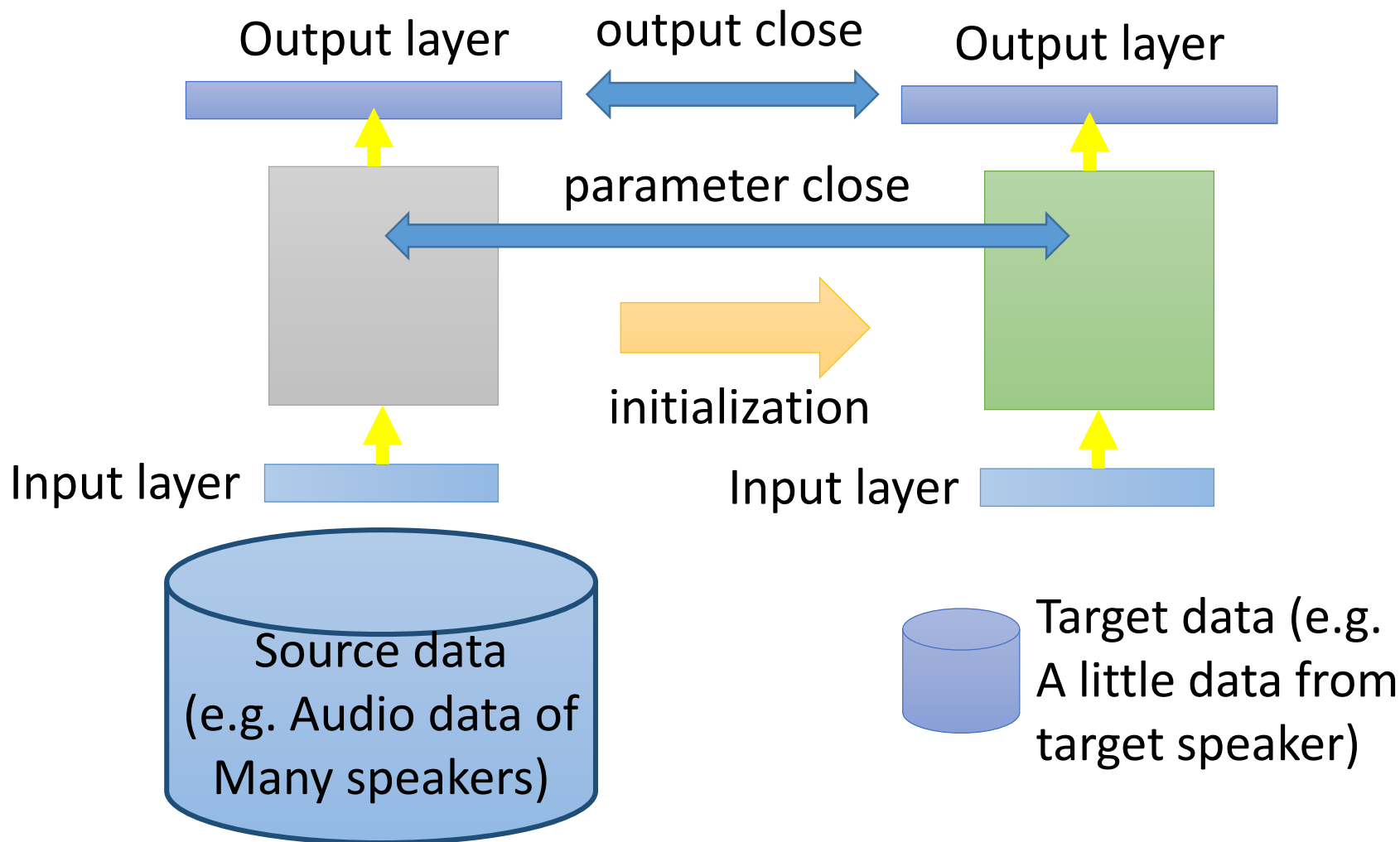
Warning: different terminology in different literature

# Model Fine-tuning

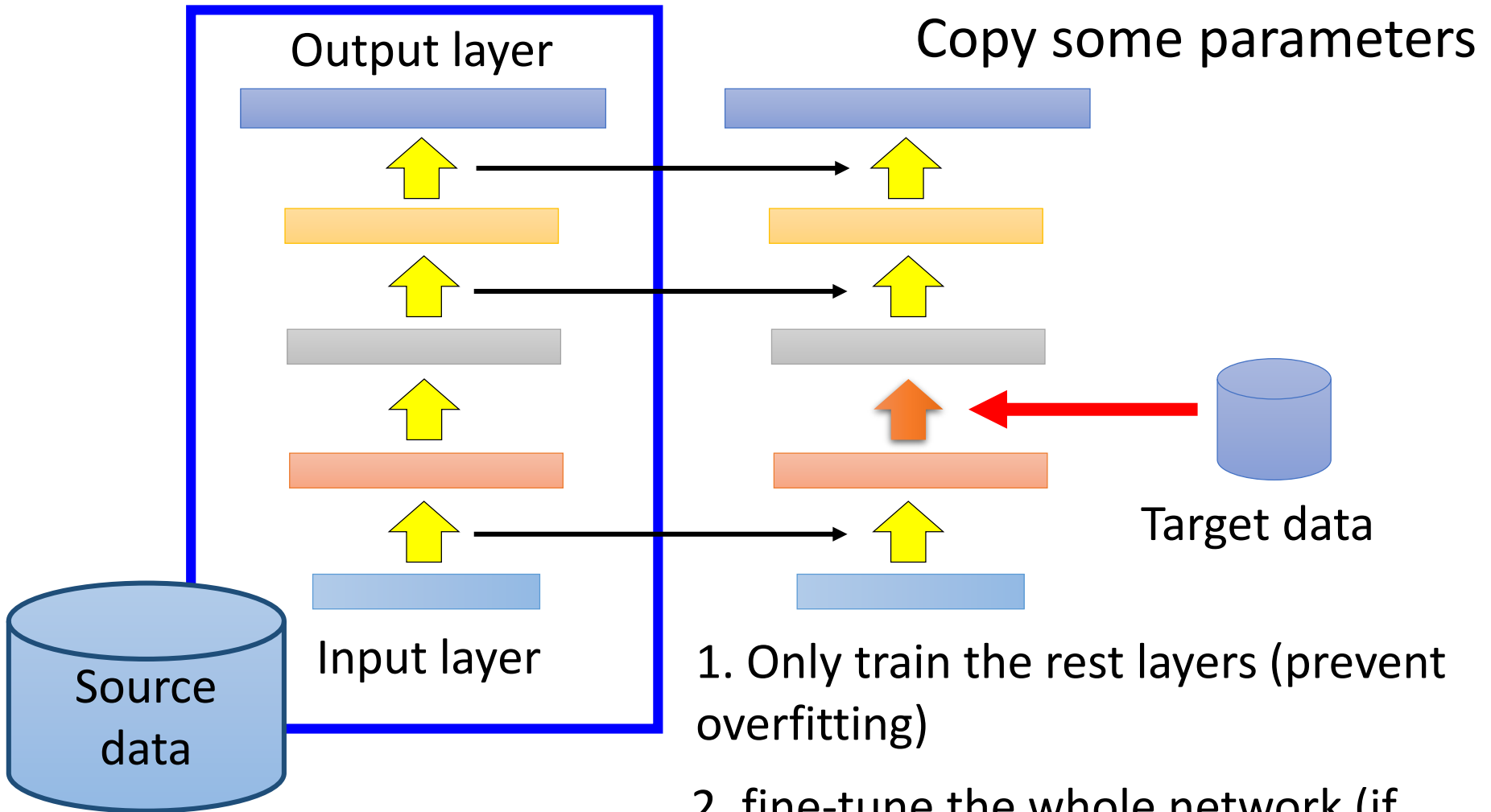
One-shot learning: only a few examples in target domain

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



# Layer Transfer



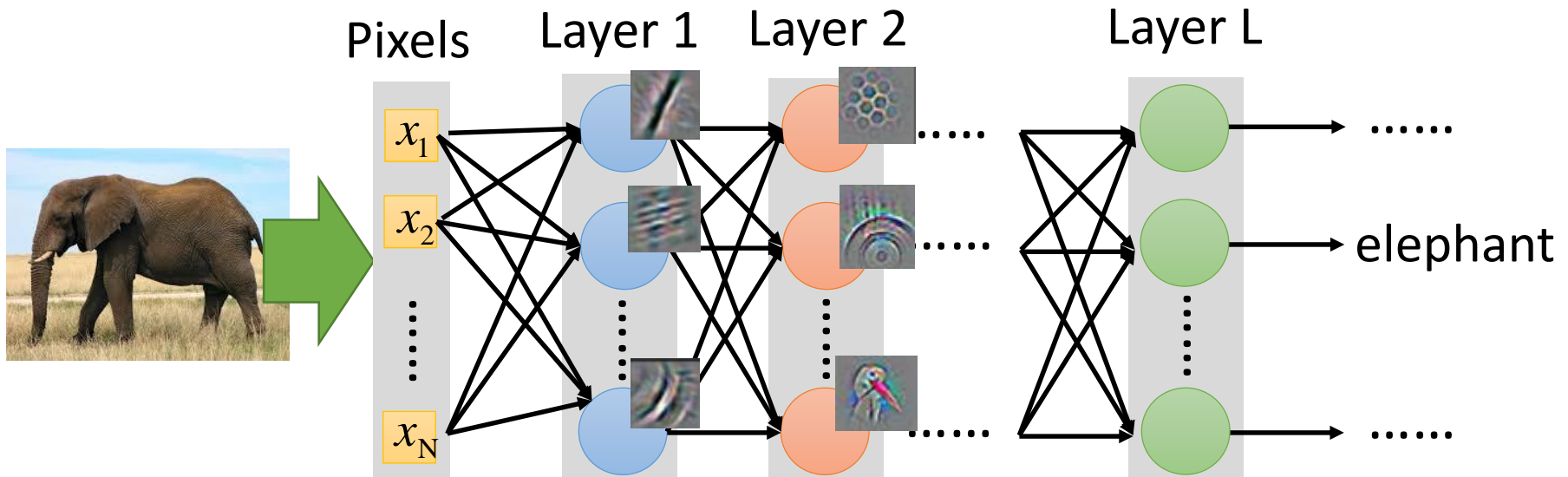
1. Only train the rest layers (prevent overfitting)

2. fine-tune the whole network (if there is sufficient data)



# Layer Transfer

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

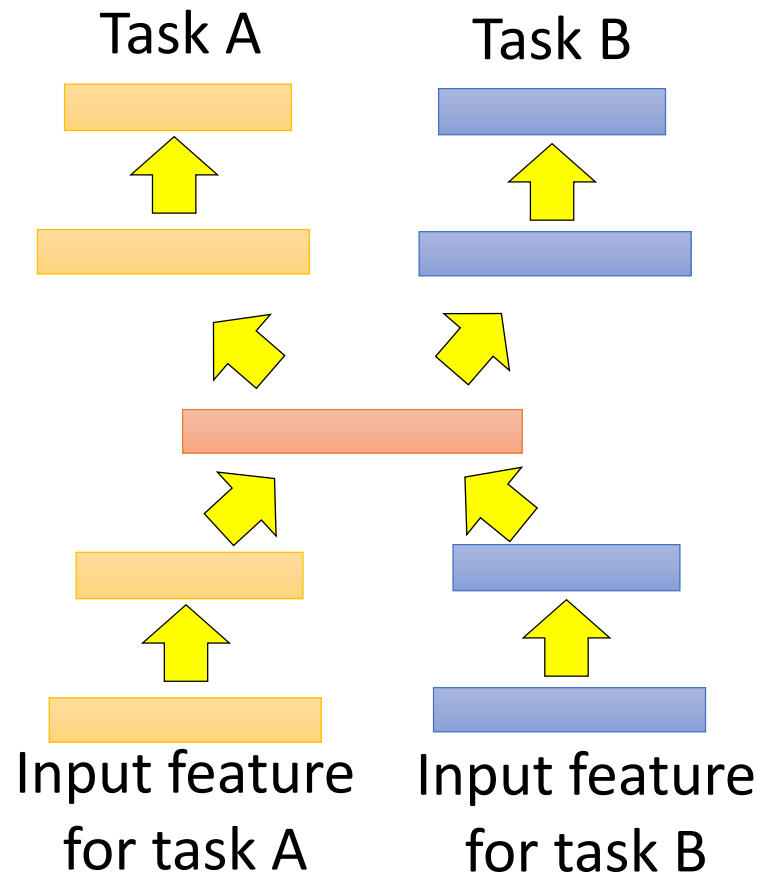
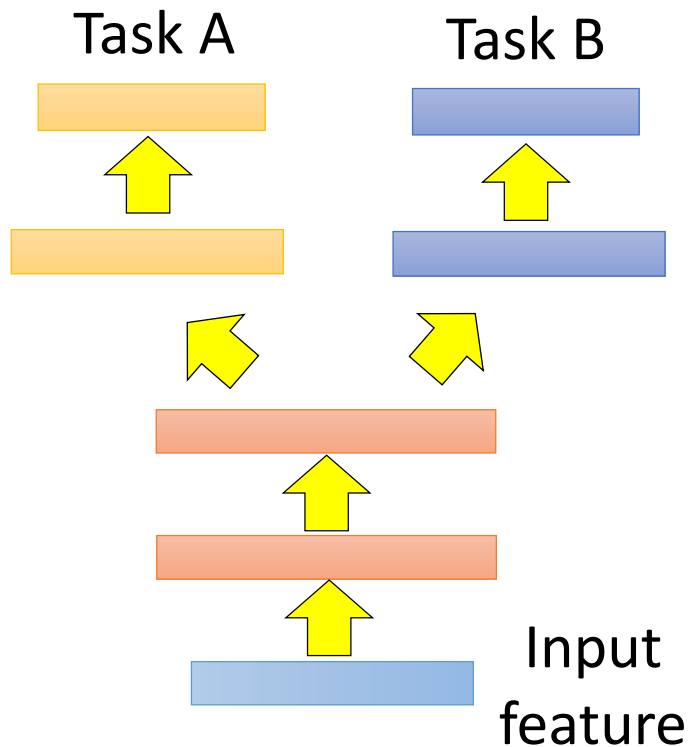


# Transfer Learning - Overview

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

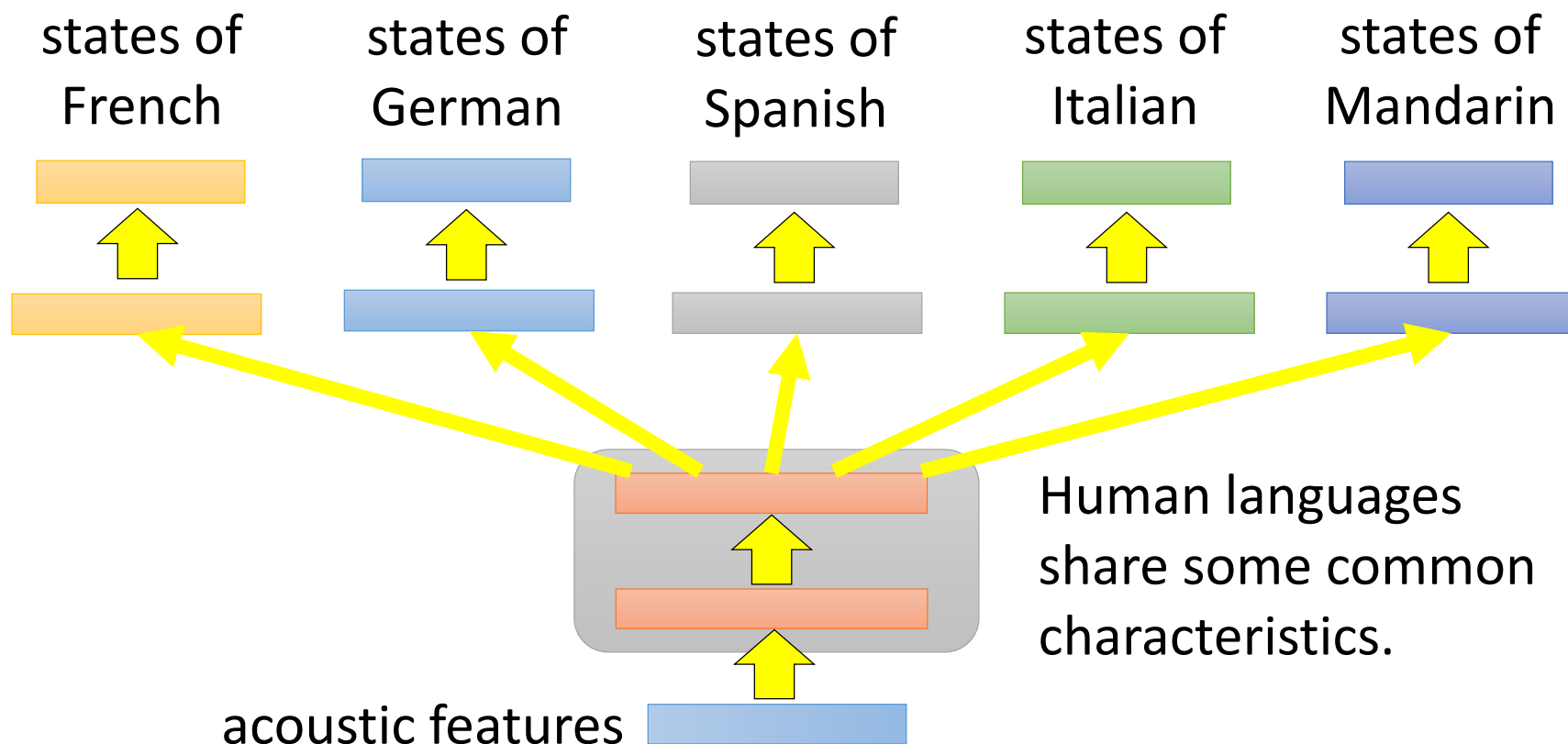
# Multitask Learning

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



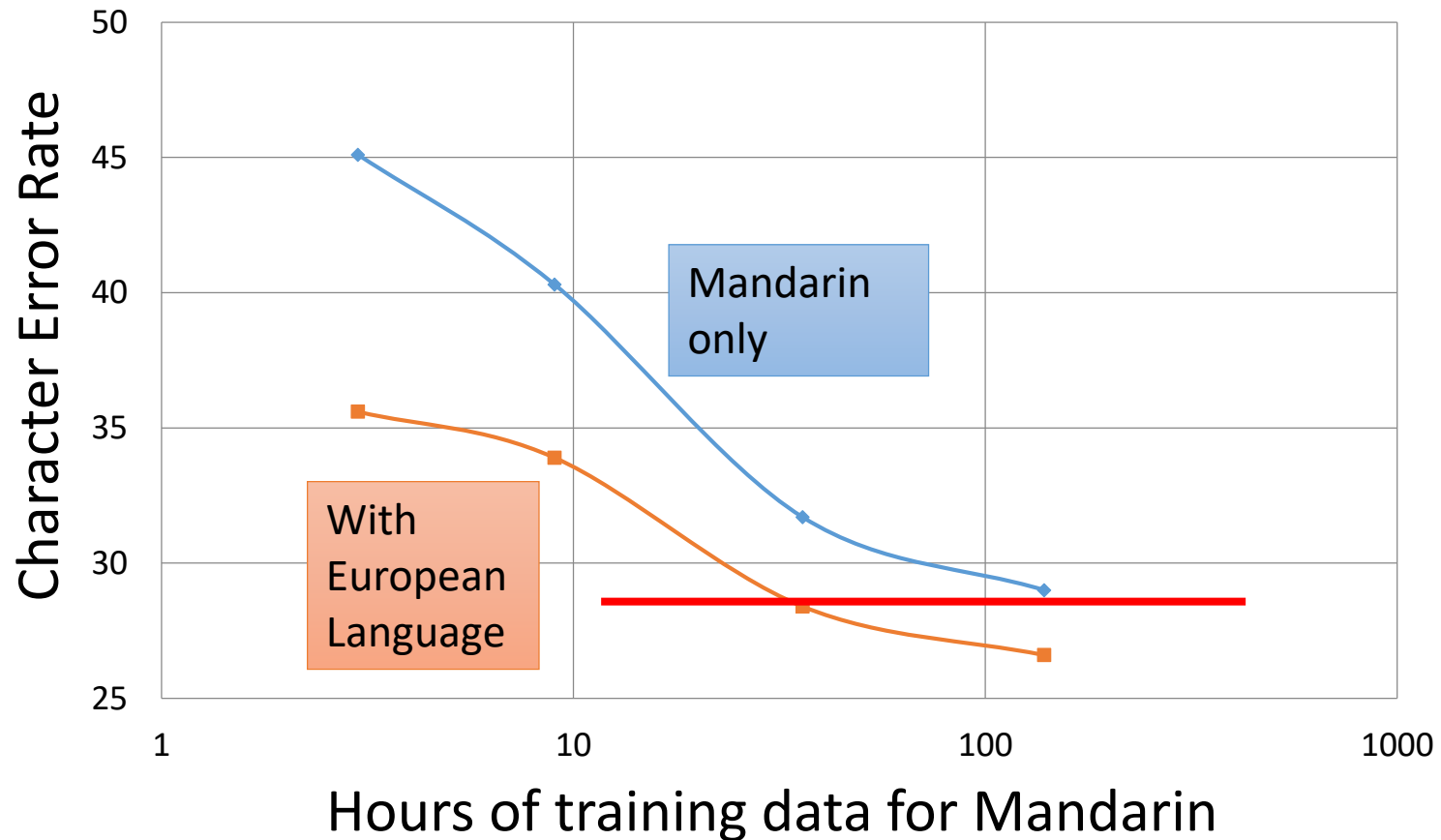
# Multitask Learning

## - Multilingual Speech Recognition



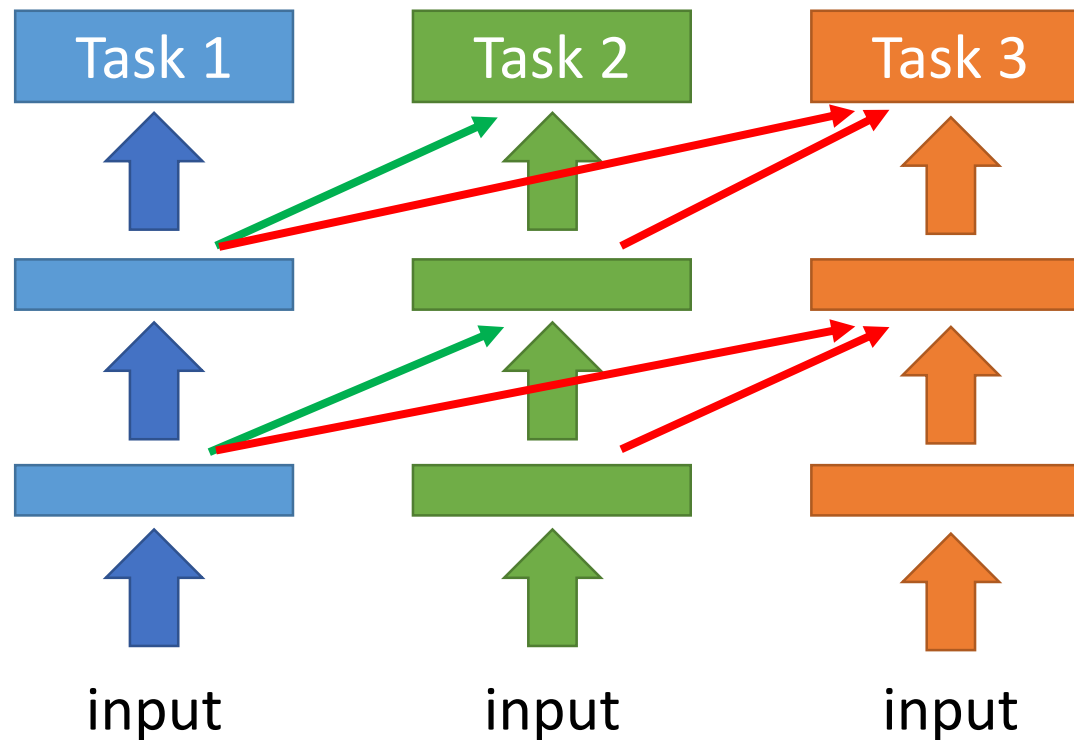
***Similar idea in translation:*** Daxiang Dong, Hua Wu, Wei He, Dianhai Yu and Haifeng Wang, "Multi-task learning for multiple language translation.", ACL 2015

# Multitask Learning - Multilingual



Huang, Jui-Ting, et al. "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers." *ICASSP, 2013*

# Progressive Neural Networks



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

# Transfer Learning - Overview

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

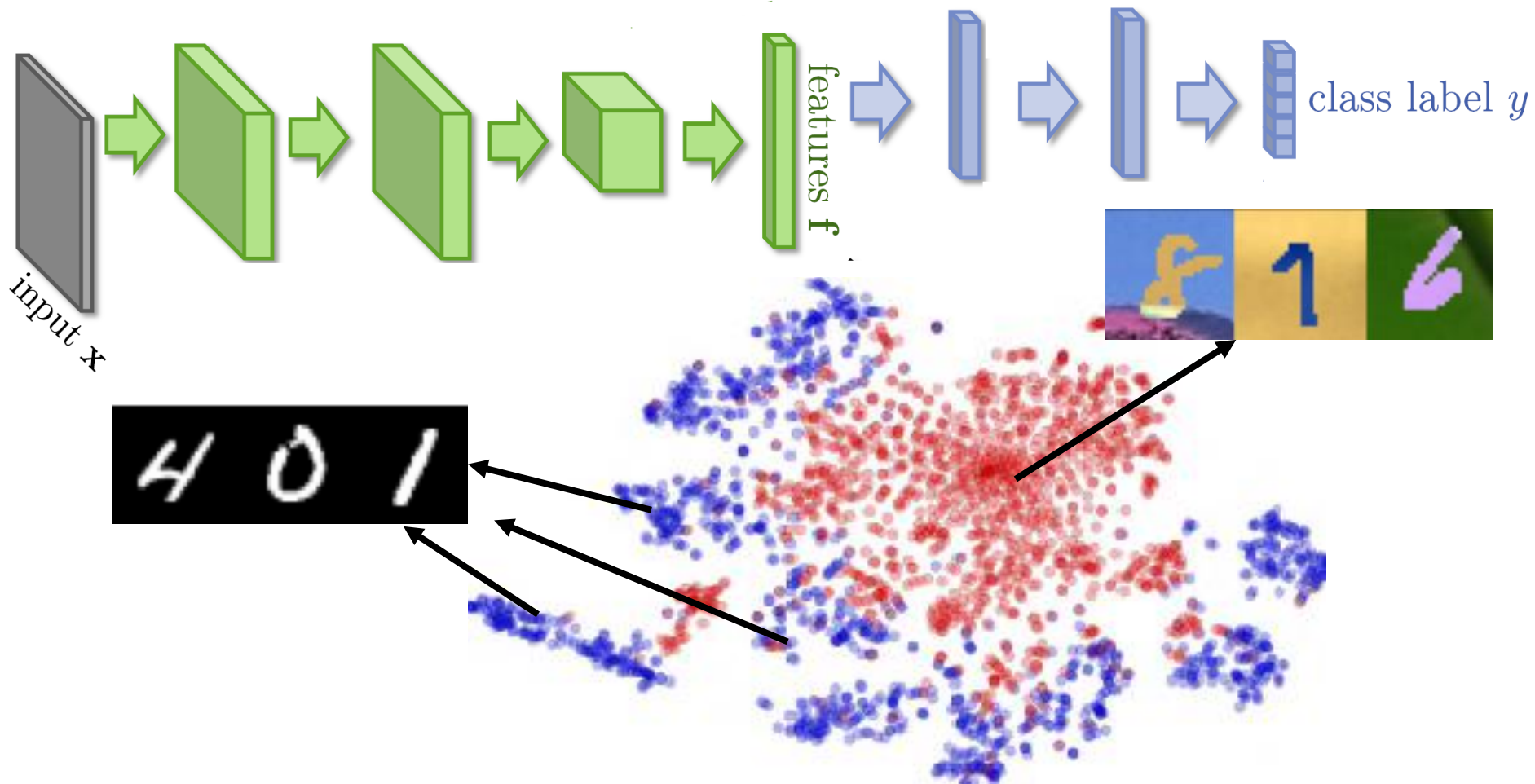
# Task description

- Source data:  $(x^s, y^s) \longrightarrow$  Training data
  - Target data:  $(x^t) \longrightarrow$  Testing data
- } Same task, mismatch

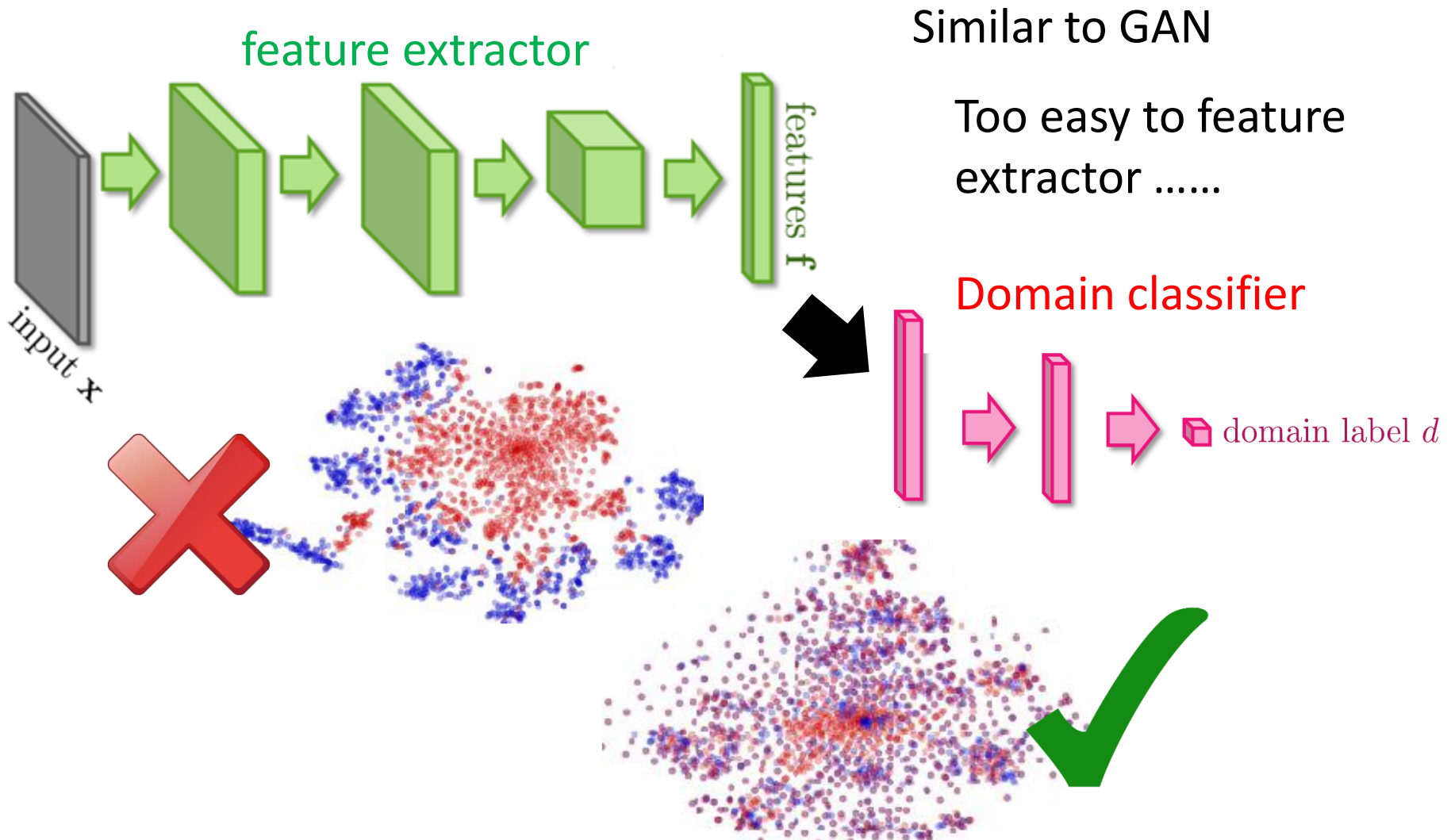




# Domain-adversarial training



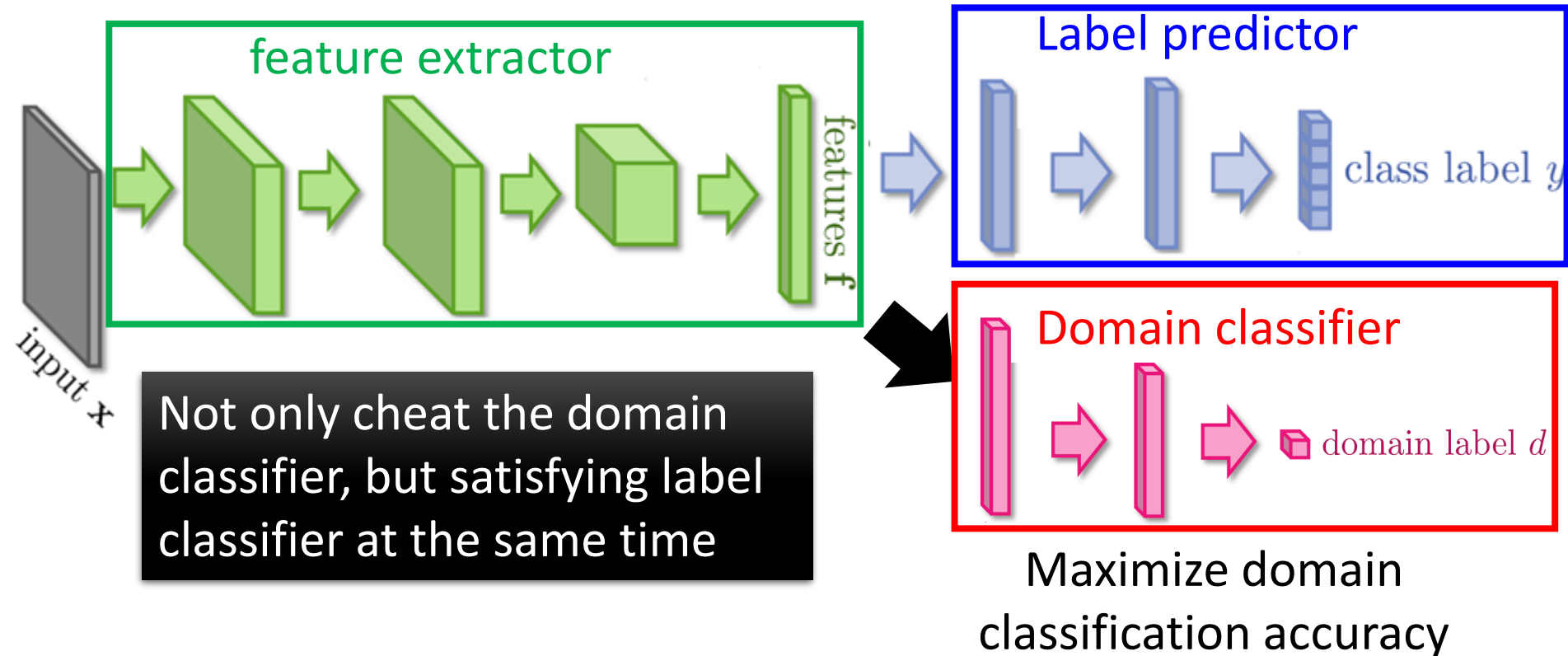
# Domain-adversarial training



# Domain-adversarial training

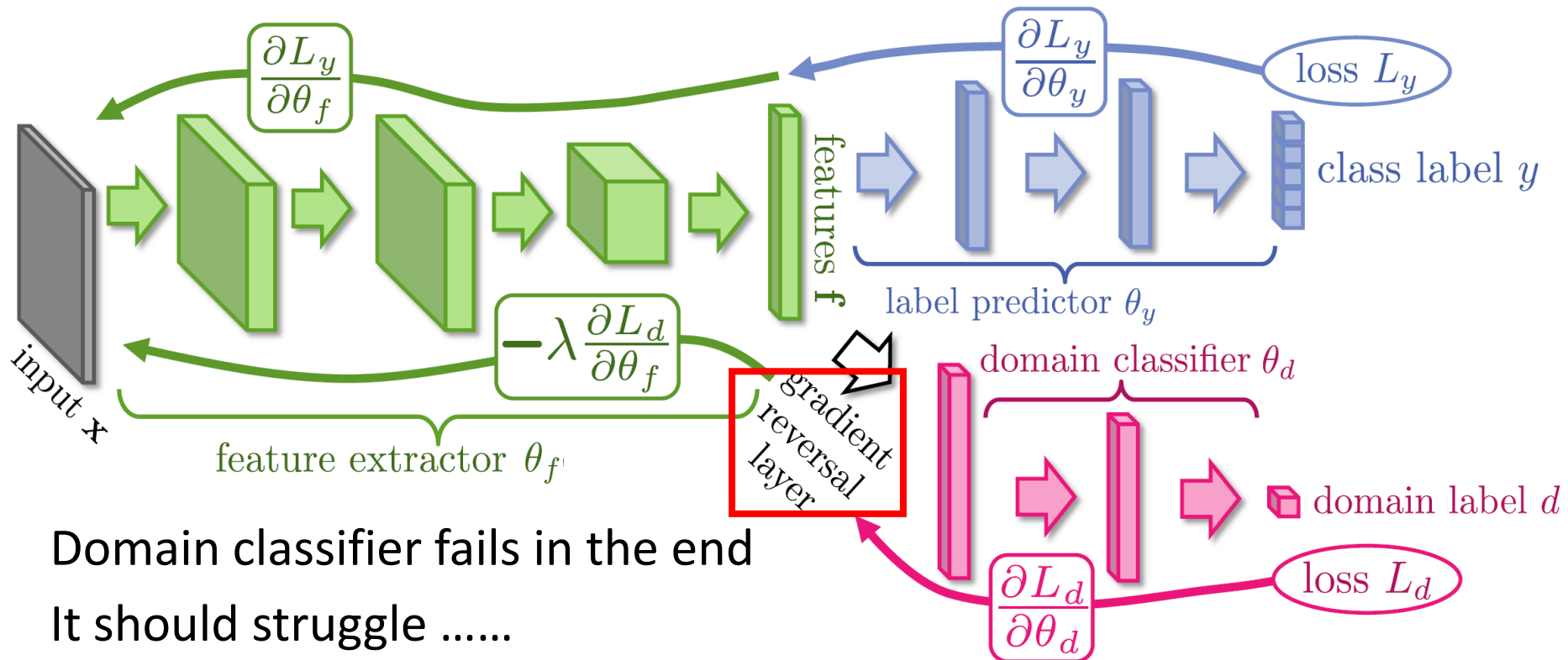
Maximize label classification accuracy +  
minimize domain classification accuracy

Maximize label  
classification accuracy



This is a big network, but different parts have different goals.

# 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



METHOD	SOURCE	MNIST	SYN NUMBERS	SVHN	SYN SIGNS
	TARGET	MNIST-M	SVHN	MNIST	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		<b>.8149</b> (57.9%)	<b>.9048</b> (66.1%)	<b>.7107</b> (29.3%)	<b>.8866</b> (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

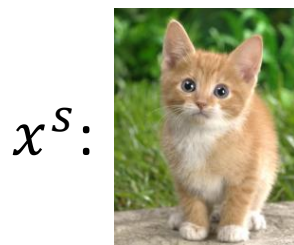
# Transfer Learning - Overview

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

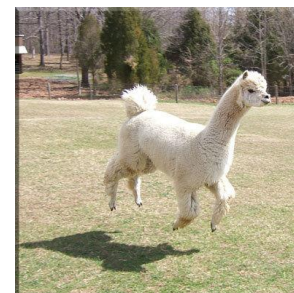
# Zero-shot Learning

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

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



.....



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

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

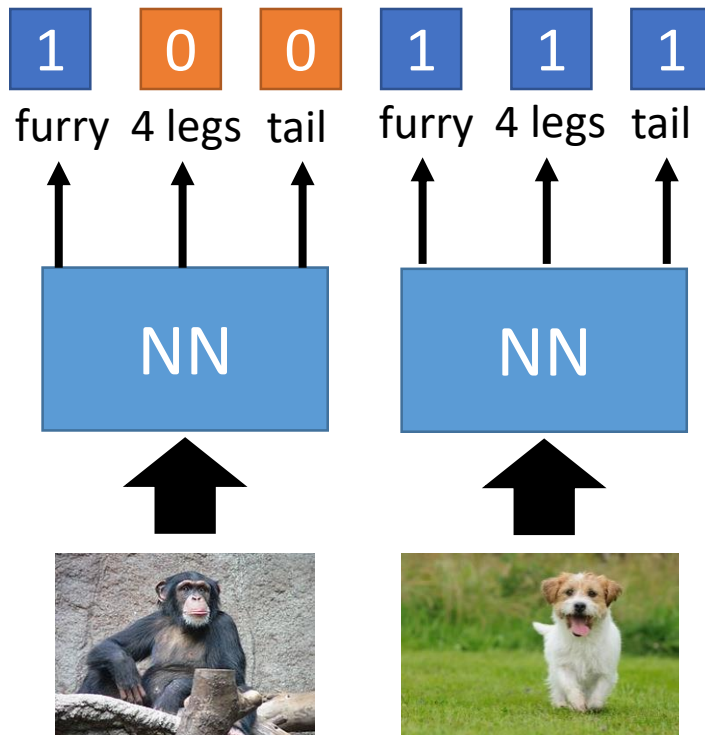
How we solve this problem in speech recognition?



# Zero-shot Learning

- Representing each class by its attributes

## Training



## Database

### attributes

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

class

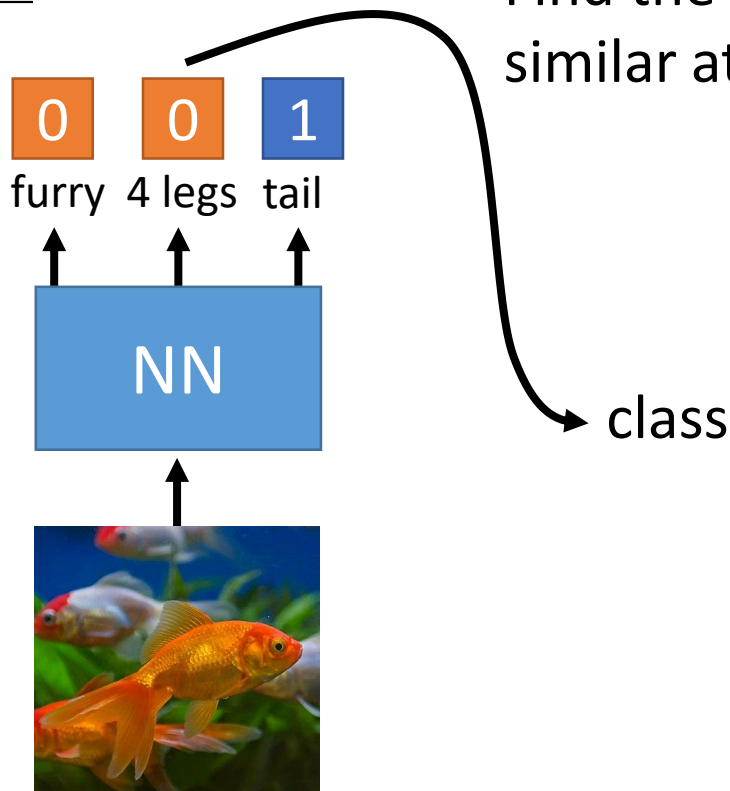
sufficient attributes for one to one mapping



# Zero-shot Learning

- Representing each class by its attributes

## Testing



Find the class with the most similar attributes

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

sufficient attributes for one to one mapping

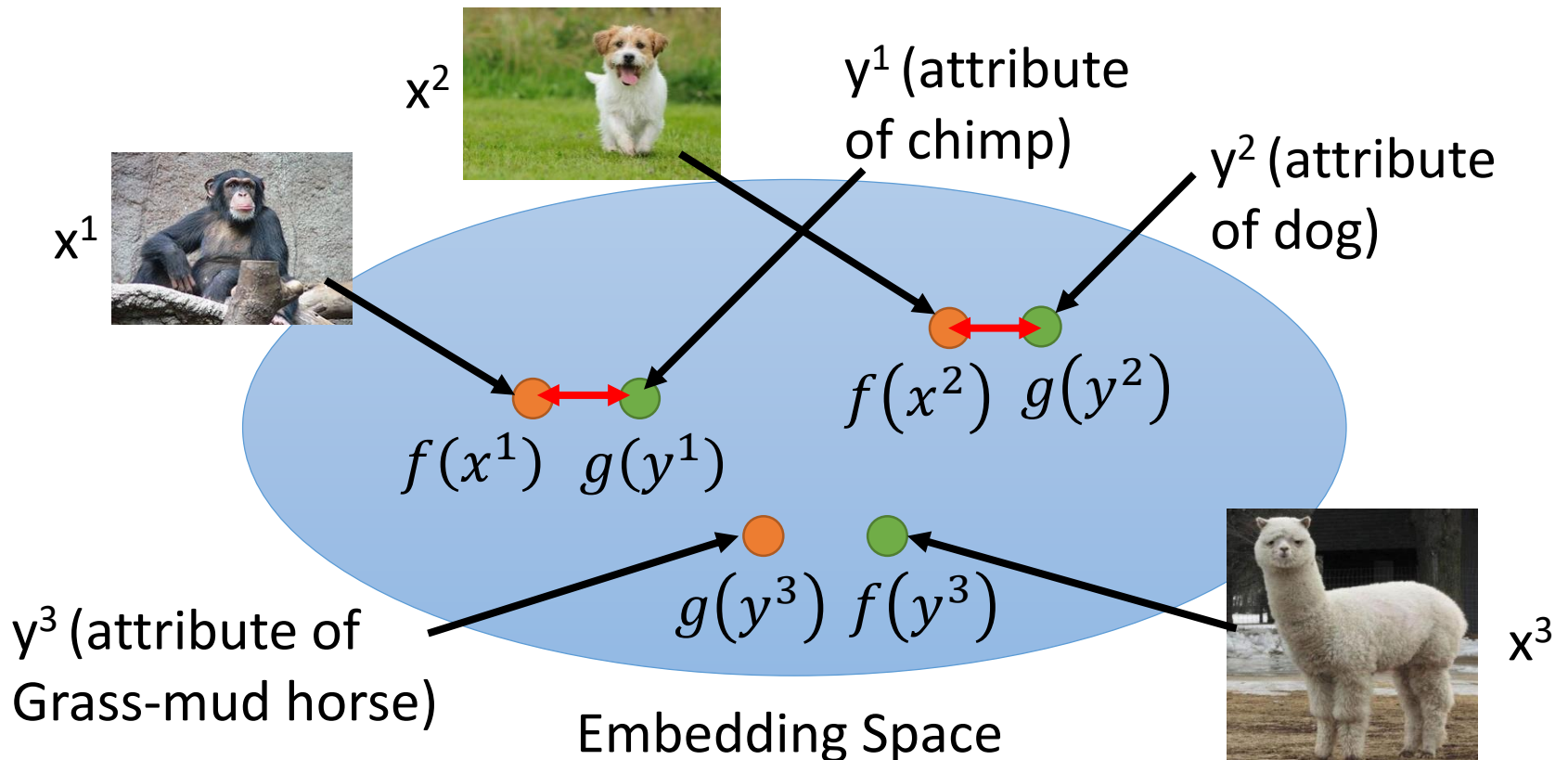
# Zero-shot Learning

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

Training target:

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

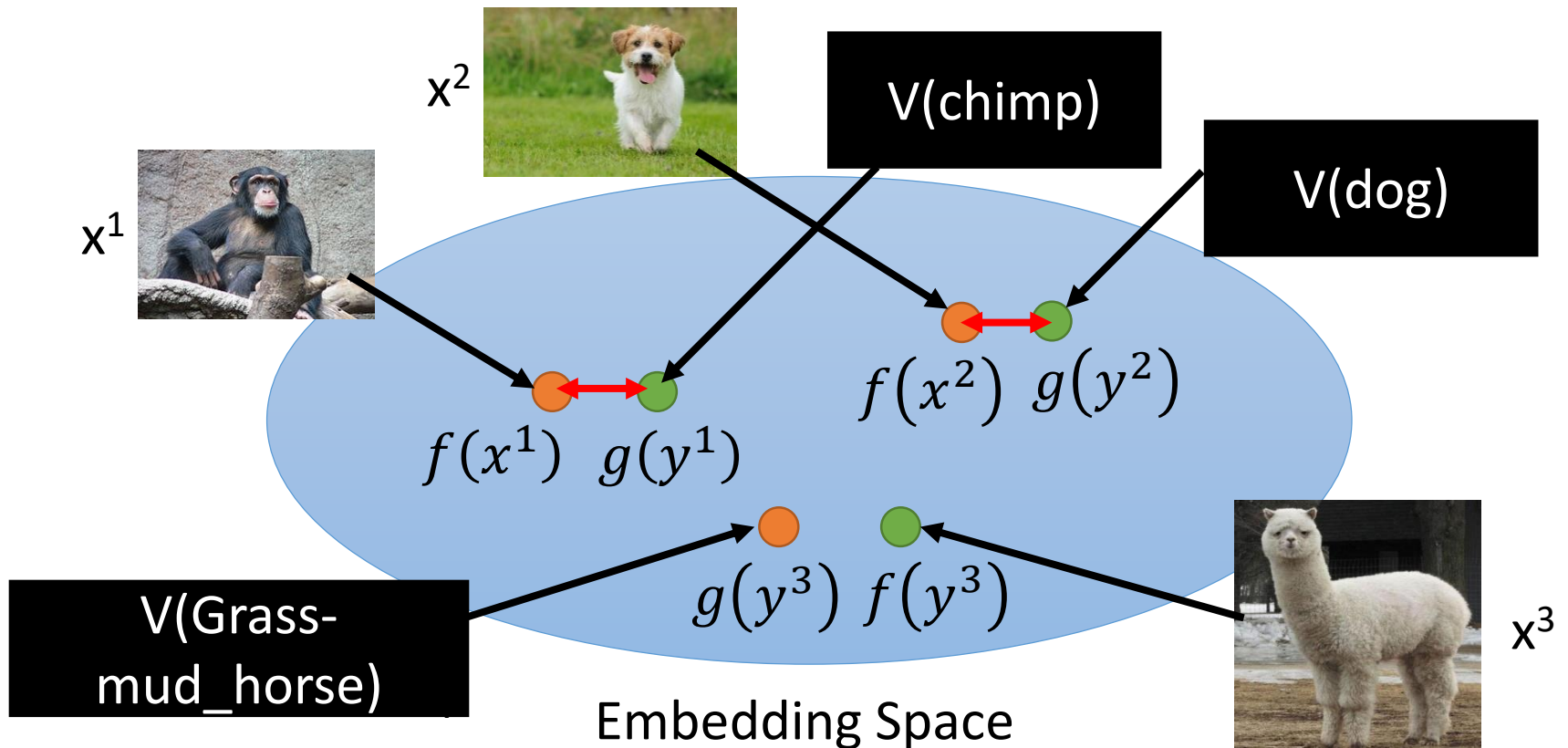
- Attribute embedding



# Zero-shot Learning

What if we don't have database

- Attribute embedding + word embedding



# Zero-shot Learning

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

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

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

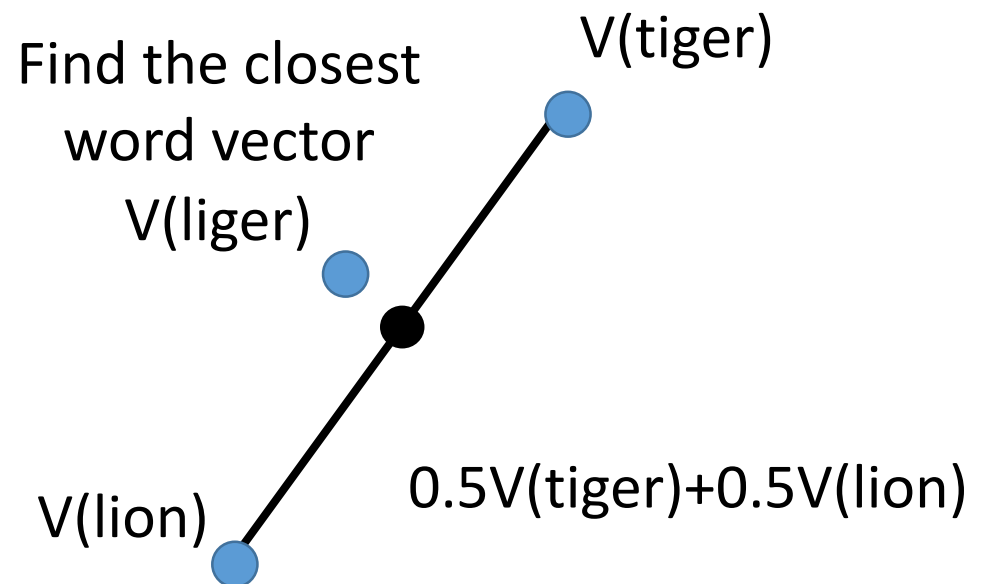
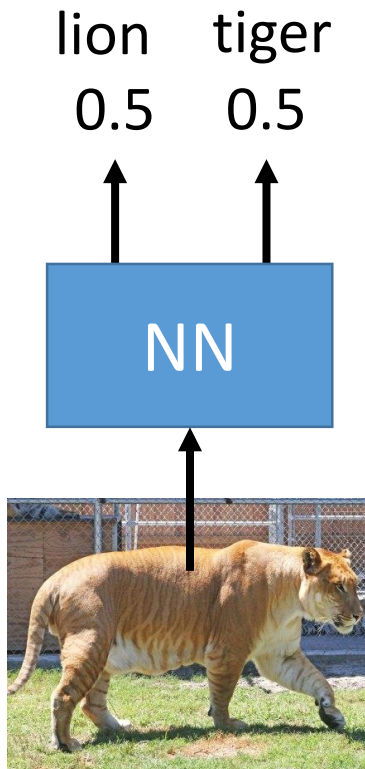
$$\underline{f(x^n) \cdot g(y^n)} - \underline{\max_{m \neq n} f(x^n) \cdot g(y^m)} > k$$

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

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

# Zero-shot Learning

- Convex Combination of Semantic Embedding

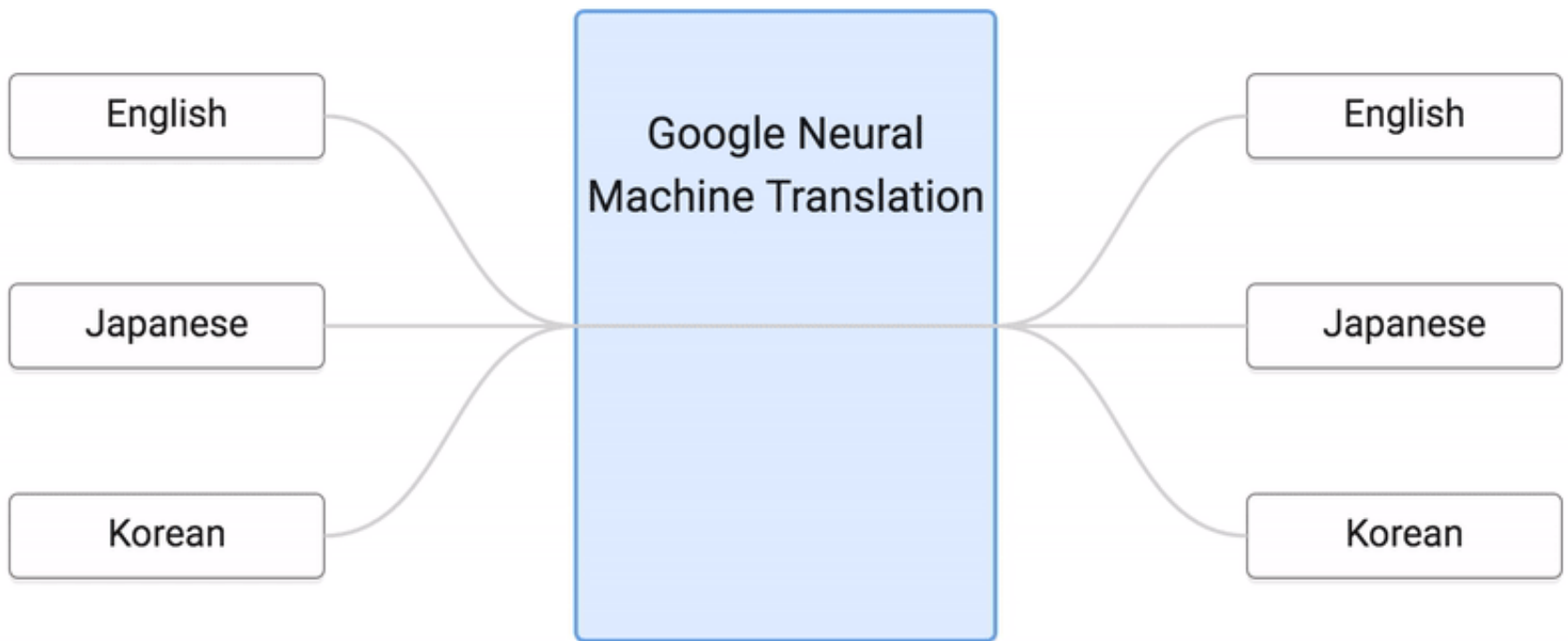


Only need off-the-shelf NN for ImageNet and word vector

Test Image	ConvNet	DeViSE	ConSE(10)

# Example of Zero-shot Learning

Training



Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation, arXiv preprint 2016

# Transfer Learning - Overview

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

**Different from semi-supervised learning**



# 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 <i>different</i> 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

# Appendix

# 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