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

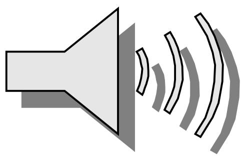
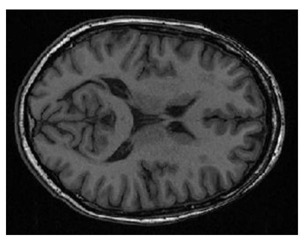
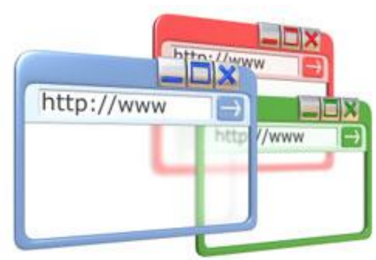
domain:feature space/probability

task:label space/objective predictive function

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	 Webpages

数据少/标记数据少

Transfer Learning

- Example in real life

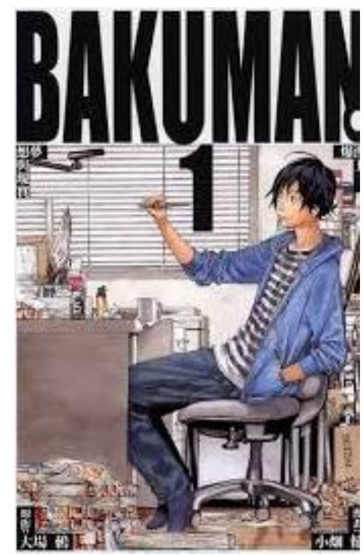
<u>*研究生 on-line*</u>

<u>*漫畫家 on-line*</u> 真城/高木

研究生
生存守則

研究生	←	漫畫家
指導教授	←	責編
跑實驗	←	畫分鏡
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

transfer Learning 是用来做one-shot learning的一种方式

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

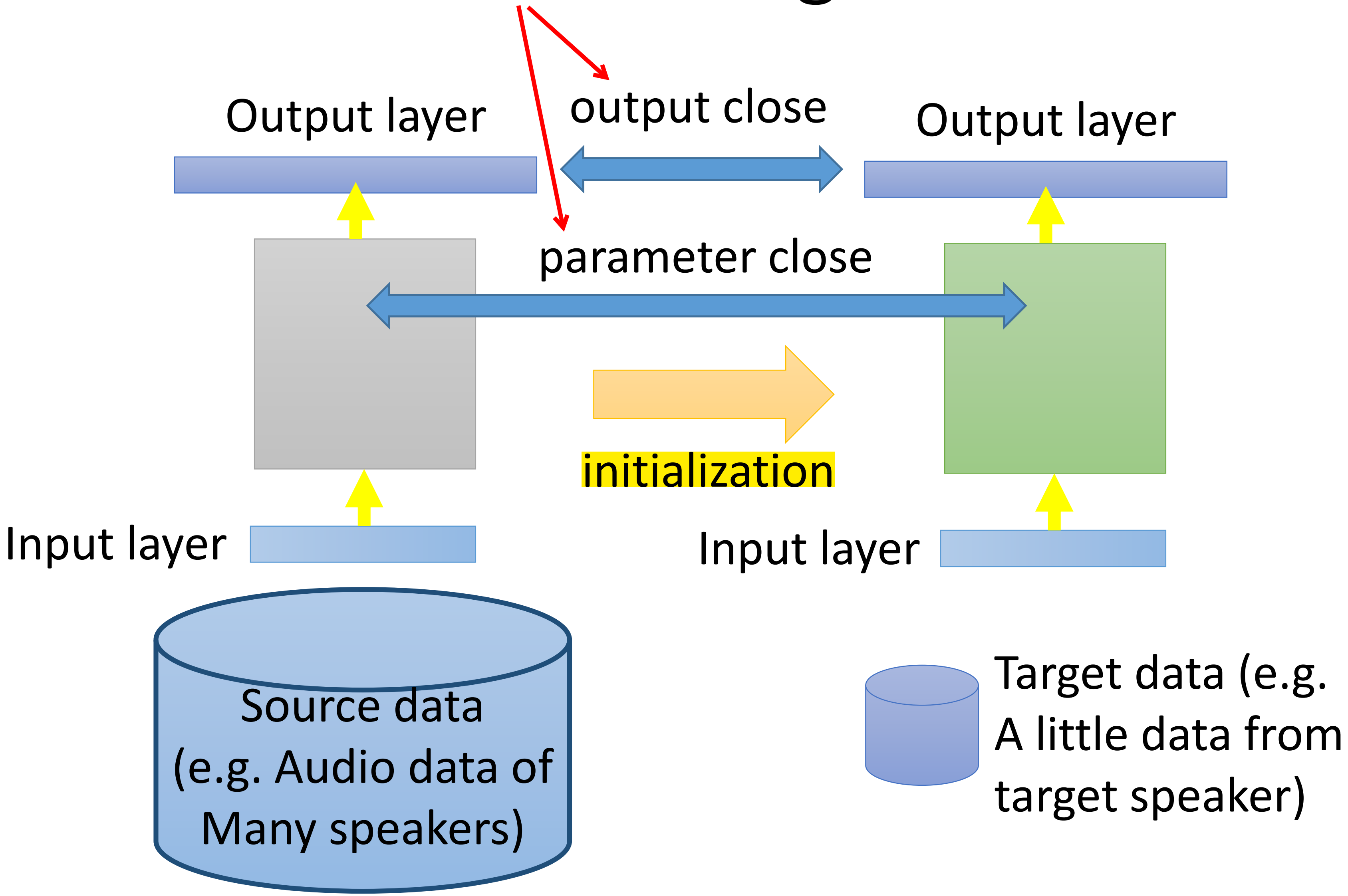
- Task description
  - Target data: (x^t, y^t) Very little
  - Source data: (x^s, y^s) A large amount

不一定shot one：可以有少量多个

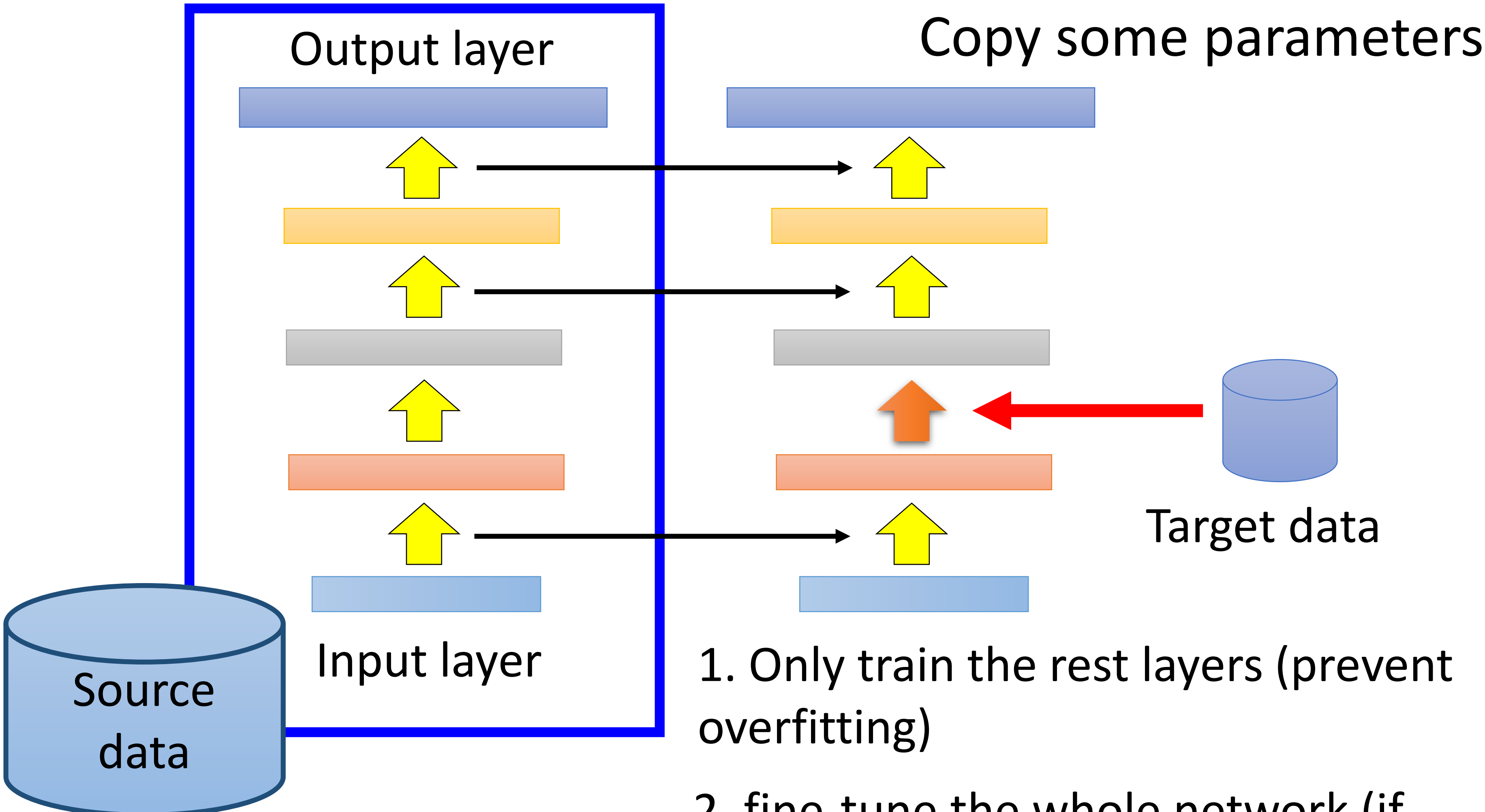
- Example: (supervised) speaker adaption
  - Target data: audio data and its transcriptions of specific user

    有帮助
  - 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



Layer Transfer



1. Only train the rest layers (prevent overfitting)

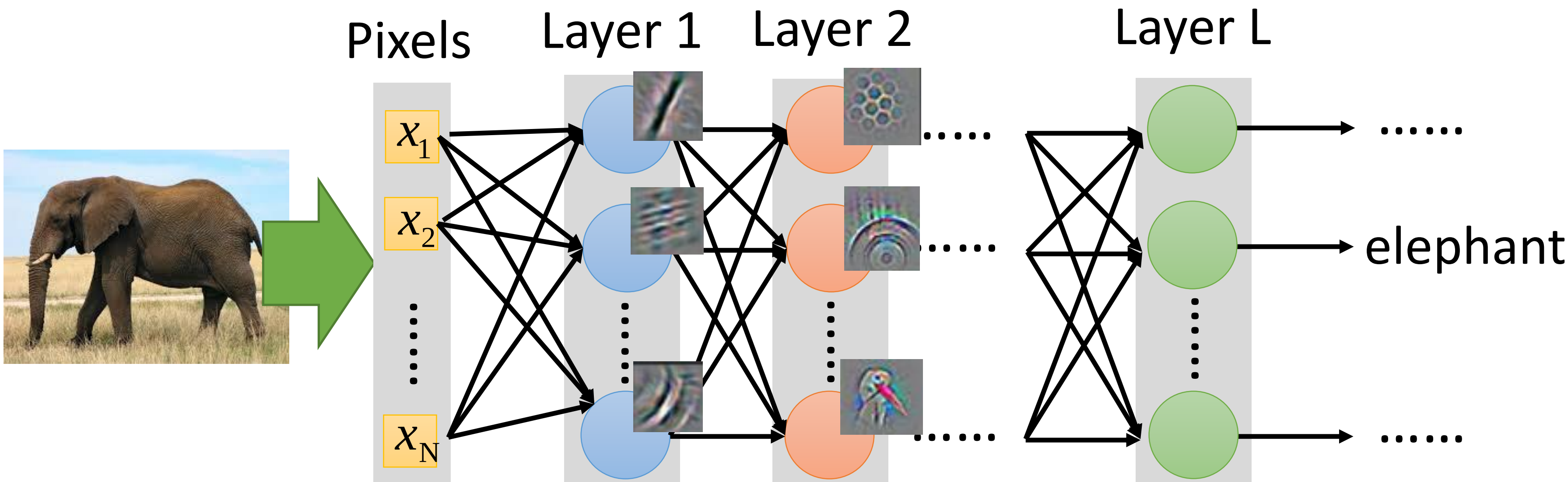
2. fine-tune the whole network (if there is sufficient data)

Layer Transfer

运用之妙，存乎一心

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

语音到发音方式往往不同，但是发音方式到词汇往往相似

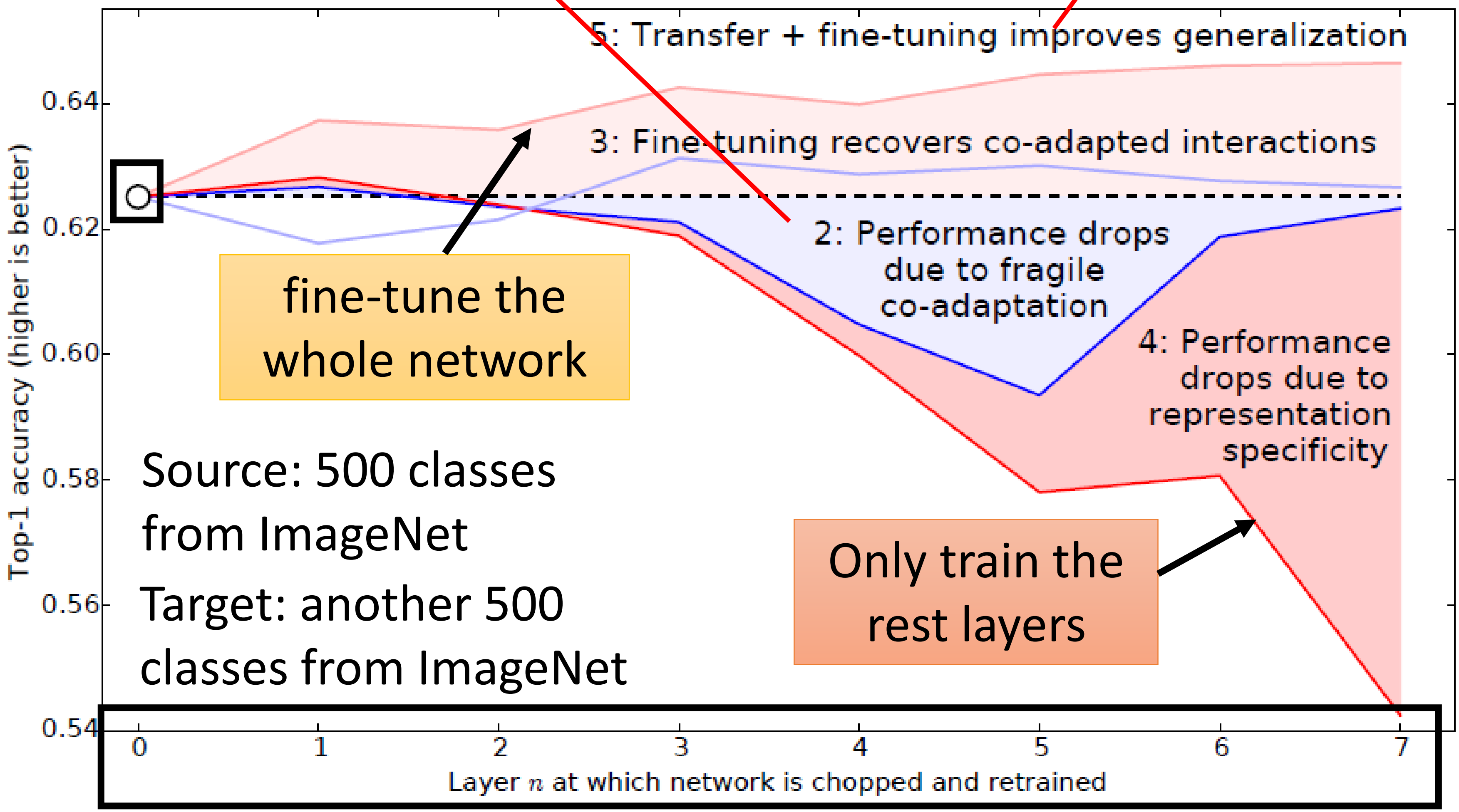


最简单的pattern往往通用

先整体train，再fix前面几层，train后面的部分：
由于前后不配，很有可能坏掉

即使target很多，
source也很有用

Layer Transfer - Image

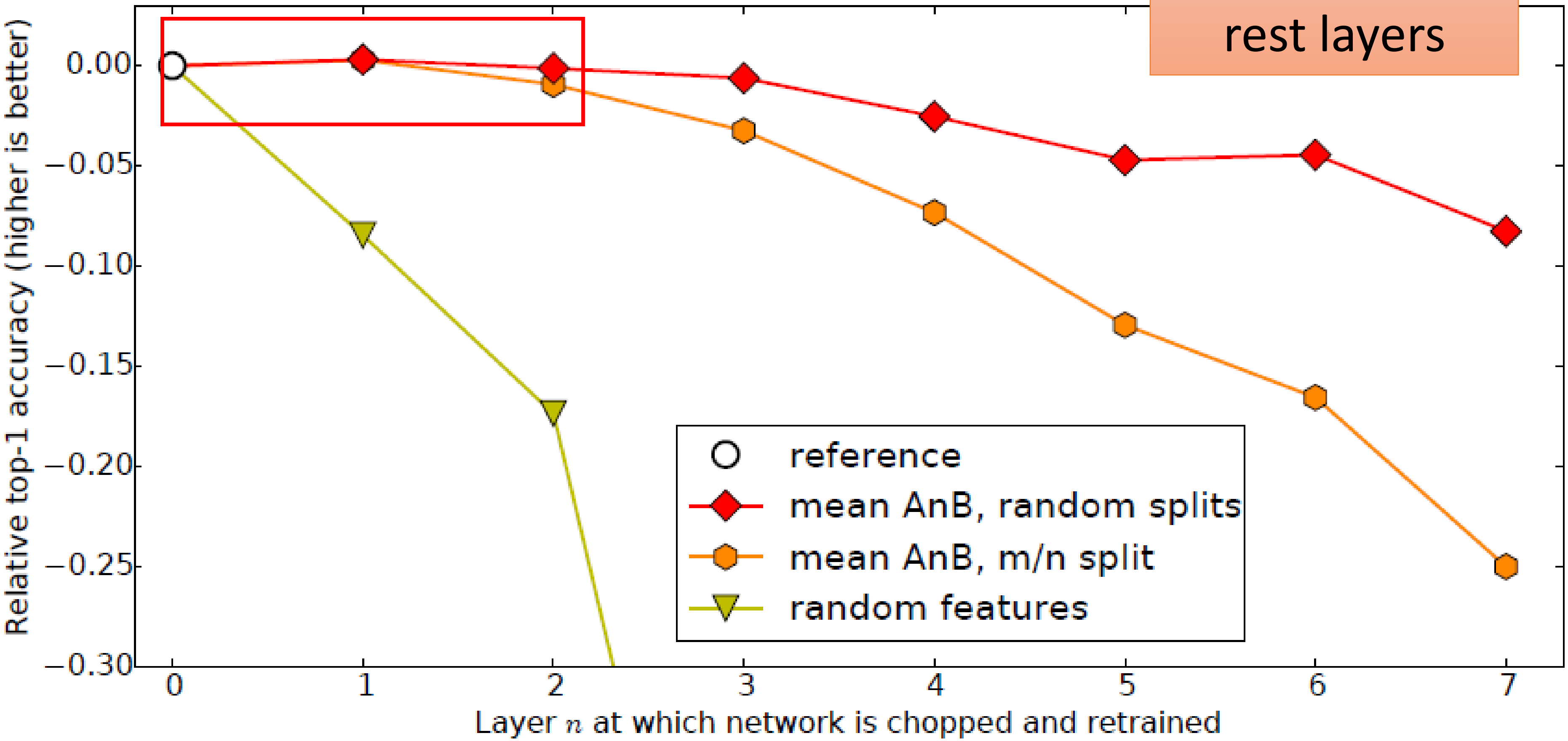


Jason Yosinski, Jeff Clune, Yoshua Bengio, Hod Lipson, “How transferable are features in deep neural networks?”, NIPS, 2014

Layer Transfer - Image

如果source和target
差别比较大，最好只
用前面的几层，不要太多

Only train the rest layers



Jason Yosinski, Jeff Clune, Yoshua Bengio, Hod Lipson, “How transferable are features in deep neural networks?”, NIPS, 2014

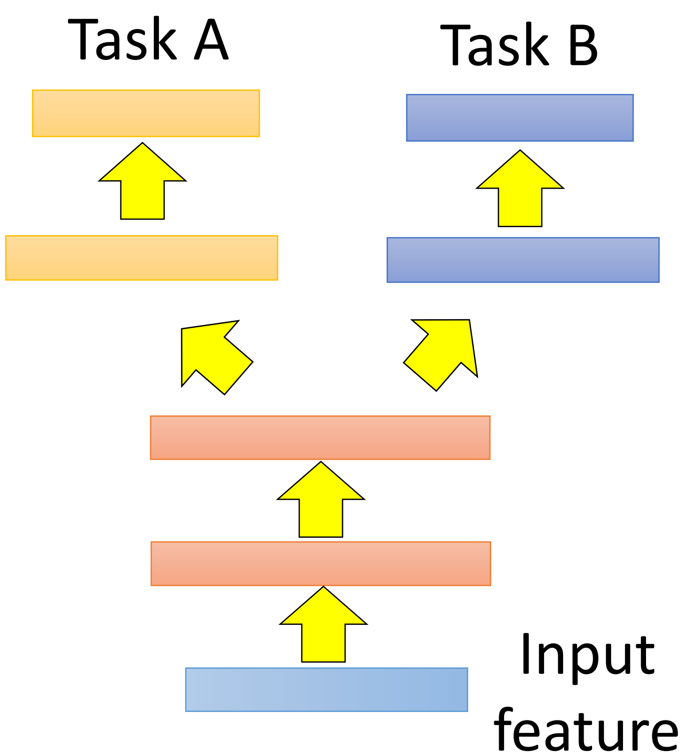
Transfer Learning - Overview

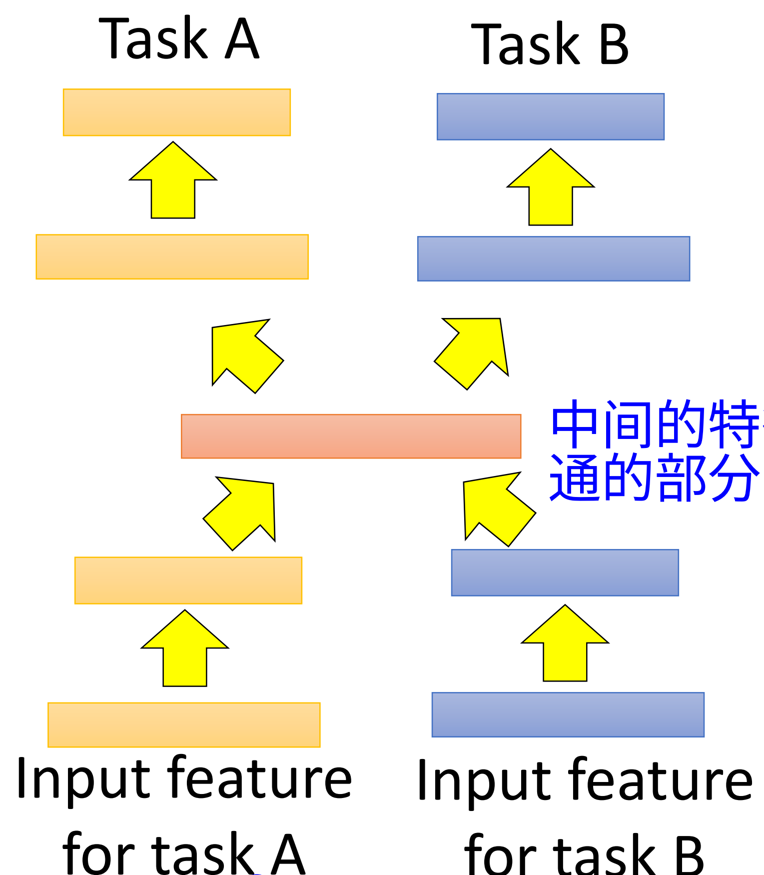
		Source Data (not directly related to the task)	
		labelled	unlabeled
Target Data	labelled	Fine-tuning 只关心target结果好不好，fine-tune后source domain 坏掉就坏掉 Multitask Learning 同时care source与target	
	unlabeled		

Multitask Learning

传统的ML不容易multi-task

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

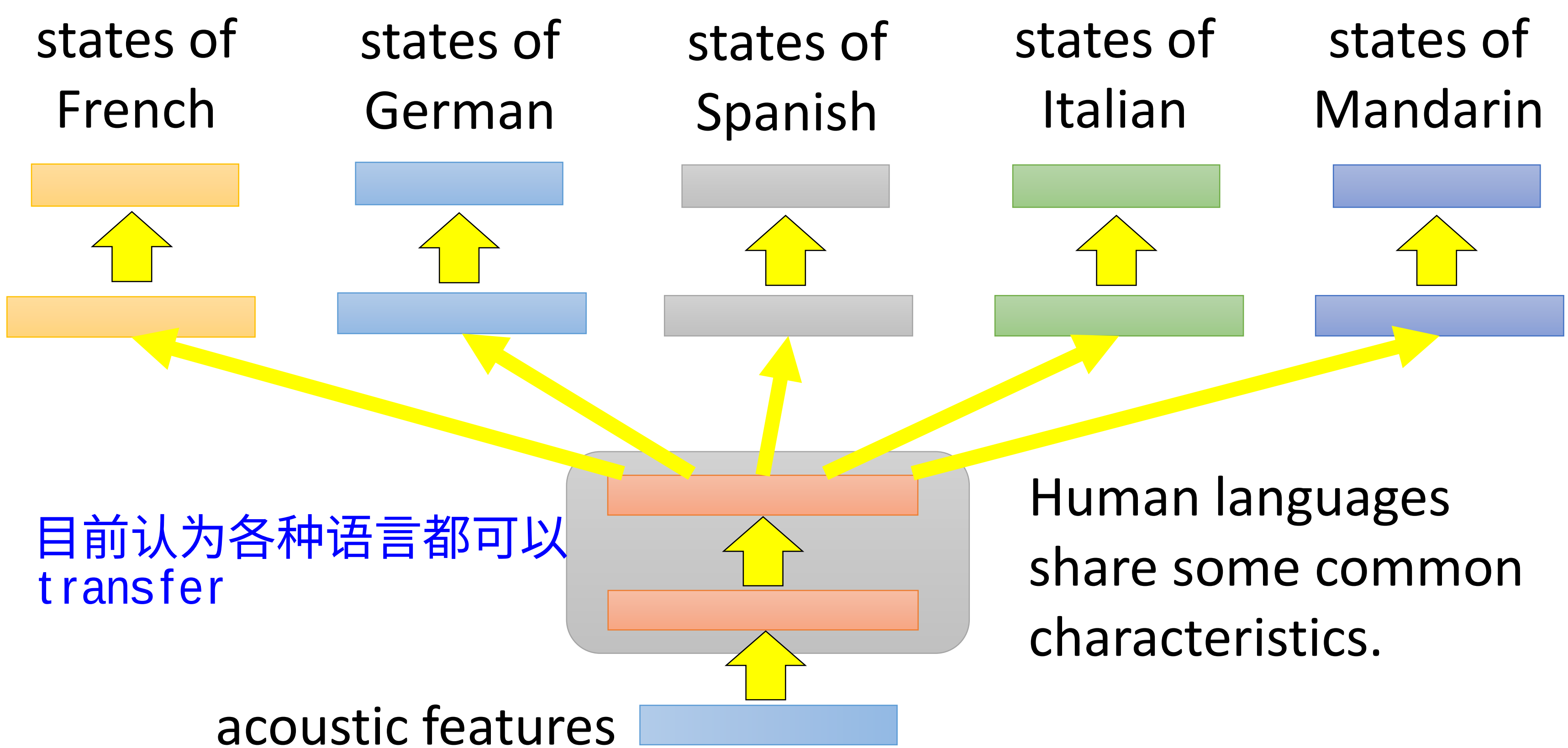




中间的特征有共通的部分

是否能share input？

Multitask Learning
- Multilingual Speech Recognition

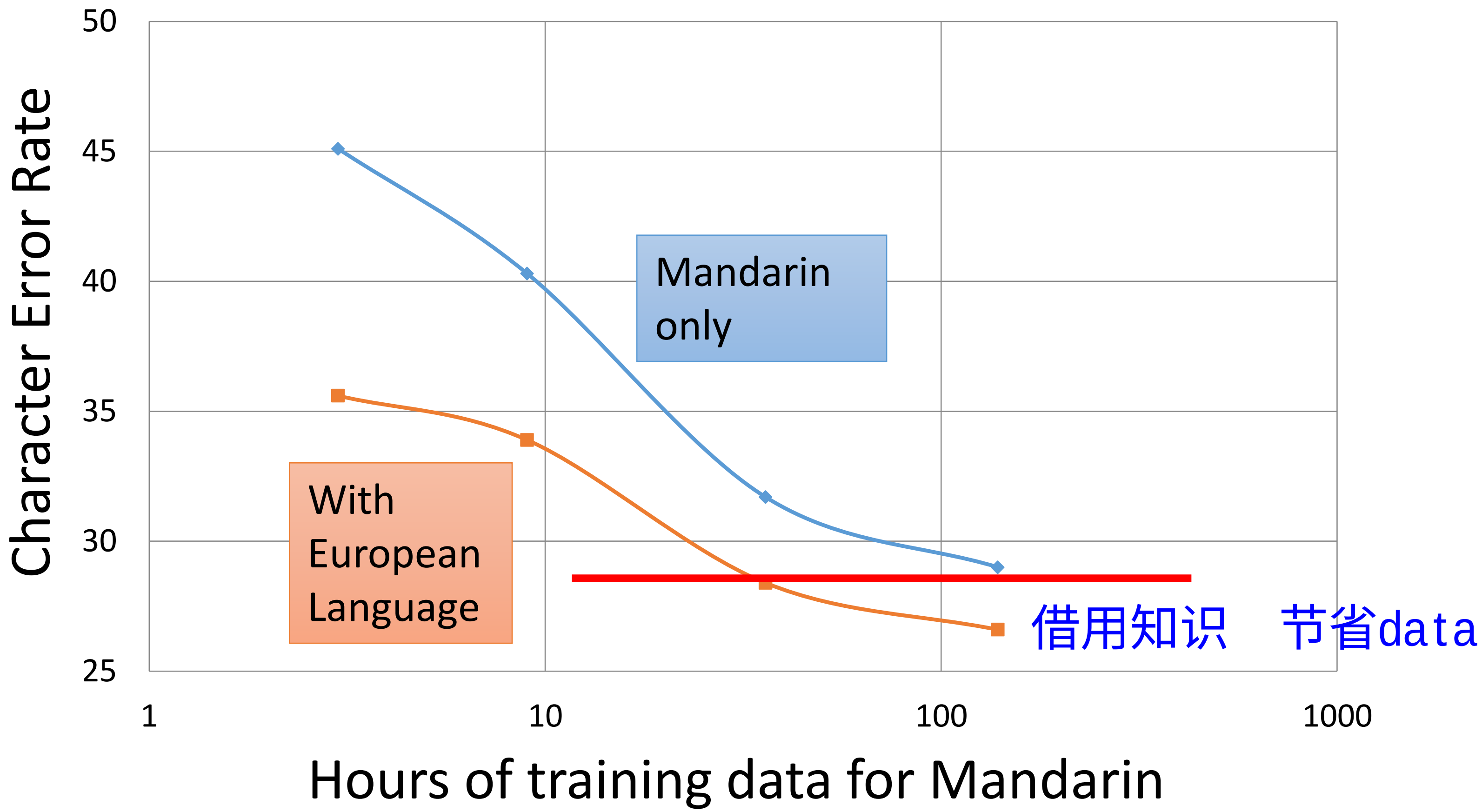


目前认为各种语言都可以 transfer

Human languages share some common characteristics.

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

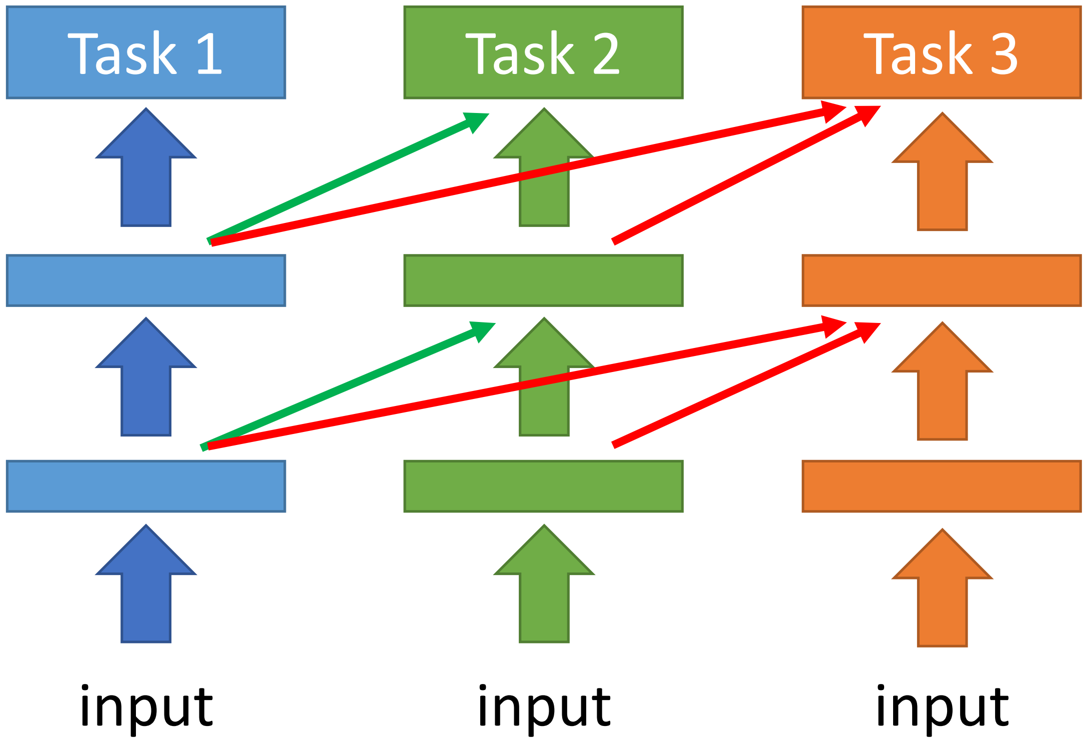
Multitask Learning - Multilingual



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

Progressive Neural Networks

一个任务与一个
点很像



后面的task不会
影响前面的
task，
如果前面的task
对后面的影响不
好，可以令参数
为0

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	Fine-tuning Multitask Learning	
	unlabeled	Domain-adversarial training	

Task description

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

mismatch

MNIST

SOURCE

with label

TARGET

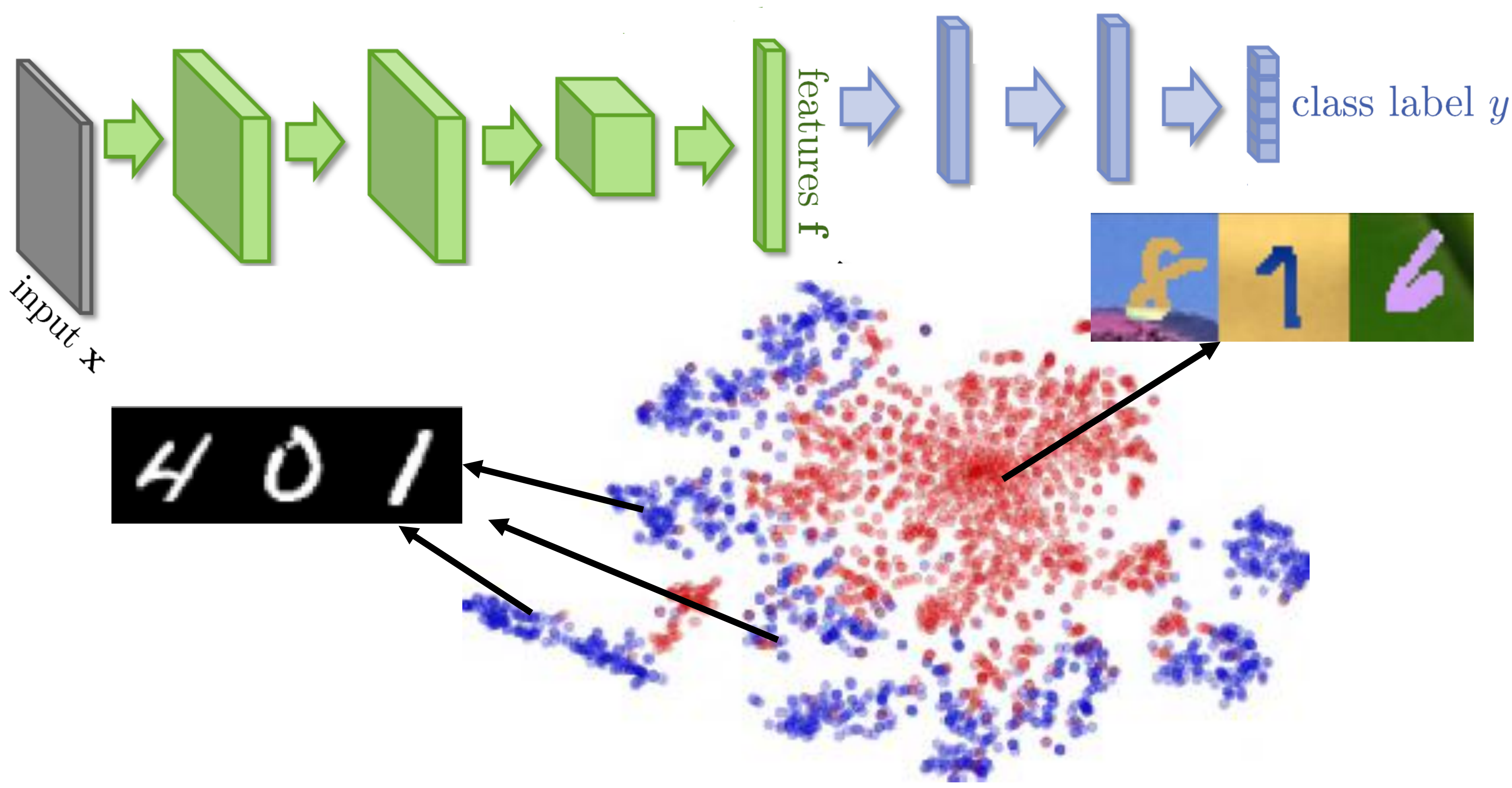
without label

MNIST-M

不同的domain

Domain-adversarial training

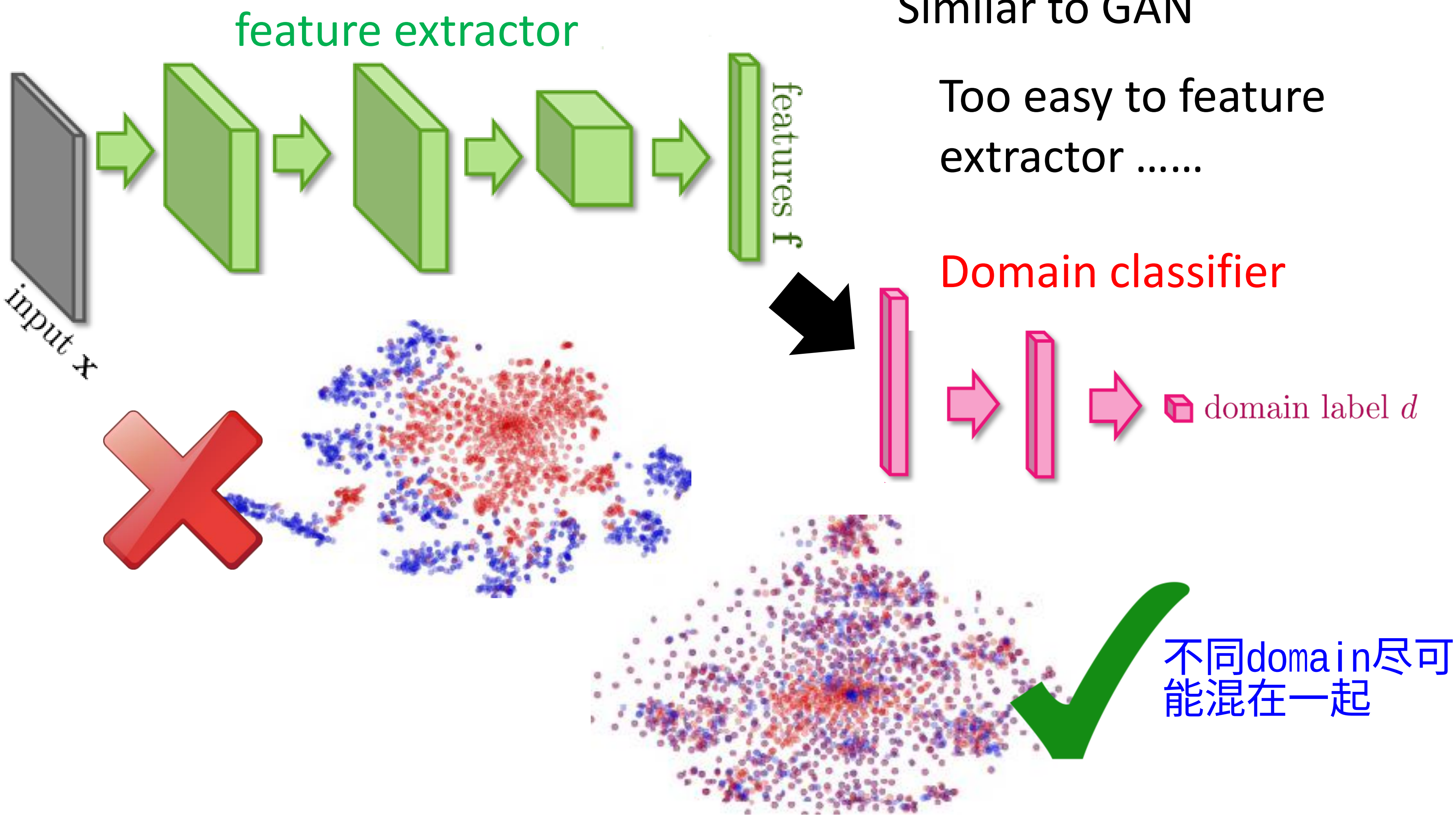
不同的domain，feature很不一样(t-SNE降维)



Domain-adversarial training

Similar to GAN

Too easy to feature extractor ……

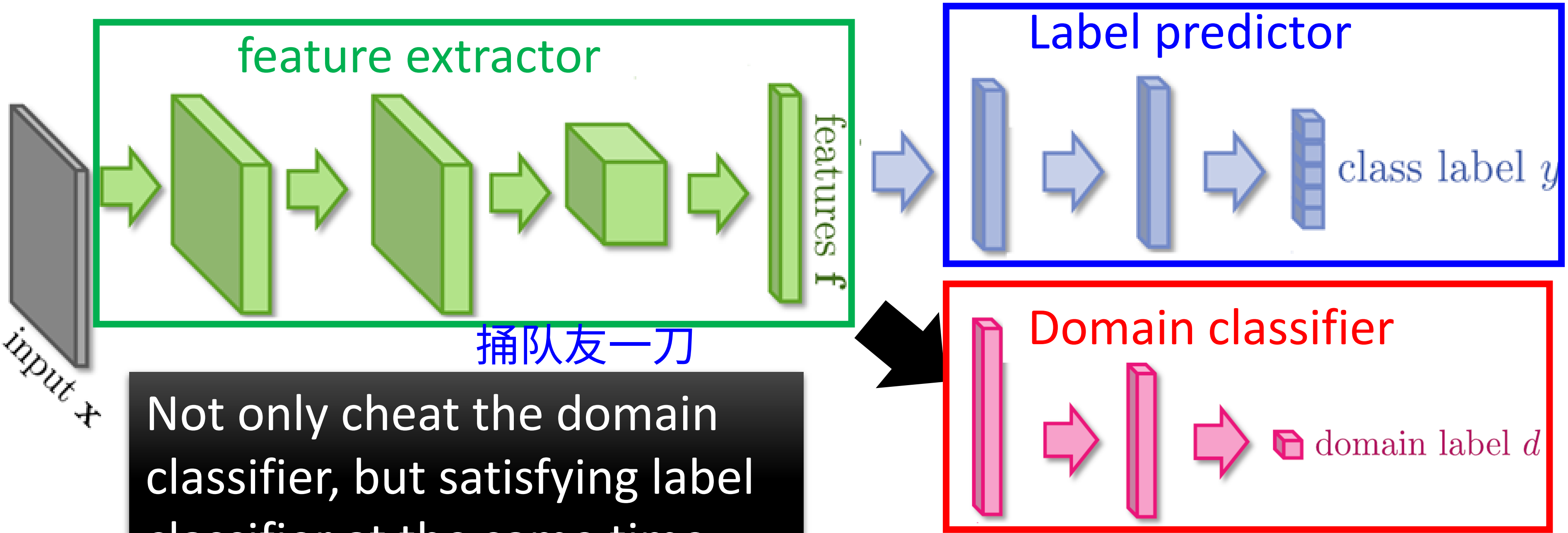


如果都是0，就可以实现domain classifier的要求，但是只有doamain classifier是不够的，要有label predictor

Domain-adversarial training

Maximize label classification accuracy + minimize domain classification accuracy

Maximize label classification accuracy



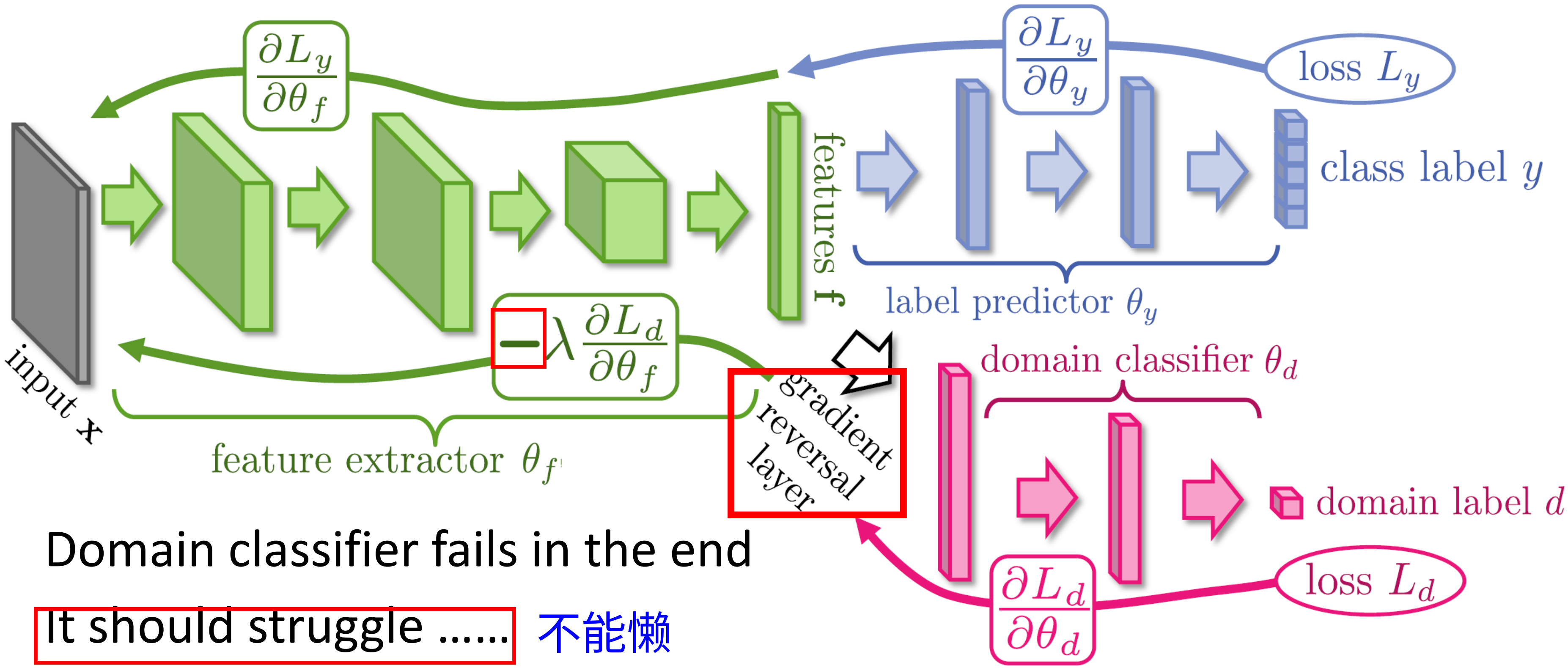
捅队友一刀

Not only cheat the domain classifier, but satisfying label classifier at the same time

Maximize domain classification accuracy

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

Domain-adversarial training



Domain classifier fails in the end

It should struggle …… 不能懶

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 / 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

填充了gap

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

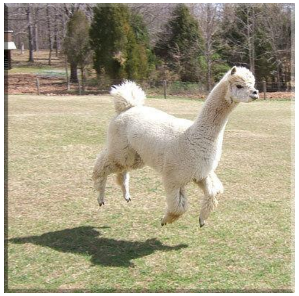
http://evchk.wikia.com/wiki/%E8%8D%89%E6%B3%A5%E9%A6%AC

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

Different tasks

x^s:   ……

y^s: cat dog ……

x^t :  

强机所难

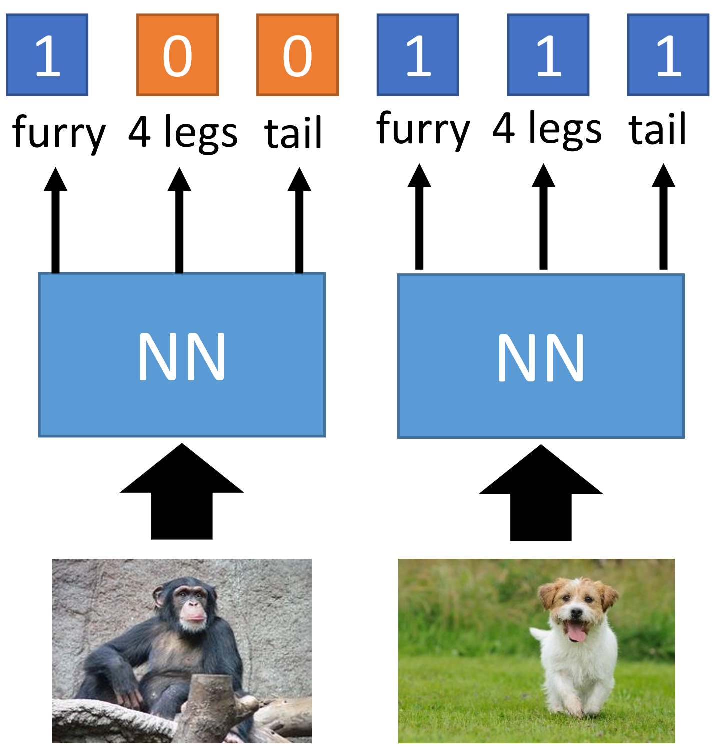
In speech recognition, we can not have all possible words in the source (training) data.

How we solve this problem in speech recognition?
对于训练集没有的词，预测的不是word，而是phoneme，然后再加lexicon去查字典

Zero-shot Learning

- Representing each class by its attributes

Training



Database 丰富且唯一确定

attributes

class

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

chimp没有尾巴

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

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

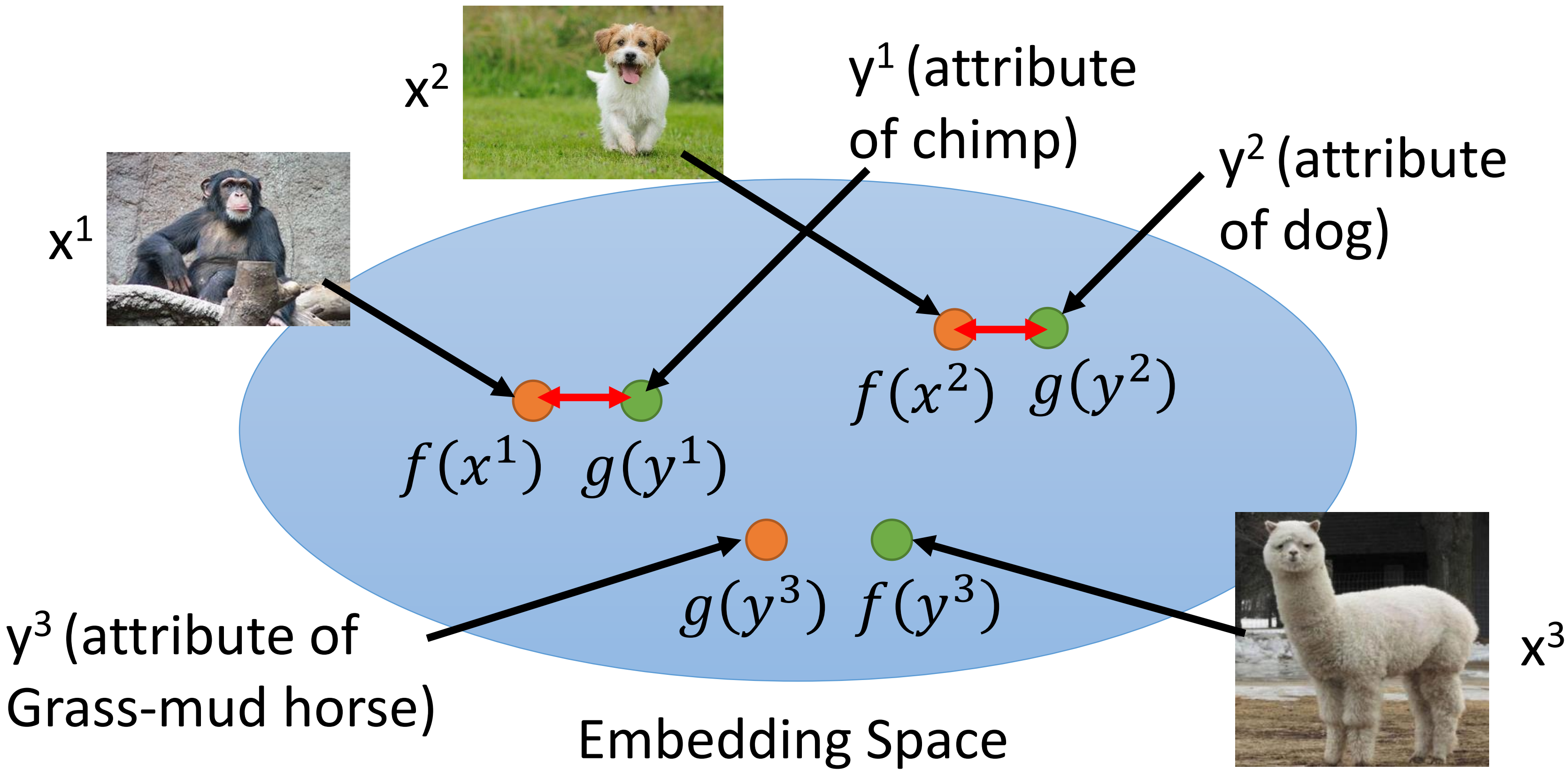
Zero-shot Learning

$f(*)$ and g$(*)$ can be $\boxed{\text{NN.}}$

Training target:

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

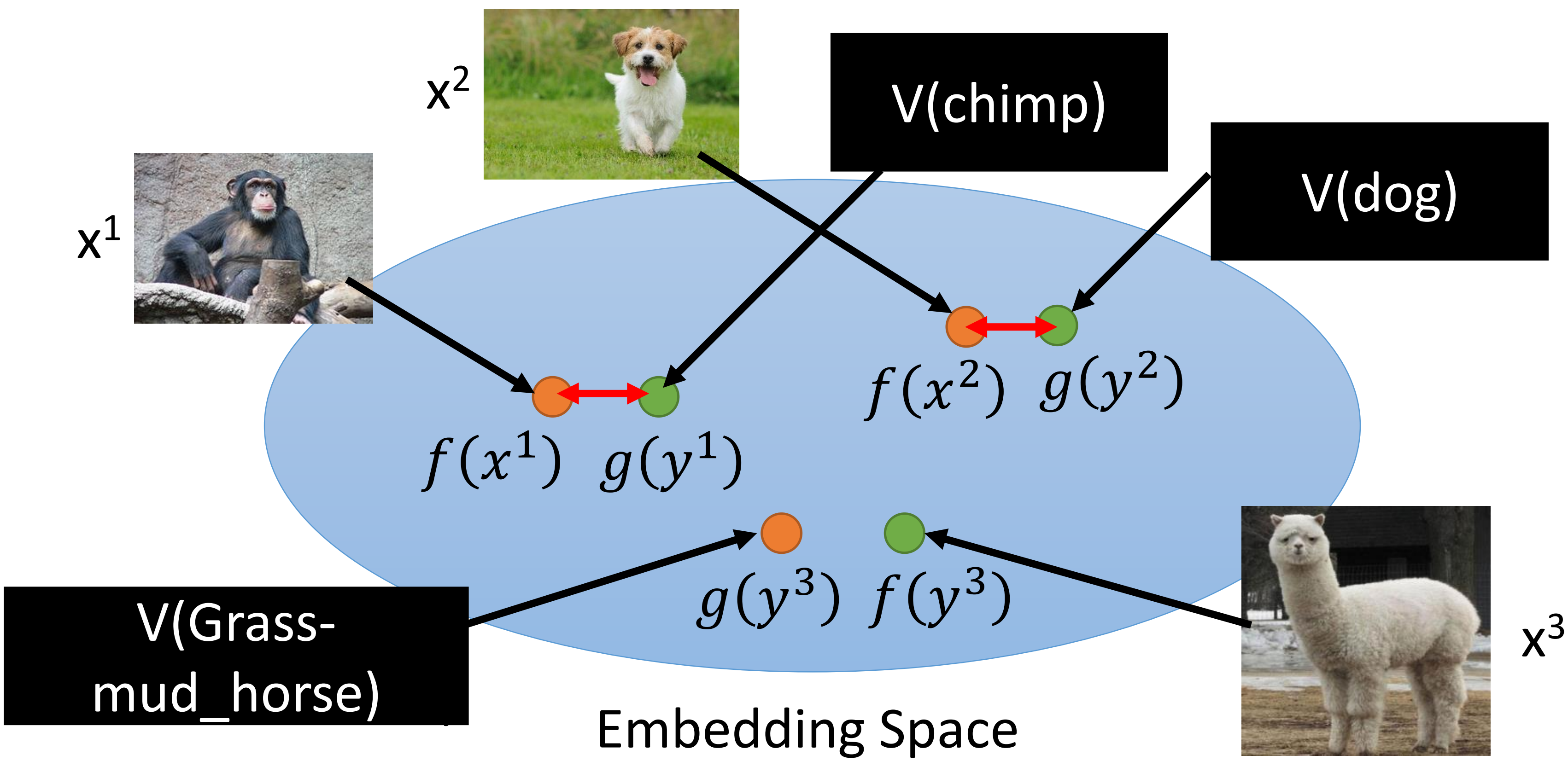
- Attribute embedding



Zero-shot Learning

What if we don't have database

- Attribute embedding + word embedding



Zero-shot Learning

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

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

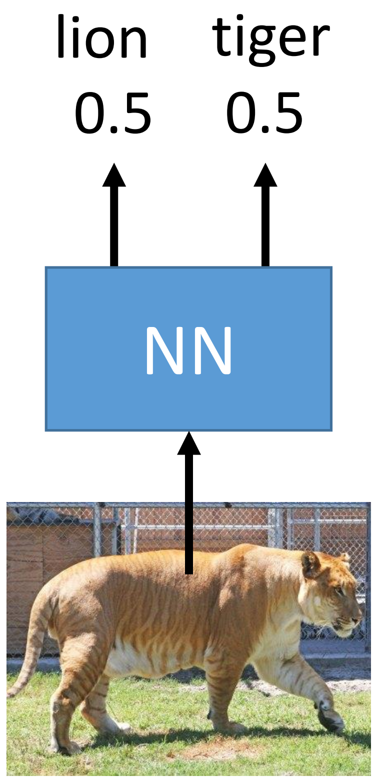
$$\underline{f(x^n) \cdot g(y^n)} - \underline{\max_{m \neq n} f(x^n) \cdot g(y^m)} > k$$

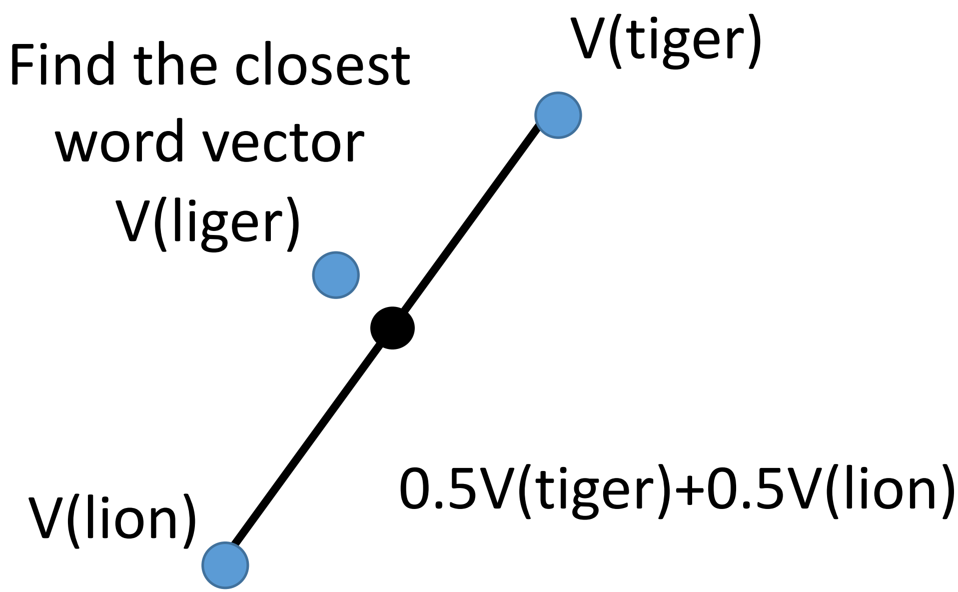
$f(x^n)$ and $g(y^n)$ as close

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

Zero-shot Learning

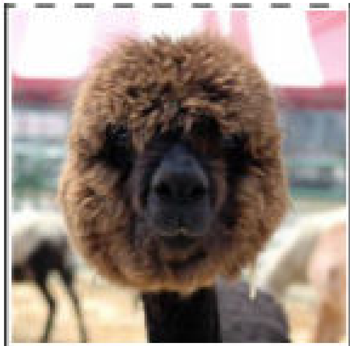
- Convex Combination of Semantic Embedding





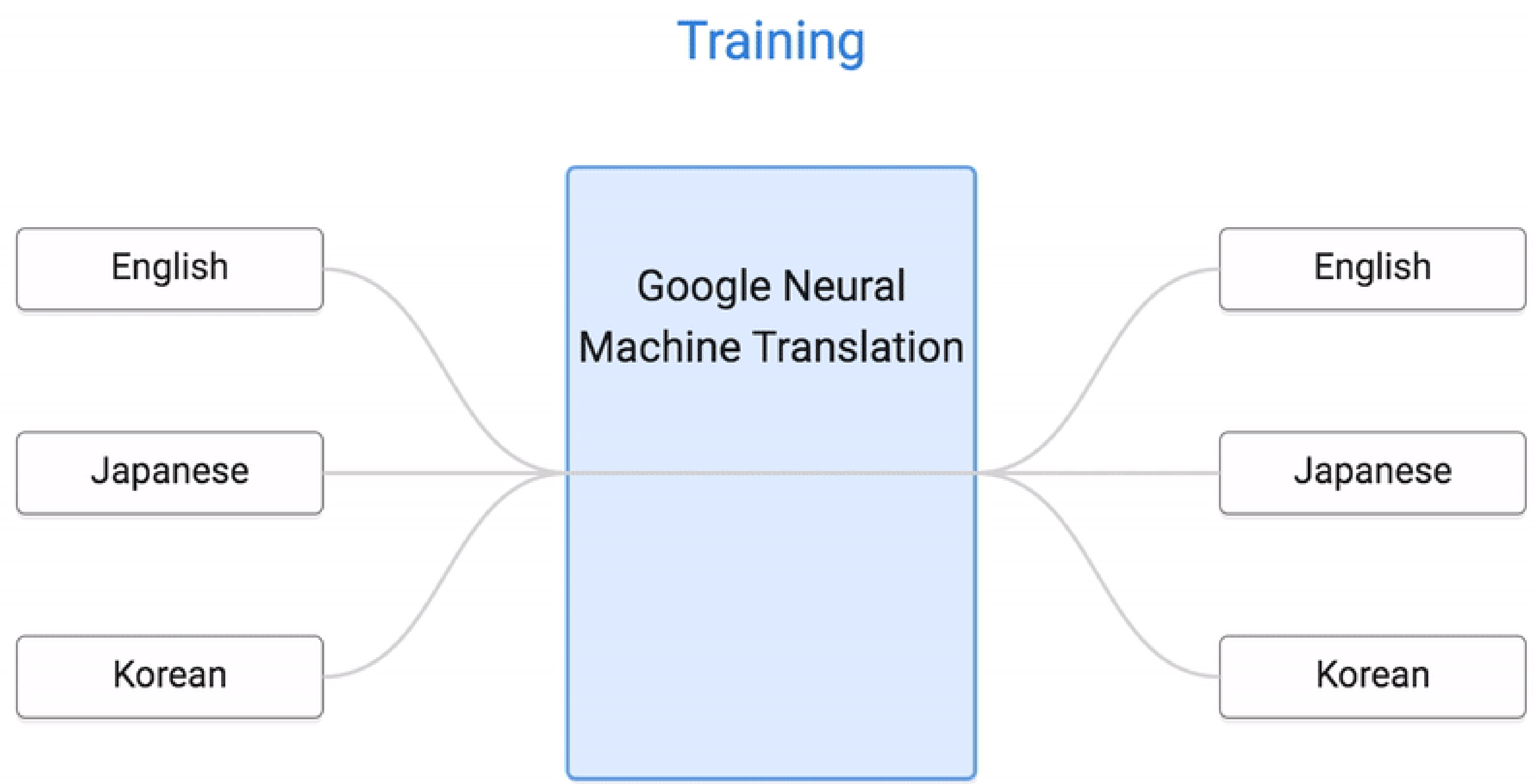
Only need off-the-shelf NN for ImageNet and word vector

不用嵌入空间

Test Image	ConvNet	DeViSE	ConSE(10)
	sea lion plane, carpenter's plane cowboy boot loggerhead, loggerhead turtle goose	elephant turtle turtleneck, turtle, polo-neck flip-flop, thong handcart, pushcart, cart, go-cart	California sea lion **Steller sea lion** Australian sea lion South American sea lion eared seal
 **(alpaca, Lama pacos)**	Tibetan mastiff titi, titi monkey koala, koala bear, kangaroo bear llama chow, chow chow	kernel littoral, litoral, littoral zone, sands carillon Cabernet, Cabernet Sauvignon poodle, poodle dog	dog, domestic dog domestic cat, house cat schnauzer Belgian sheepdog domestic llama, Lama peruana

https://arxiv.org/pdf/1312.5650v3.pdf

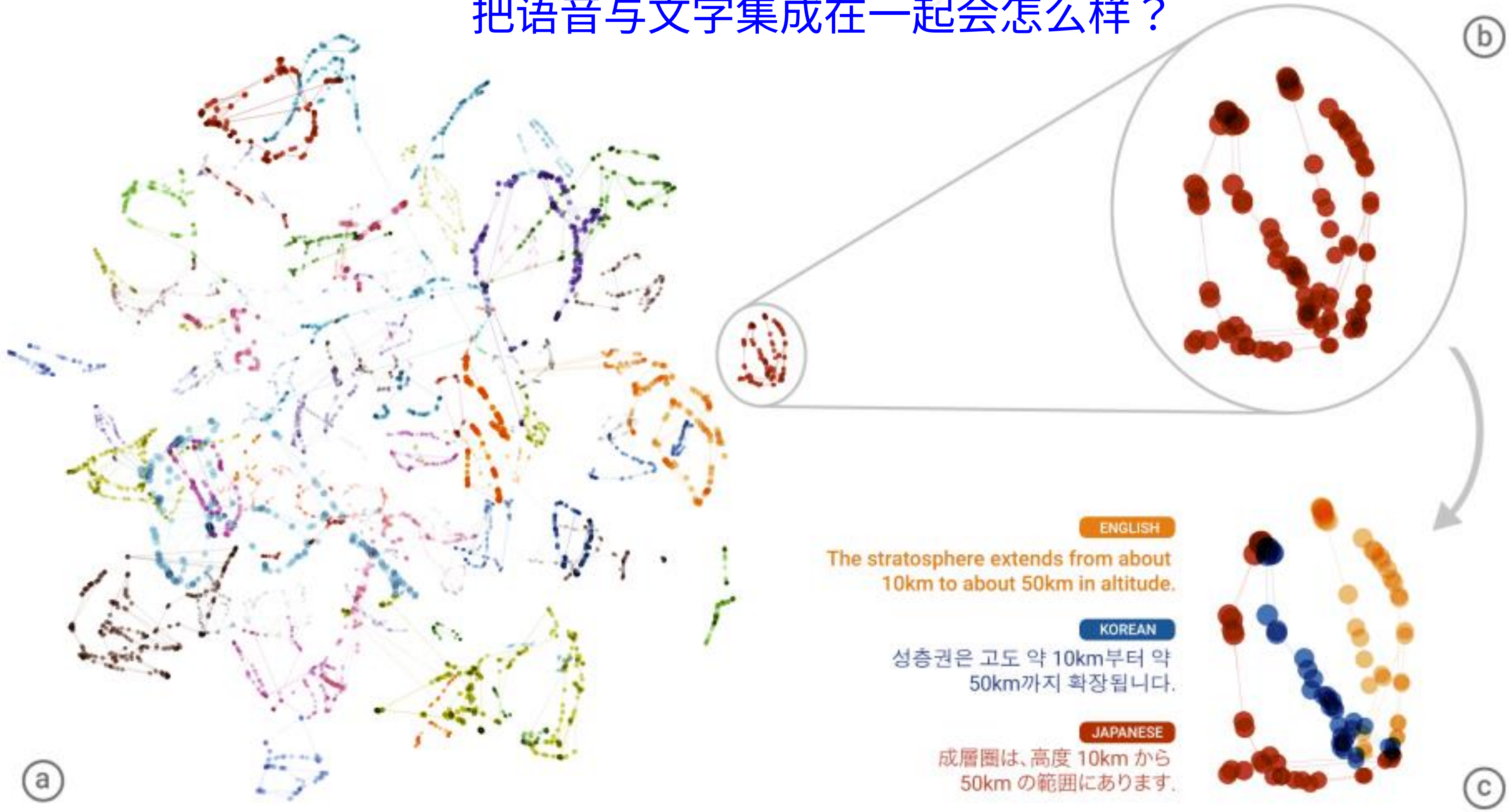
Example of Zero-shot Learning



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

Example of Zero-shot Learning

另外一种语言space
把语音与文字集成在一起会怎么样？



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	Source Data (not directly related to the task): unlabeled 半监督
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
Target Data	unlabeled	Domain-adversarial training Zero-shot learning	Self-taught Clustering Wenyuan Dai, Qiang Yang, Gui-Rong Xue, Yong Yu, "Self-taught clustering", ICML 2008

Self-taught learning

原来是sparse coding 现在多是auto-encoder

- 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

Acknowledgement

- 感謝 劉致廷 同學於上課時發現投影片 上的錯誤