Transfer Learning

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

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

Transfer Learning

原本在做貓跟狗的分類器，需要很多label data做監督式學習

Dog/Cat Classifier

cat

dog

搜集的資料不一定完全與我們想要的task有關

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

elephant

tiger

dog

cat

Similar domain, different tasks

Different domains, same task

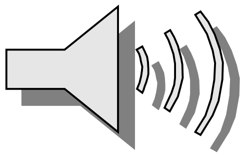
Why?

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

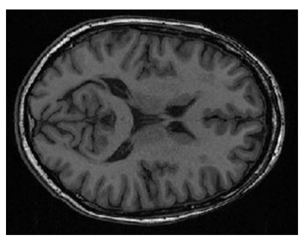


Taiwanese

YouTube: English, Chinese ……

辨識腦瘤

Image Recognition



Medical Images

一堆圖片

Text Analysis

Specific domain

分類文件

Webpages

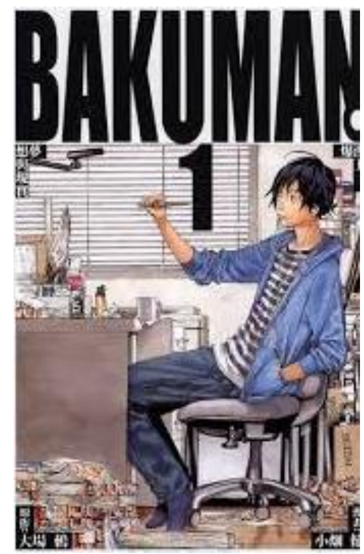
網路上的文章

Transfer Learning

- Example in real life

研究生		*漫畫家*
研究生	←	漫畫家
跑實驗	←	畫分鏡
指導教授	←	責編
投稿期刊	←	投稿 jump

(word embedding knows that)



爆漫王

Transfer Learning - Overview

有一些關聯性的資料

Source Data (not directly related to the task)

		labelled	unlabeled
直接相關的資料 Target Data	labelled	Model Fine-tuning	
	unlabeled		

Warning: different terminology in different literature

這些資料都可以是label以及unlabel，以下先介紹都有label的情況

舉例：語音辨識

Model Fine-tuning

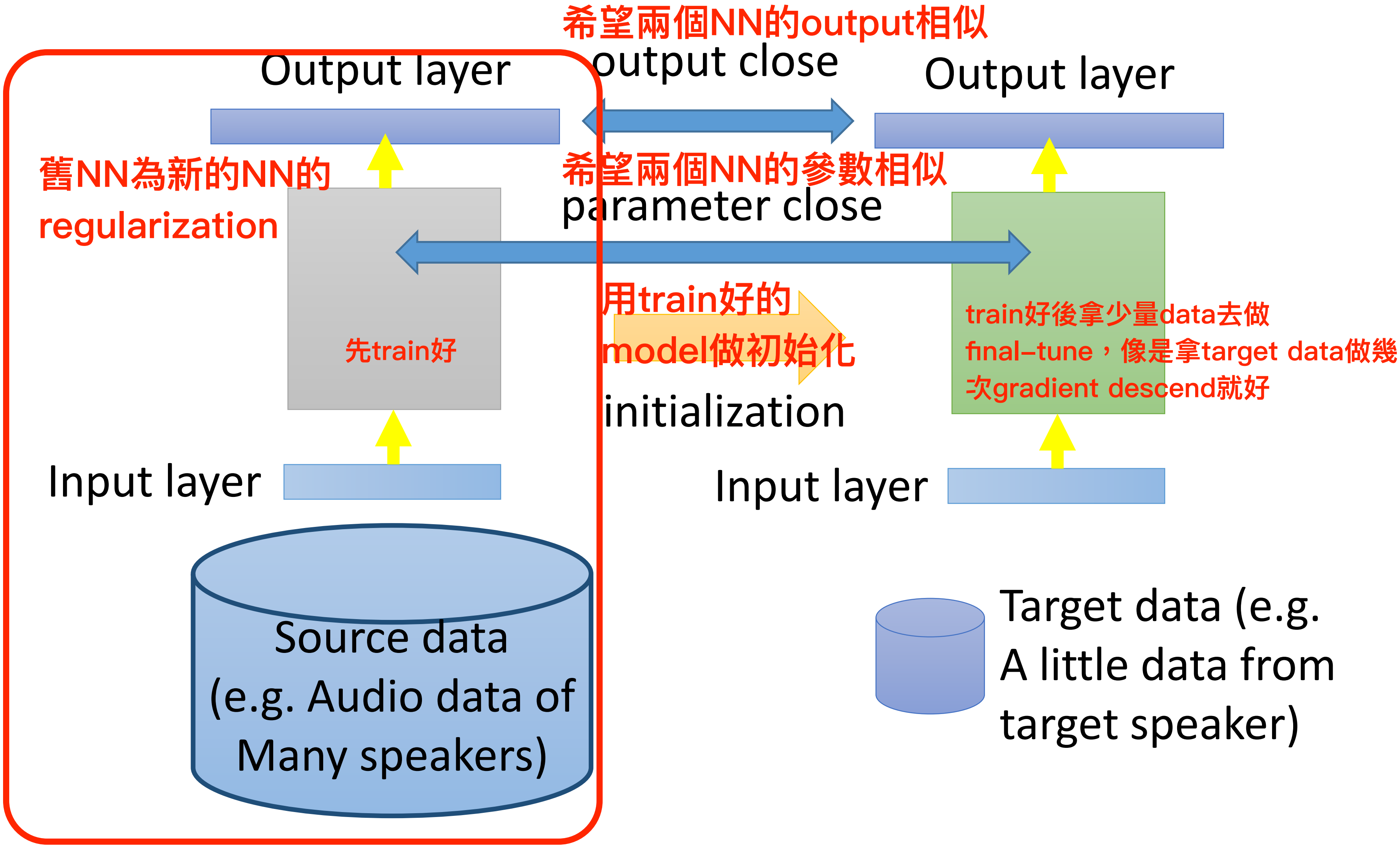
data量少到不行剩下個位數個資料

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

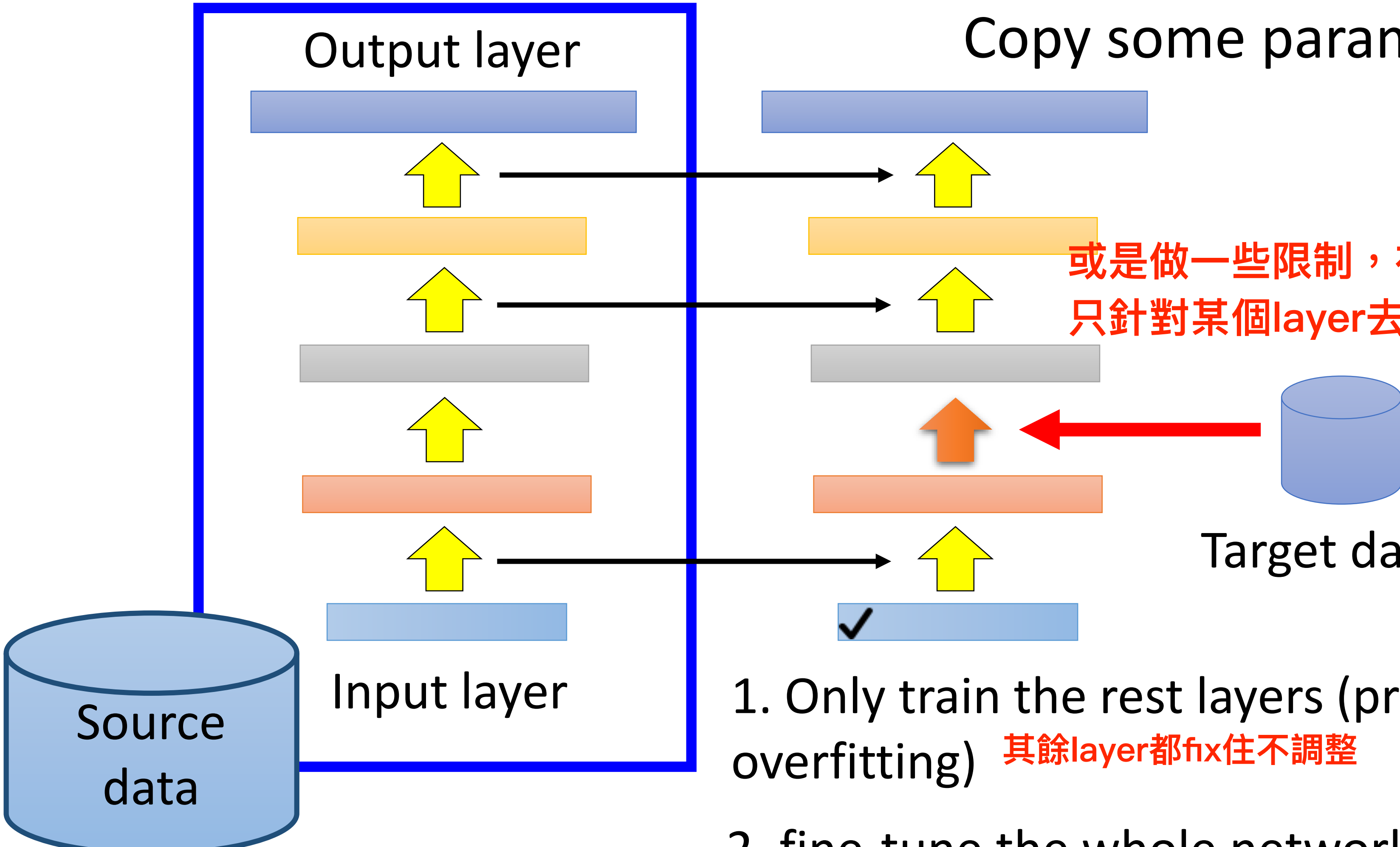
這邊通通是label過的

- 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 所有人的聲音
  - Target data: audio data and its transcriptions of specific user 想辨識的人的聲音
- 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) 其餘layer都fix住不調整

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

Layer Transfer

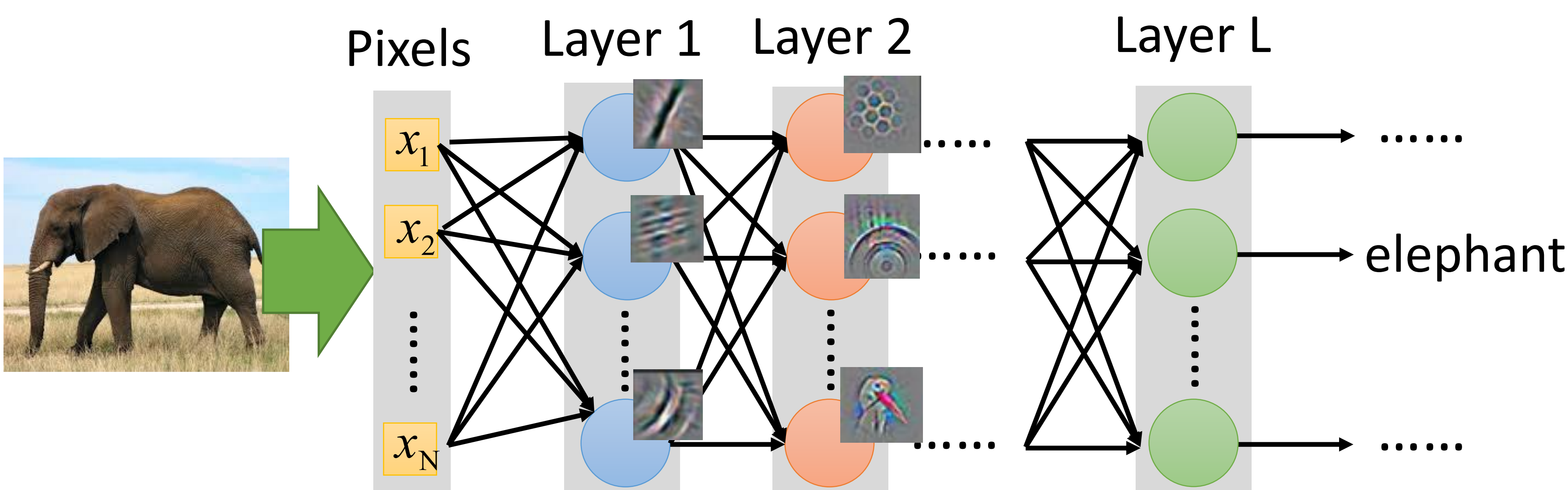
語音上：訓練第一層layer，因為希望做一個線性轉換把不同性別的聲音都變成中性的

影像上：訓練倒最後幾層layer，前幾層只是在偵測幾何圖案，可以共用

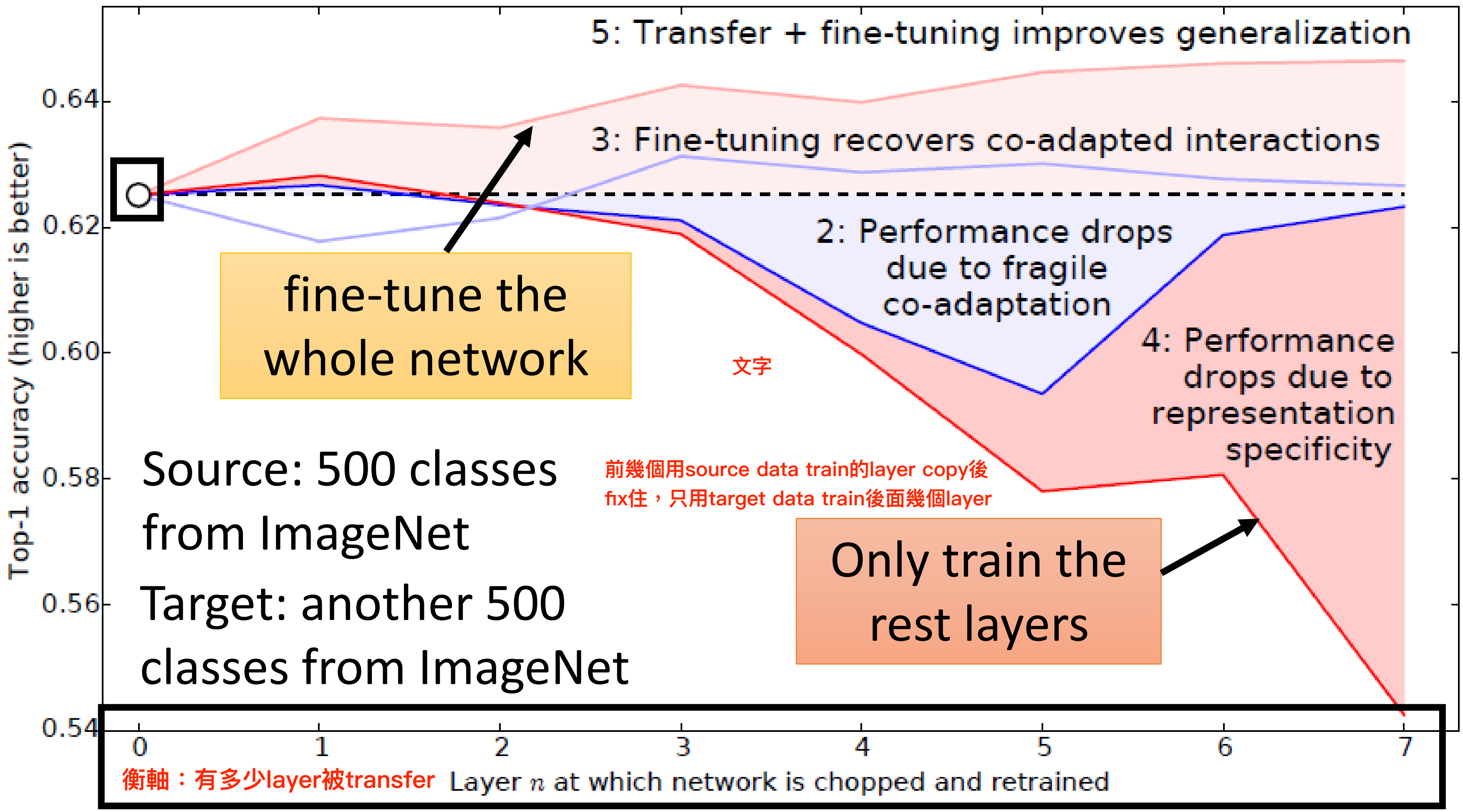
- Which layer can be transferred (copied)?
  - Speech: usually copy the last few layers

    train接近input的layer
  - Image: usually copy the first few layers

    train接近output的layer



Layer Transfer - Image

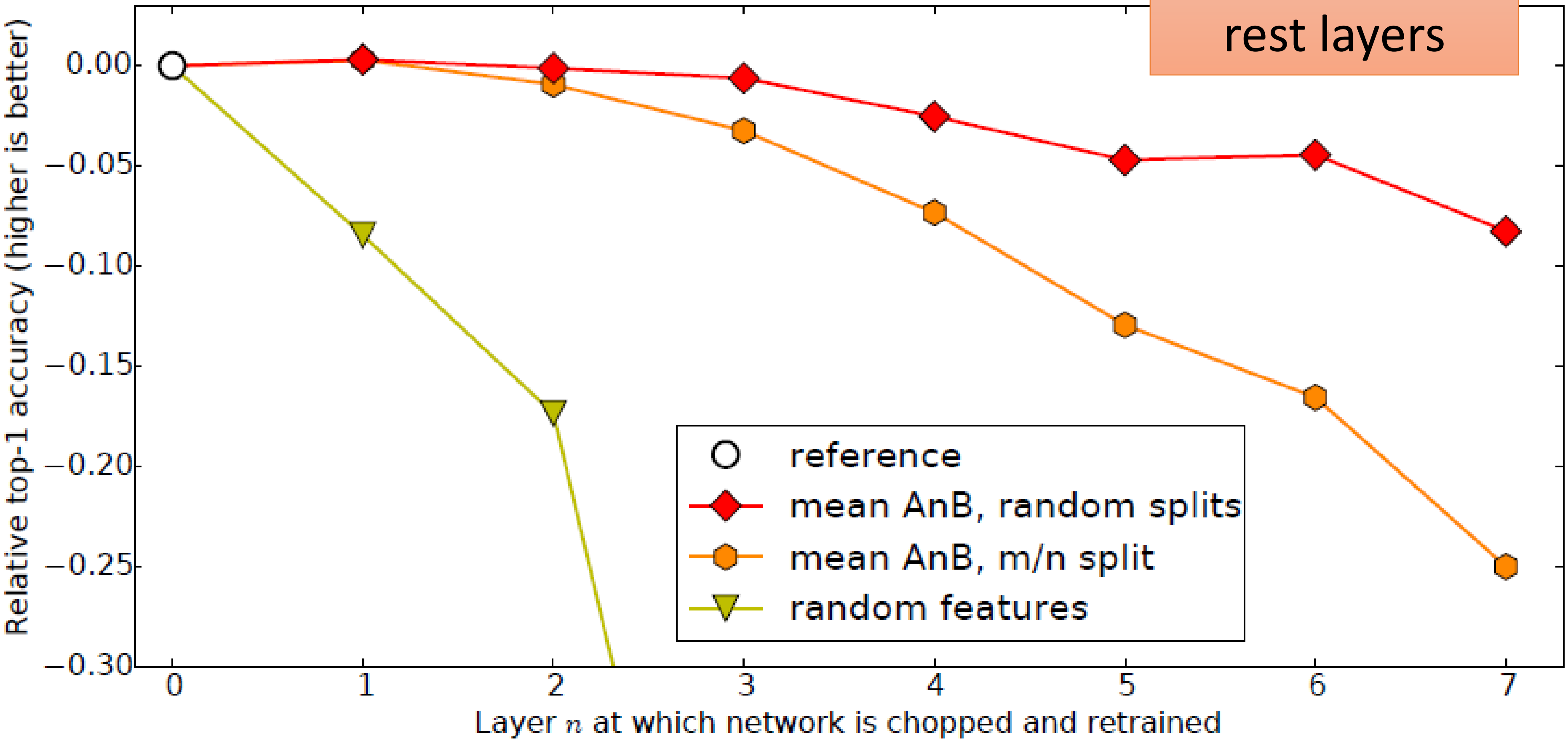


Jason Yosinski, Jeff Clune, Yoshua Bengio, Hod Lipson, “How transferable are features in deep neural networks?”, NIPS, 2014

略過

Layer Transfer - Image

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

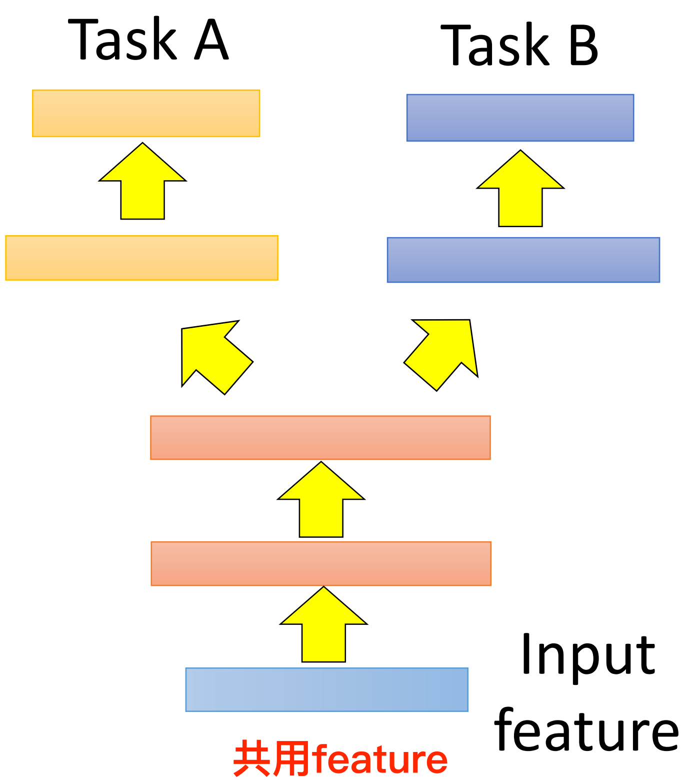
		Source Data (not directly related to the task)	
		labelled	unlabeled
Target Data	labelled	Fine-tuning Multitask Learning	
	unlabeled		

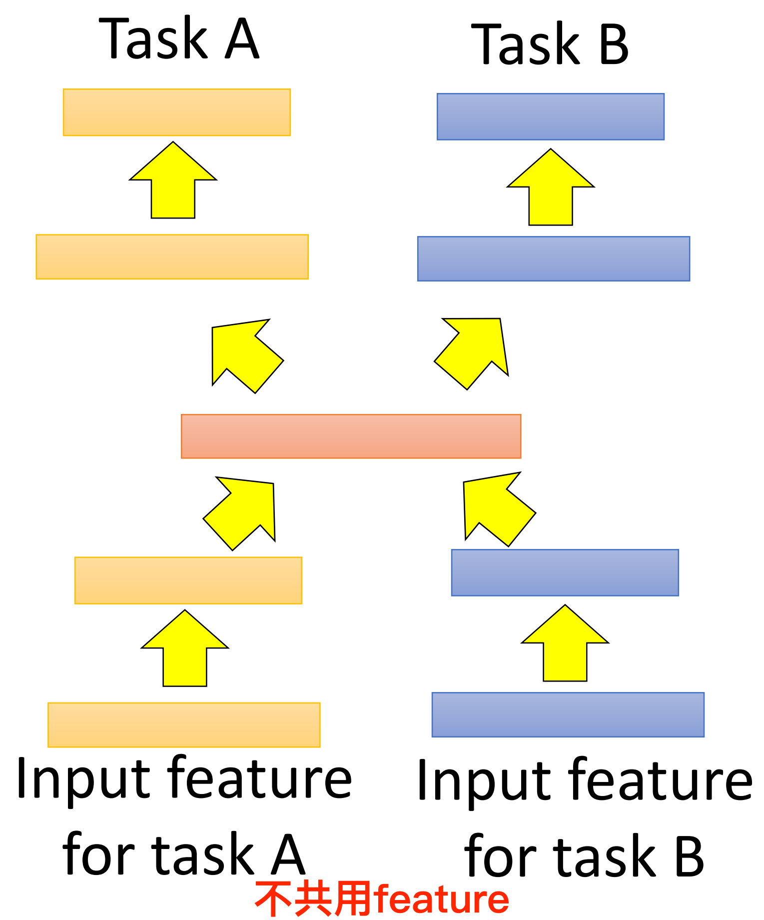
ＥＸ：同時訓練他打籃球棒球

Multitask Learning

若taskA, taskB完全無關則當然沒用，若有適當的相關的話則會有幫助

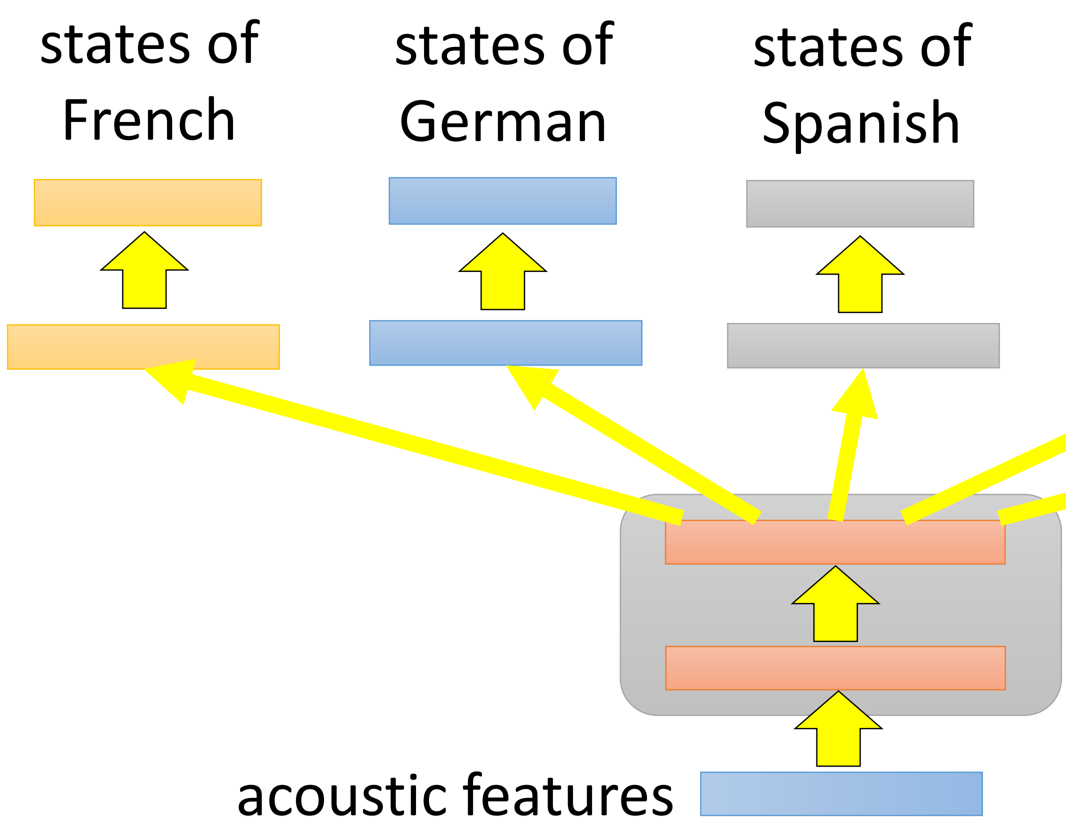
- The multi-layer structure makes NN suitable for multitask learning





辨識多國語言

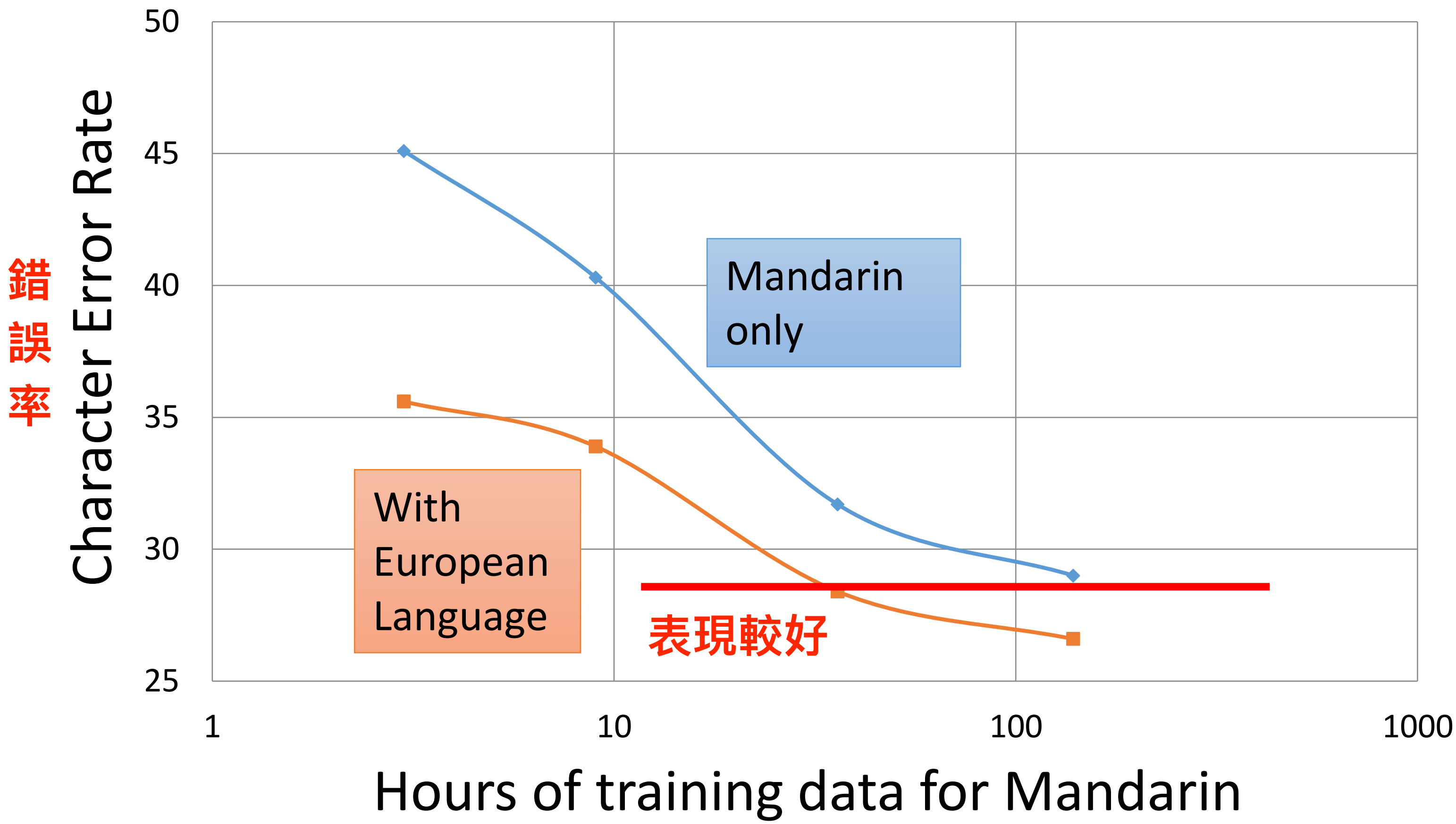
Multitask Learning
- Multilingual Speech Recognition



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

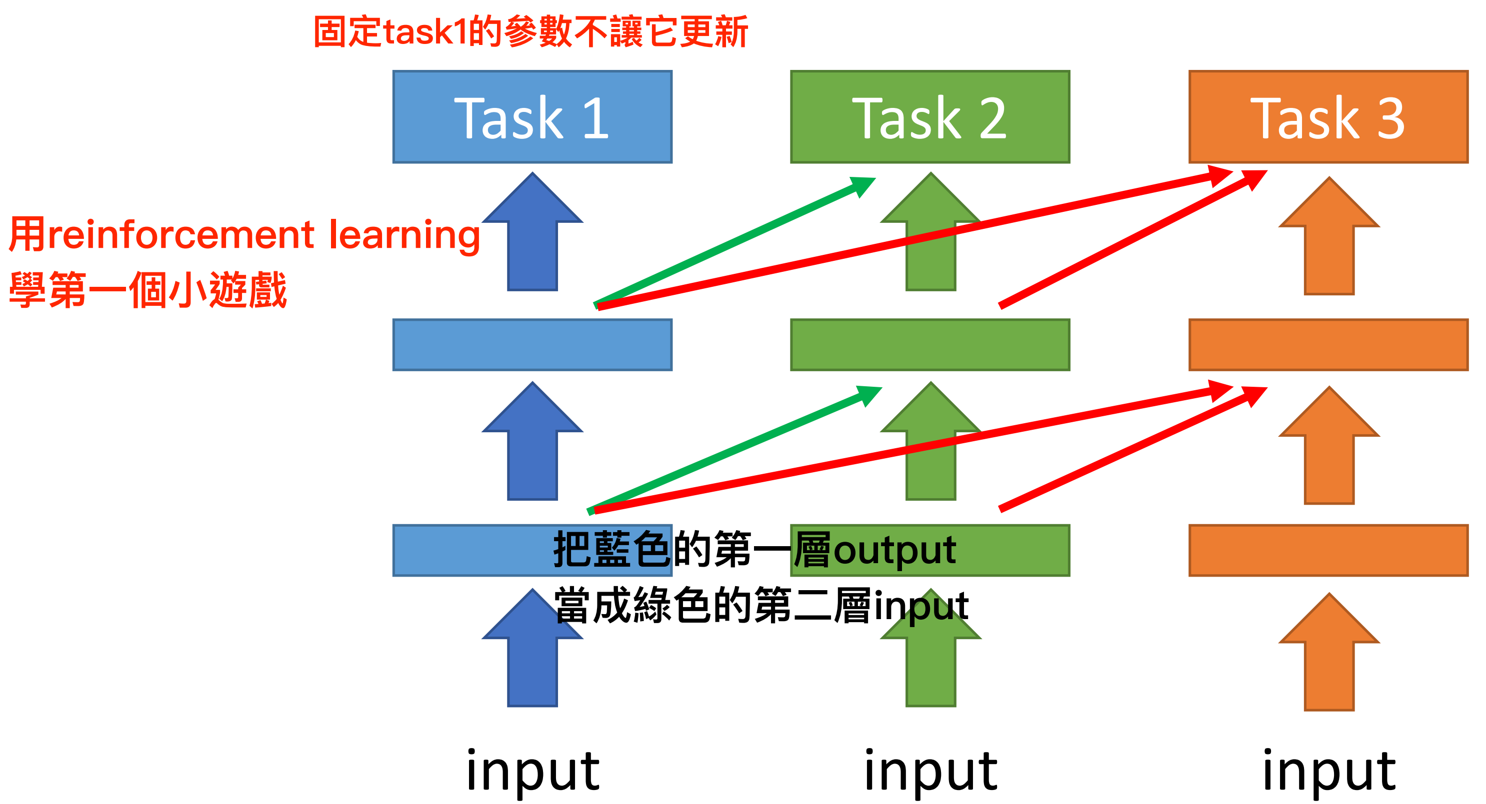
Multitask Learning - Multilingual



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

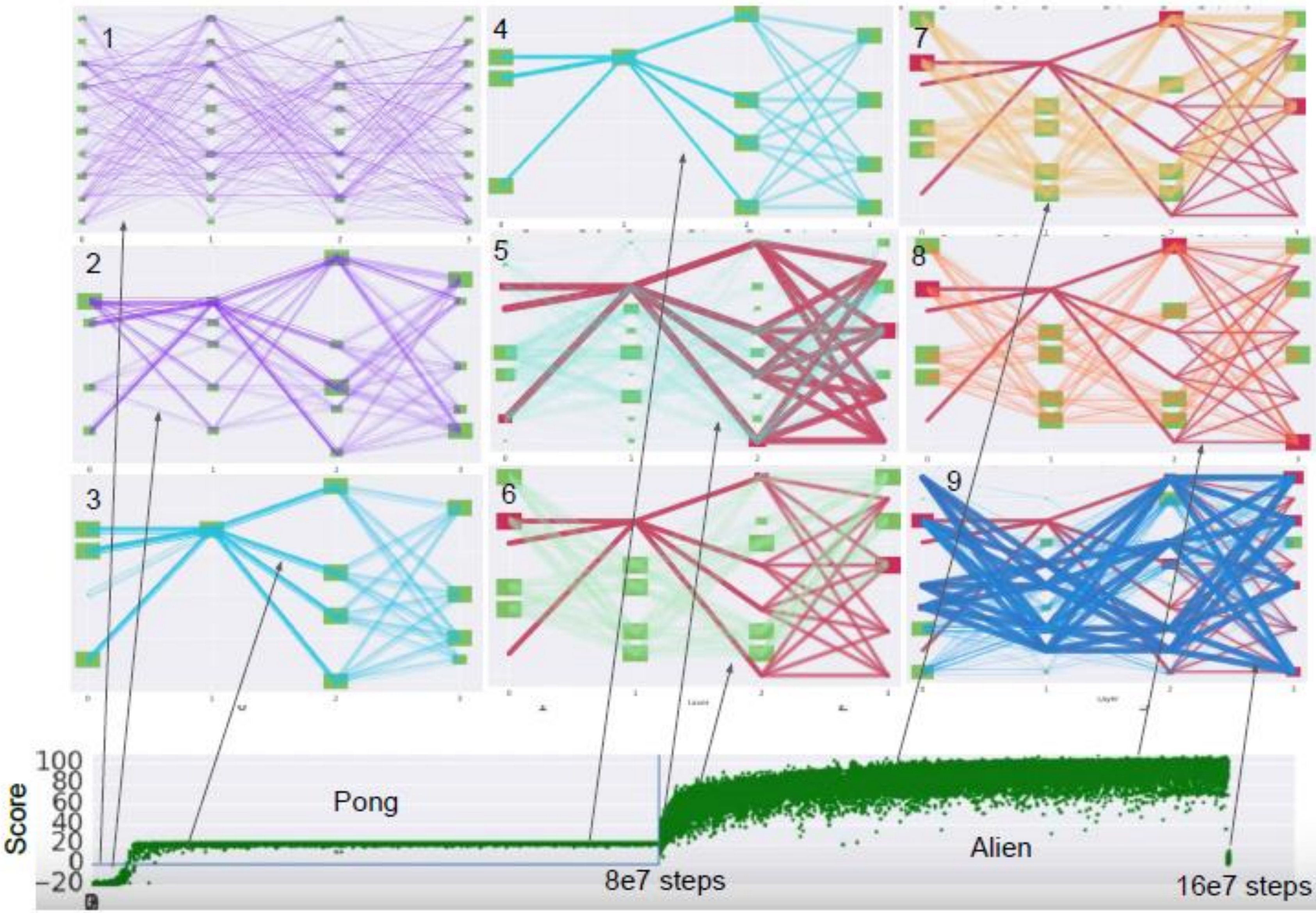
希望機器先學task1, 在學task2...

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

DeepMind



Chrisantha Fernando, Dylan Banarse, Charles Blundell, Yori Zwols, David Ha, Andrei A. Rusu, Alexander Pritzel, Daan Wierstra, “PathNet: Evolution Channels Gradient Descent in Super Neural Networks”, arXiv preprint, 2017

使用EM去train，先找比較大的network，每個task只能用一部份的參數學好後並fixed住

Transfer Learning - Overview

		Source Data (not directly related to the task)	
		labelled	unlabeled
Target Data	labelled	Fine-tuning Multitask Learning	
	unlabeled	Domain-adversarial training	接下來假設source data有label，target data是沒有label的

Task description

同一個task：辨識數字

資料分布很不一樣，直接train效果會不好

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

Same task, mismatch

MNIST

SOURCE　with label

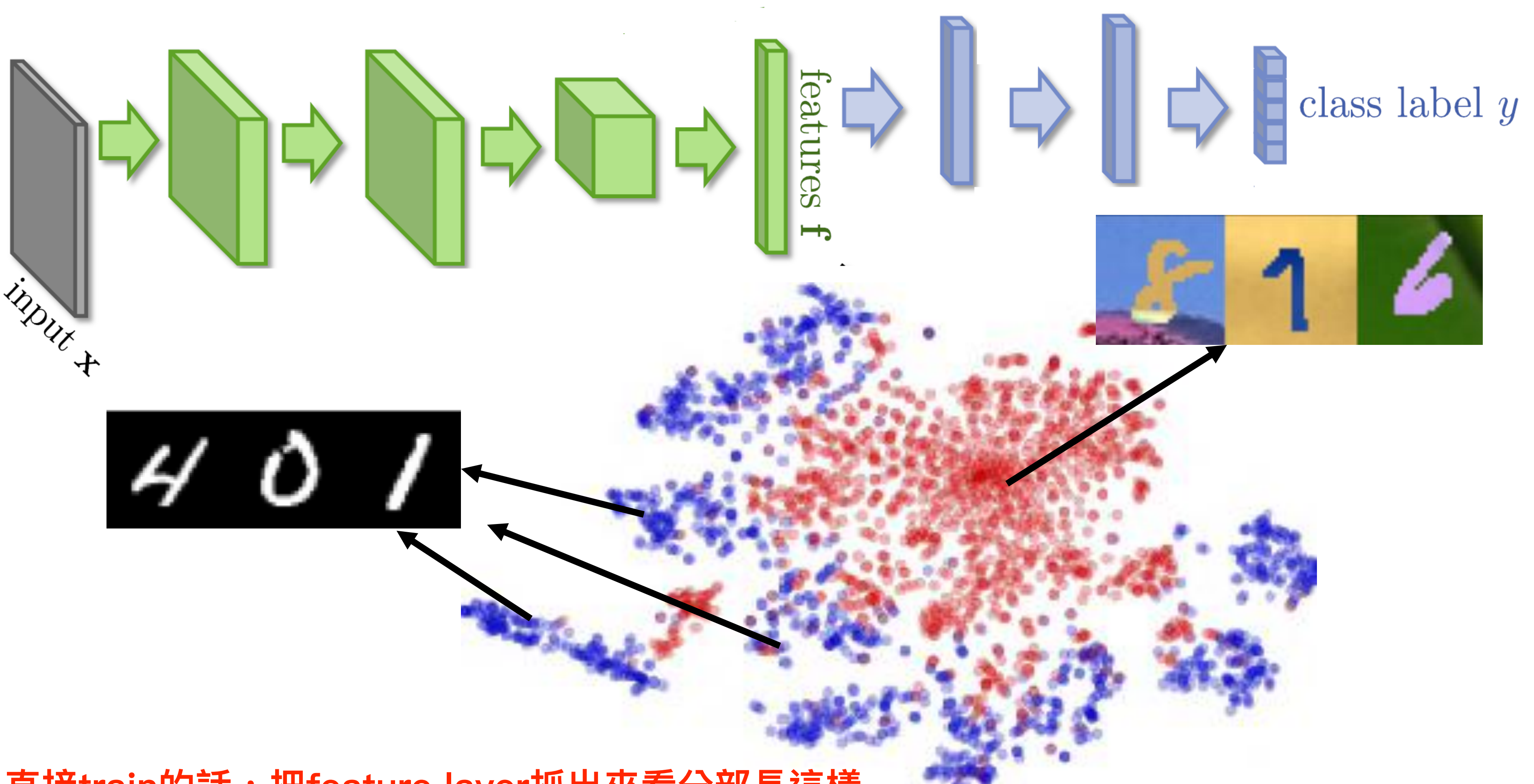
TARGET　without label

MNIST-M

想辨識的corpus

ＧＡＮ的其中一種，想把source data的domain轉到train data

Domain-adversarial training



直接train的話，把feature layer抓出來看分部長這樣

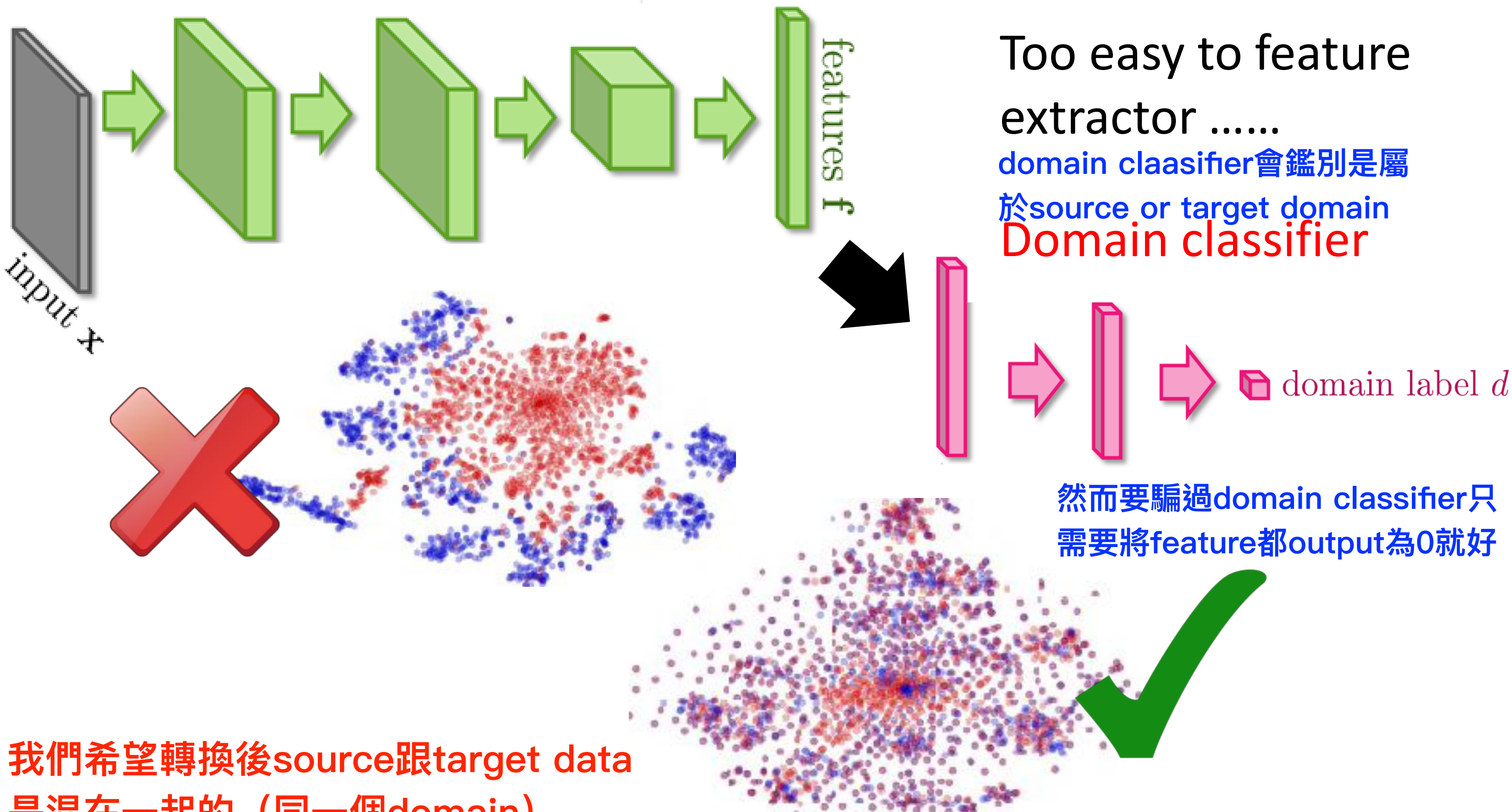
Domain-adversarial training

類似descriminator概念

Similar to GAN

Too easy to feature extractor ……

domain claasifier會鑑別是屬於source or target domain



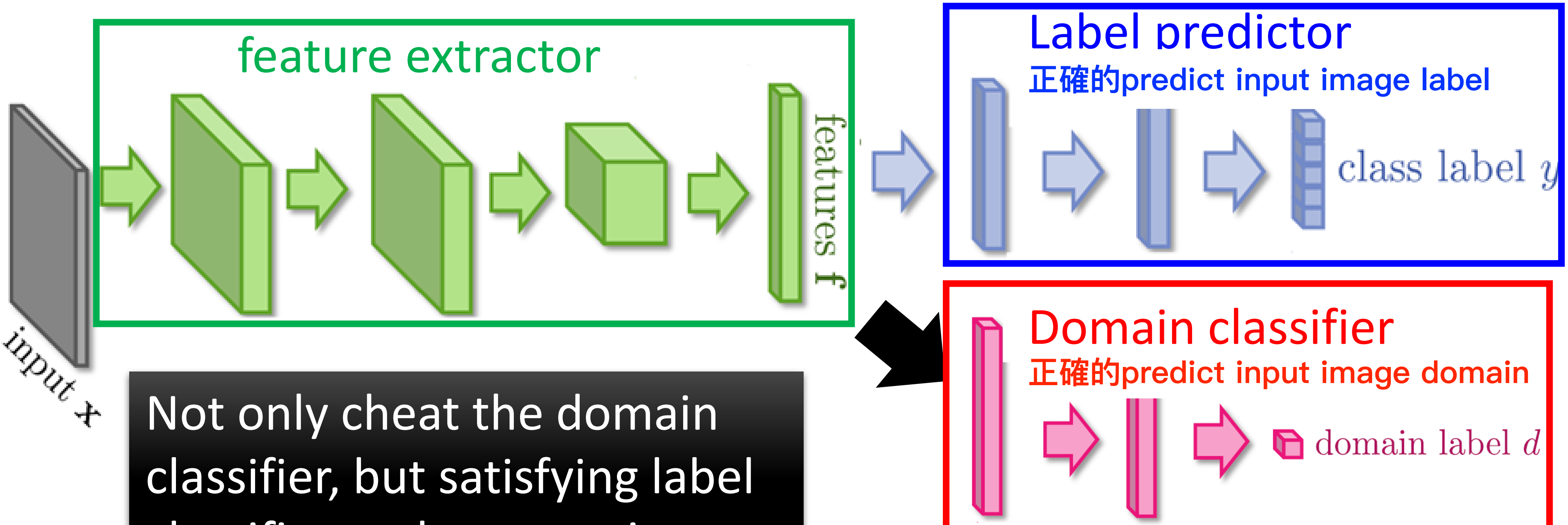
然而要騙過domain classifier只需要將feature都output為0就好

我們希望轉換後source跟target data是混在一起的（同一個domain）

Domain-adversarial training

Maximize label classification accuracy + minimize domain classification accuracy

Maximize label classification accuracy



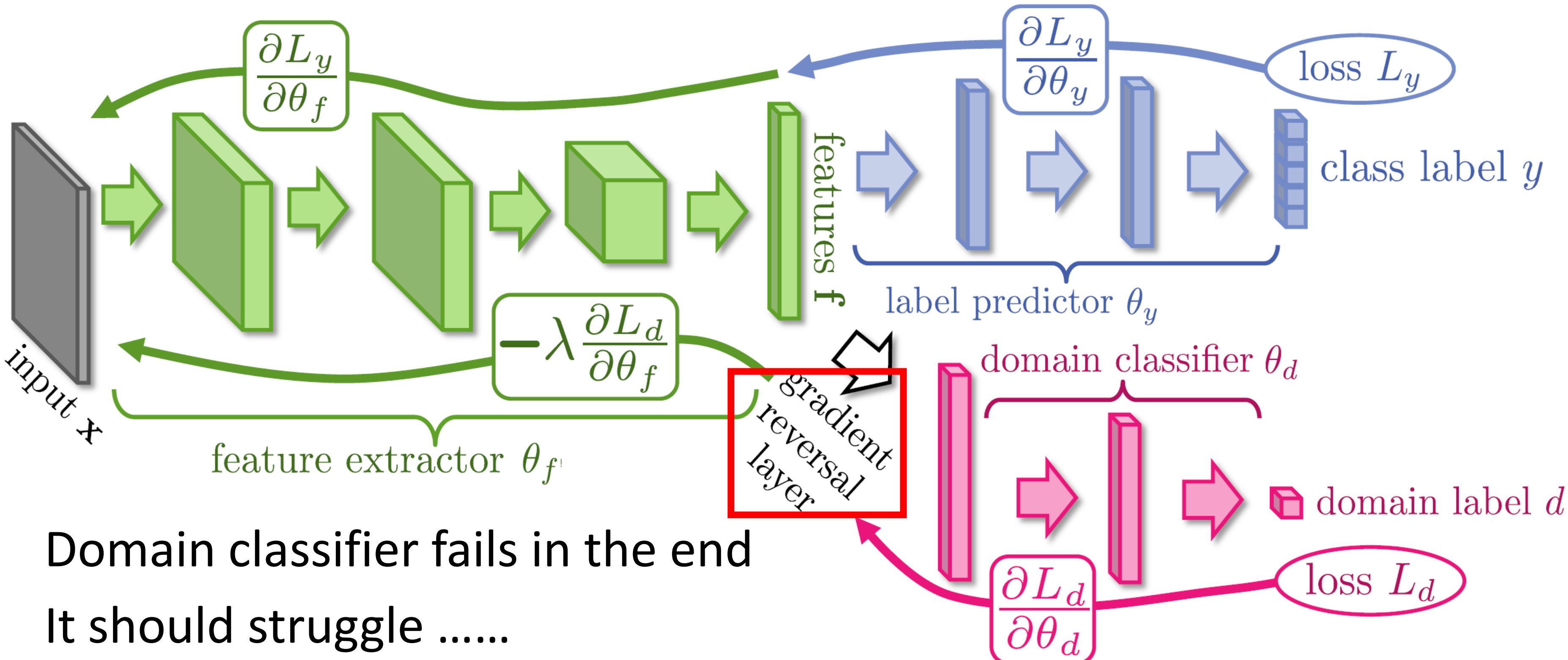
Not only cheat the domain classifier, but satisfying label classifier at the same time

Maximize domain classification accuracy

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

當做完gradient descend後紅色與藍色要做back propagation時，會把紅色的gradient乘上一個負號，藉此搞壞domain classifier而騙過他

Domain-adversarial training



Domain classifier fails in the end

It should struggle ......

而非只是讓domain classifier output 0

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 lower bound	.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 upper bound	.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

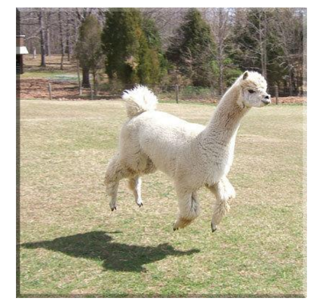
http://evchk.wikia.com/wiki/%E8%8D%89%E6%B3%A5%E9%A6%AC

完全不同的task：便是貓狗vs辨識草泥馬

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

Different tasks

x^s:   ……

x^t :  

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?

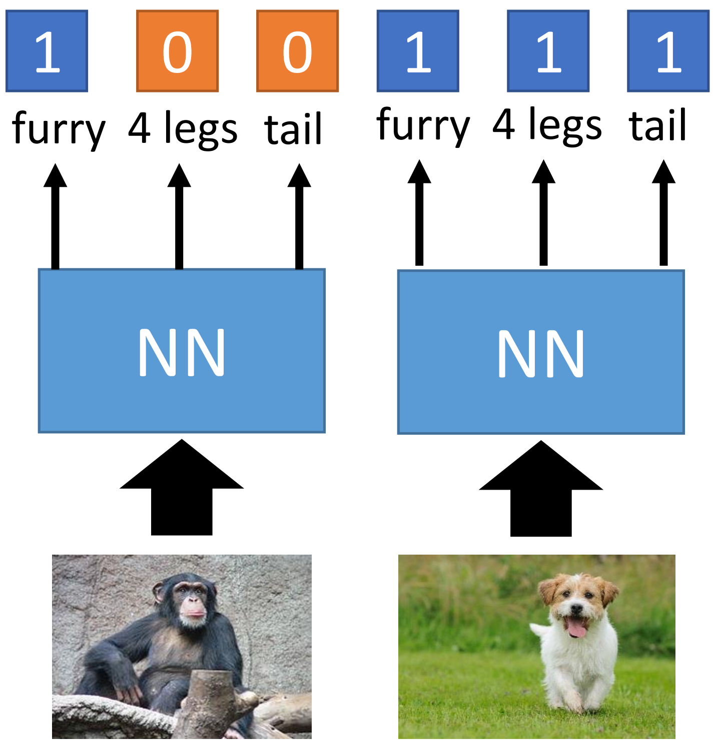
Solve：無法在training data囊括所有的詞彙，因此找出比詞彙更小的單位

Solve：無法在training data囊括所有的動物，因此找出比動物更小的特徵

Zero-shot Learning

- Representing each class by its attributes

Training



文字

Database

attributes

class

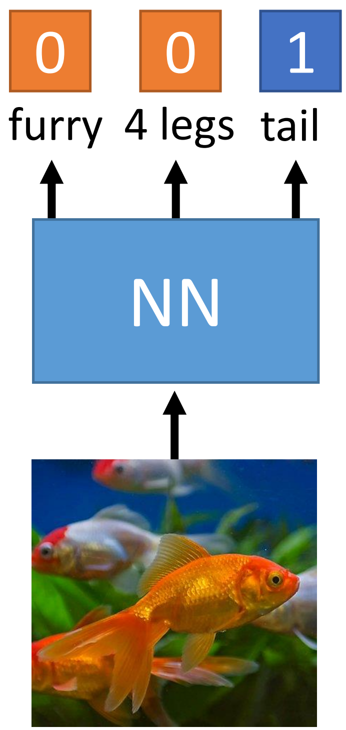
	furry	4 legs	tail	…
Dog	O	O	O	
Fish	X	X	O	
Chimp	O	X	X	
…				

sufficient attributes for one to one mapping

Zero-shot Learning

- Representing each class by its attributes

Testing



Find the class with the most similar attributes

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

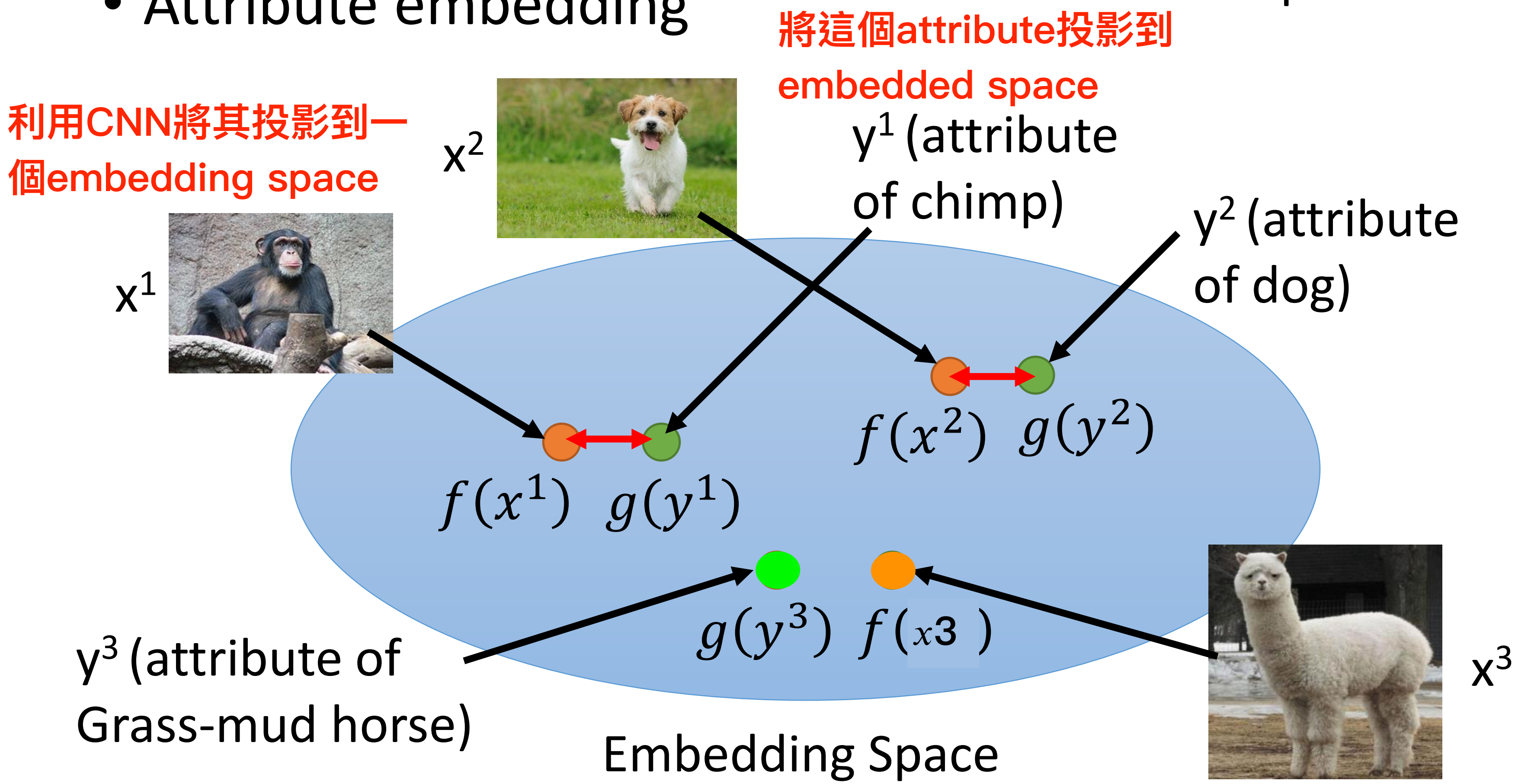
Zero-shot Learning

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

Training target:

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

- Attribute embedding



Zero-shot Learning

f,g都output零就好了

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

Problem?

把相關的距離差得越近越好

把相關的距離差得越近越好

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

把不相關的距離差得越遠越好

Margin you defined

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

若成立則loss value = 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

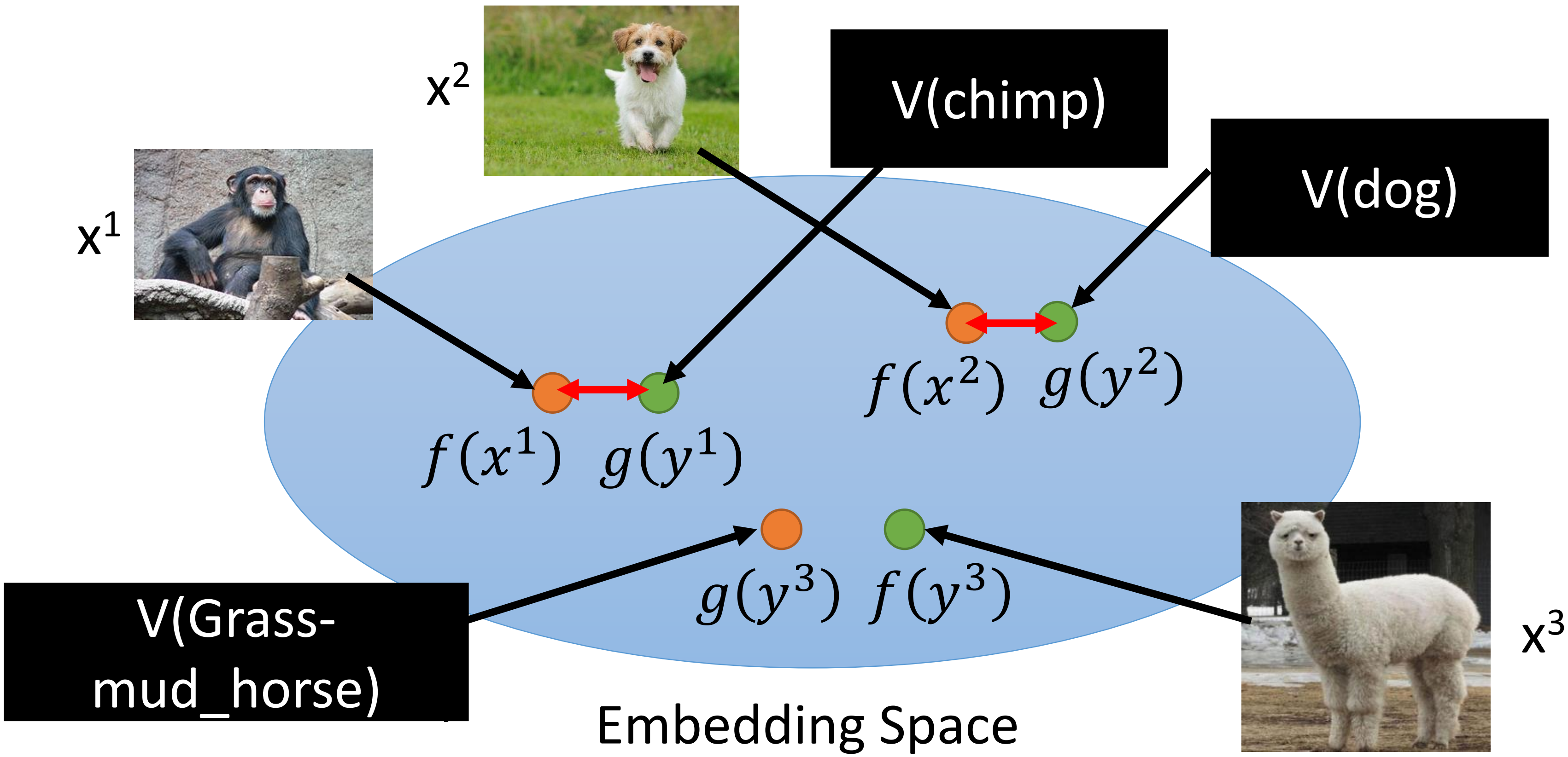
$f(x^n)$ and $g(y^m)$ not as close

Zero-shot Learning

What if we don’t have database

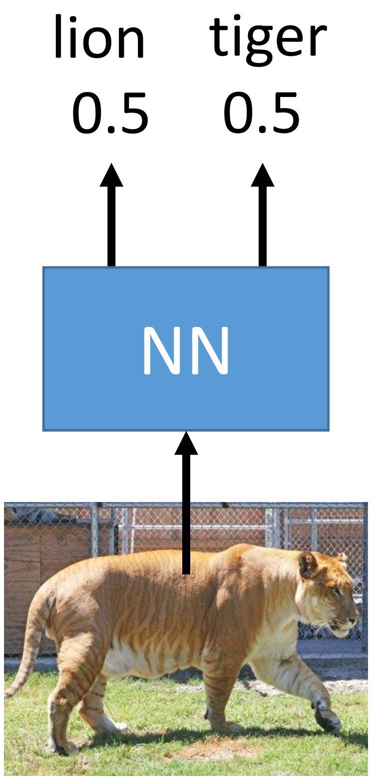
把動物名字利用word embedding找出一個vector

- Attribute embedding + word embedding

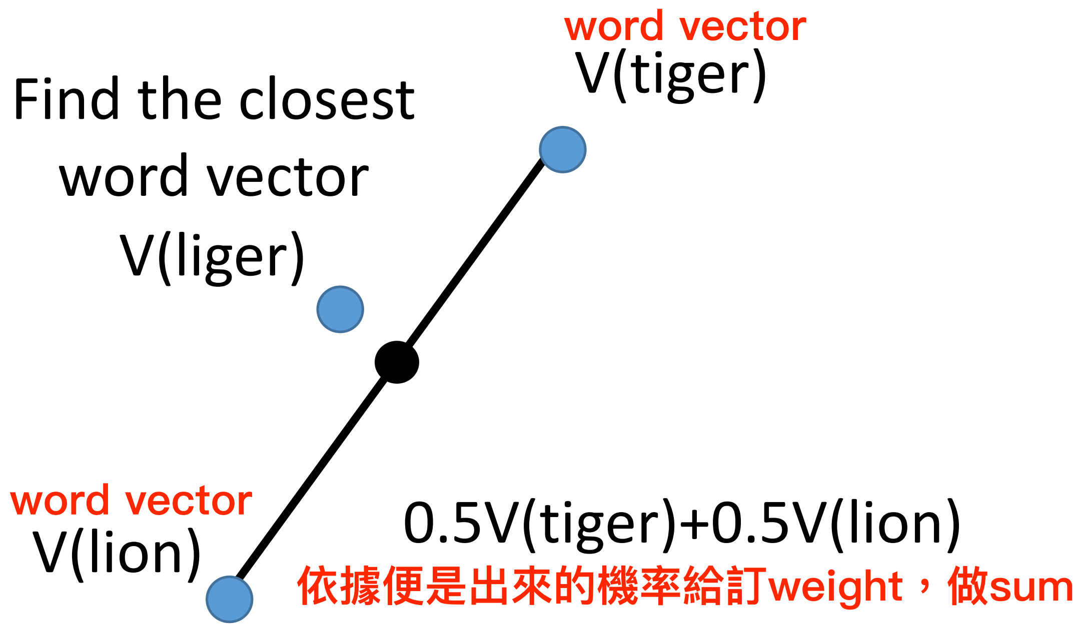


Zero-shot Learning

- Convex Combination of Semantic Embedding



現成的影像辨識model



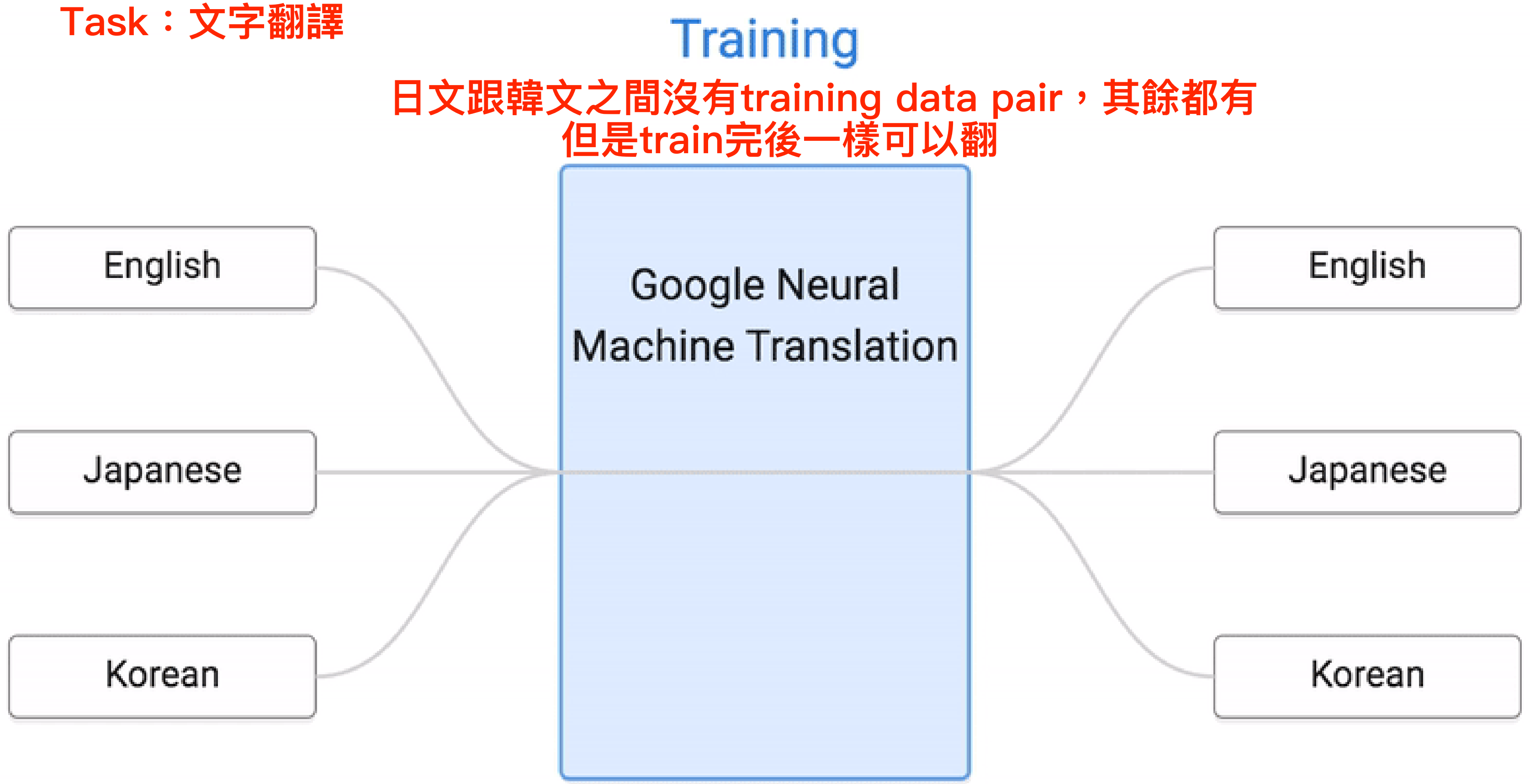
Only need off-the-shelf NN for
ImageNet and word vector

Test Image	ConvNet	DeViSE	ConSE(10)
	CNN	word vector	word vector+weight sum
	CNN	word vector	word vector+weight sum

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

Example of Zero-shot Learning

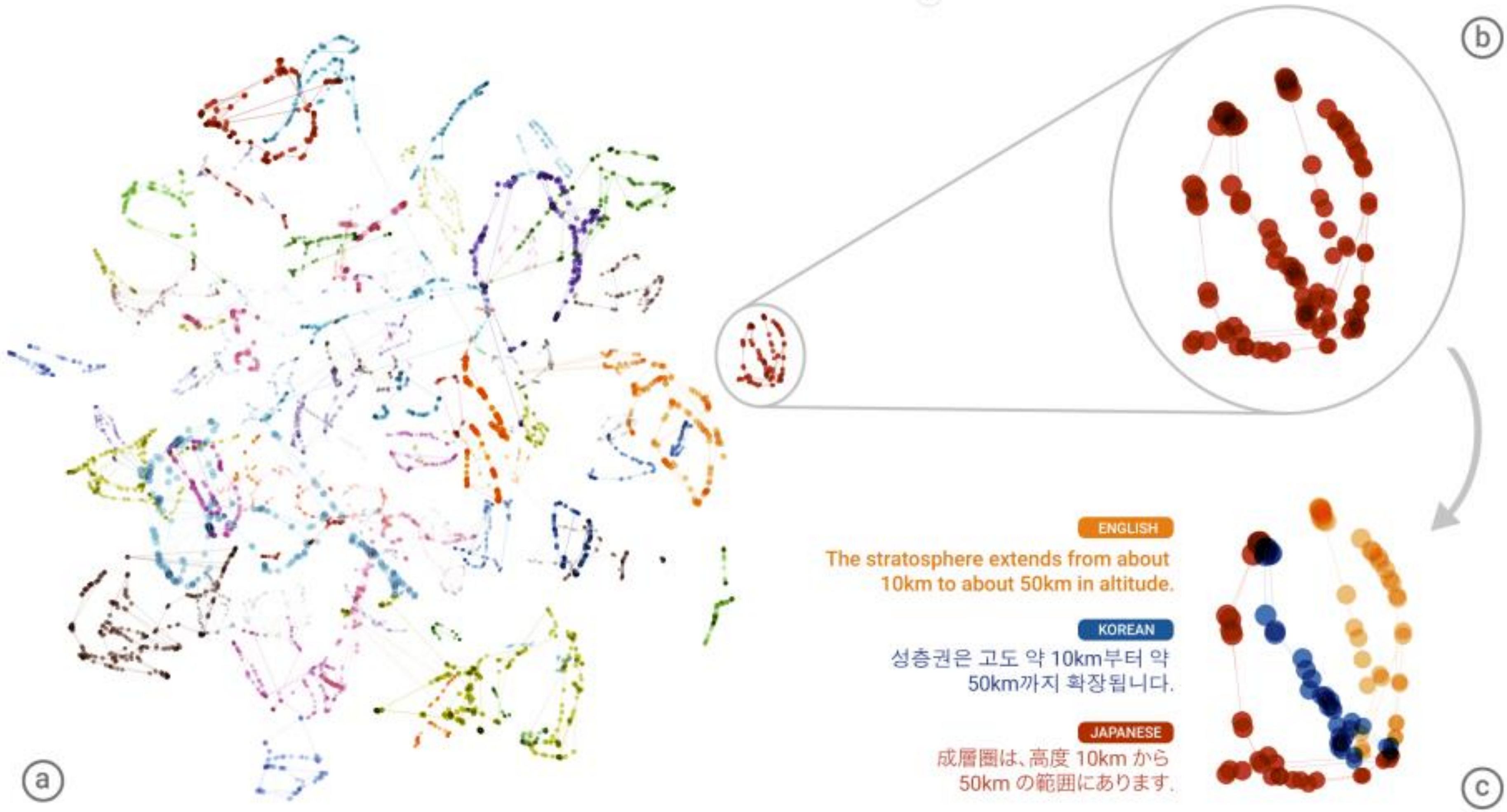
Task：文字翻譯



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

上例的結果做visualization，每個點代表一個句子(data)



Transfer Learning - Overview

self–taught learning：與semi–supervise相近，然而其
unlabel data與label data可能無關（不同類）

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

藉由大量source data還是可以learn出許多不錯的feature repusentation

Self-taught learning

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

source 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

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