Weak Points

- 待更新的参数有N*V个，N为词向量的维度，V为单词表中单词的数量
- 每次只能更新一个单词
  - 对于大语料来说，V很大，所以更新很慢
- 解决问题的途径：
  - Hierarchical softmax（层次软最大）
  - Negative sampling（负采样）

Mikolov, Tomas; et al. "Efficient Estimation of Word Representations in Vector Space". arXiv:1301.3781

Hierarchical Softmax

- 解决思路是，将输出端做成一个Hauffman树——可以对所有单词进行最短编码的树状结构
- 分为两个步骤：
  - 构建Hauffman树：每个节点都有一个可训练的参数，所有的单词都位于叶节点上
  - 训练
- 好处：每一次训练会更新一条分支上的所有参数，这些参数会影响2^{l-1}单词

$$p(w|w_I) = \prod_{j=1}^{L(w)-1} \sigma\left([\![n(w, j+1) = \mathrm{ch}(n(w,j))]\!] \cdot {v'_{n(w,j)}}^\top v_{w_I} \right)$$

n(w, j) be the j-th node on the path from the root to w, let ch(n) be an arbitrary fixed child of n and let [[x]] be 1 if x is true and -1 otherwise

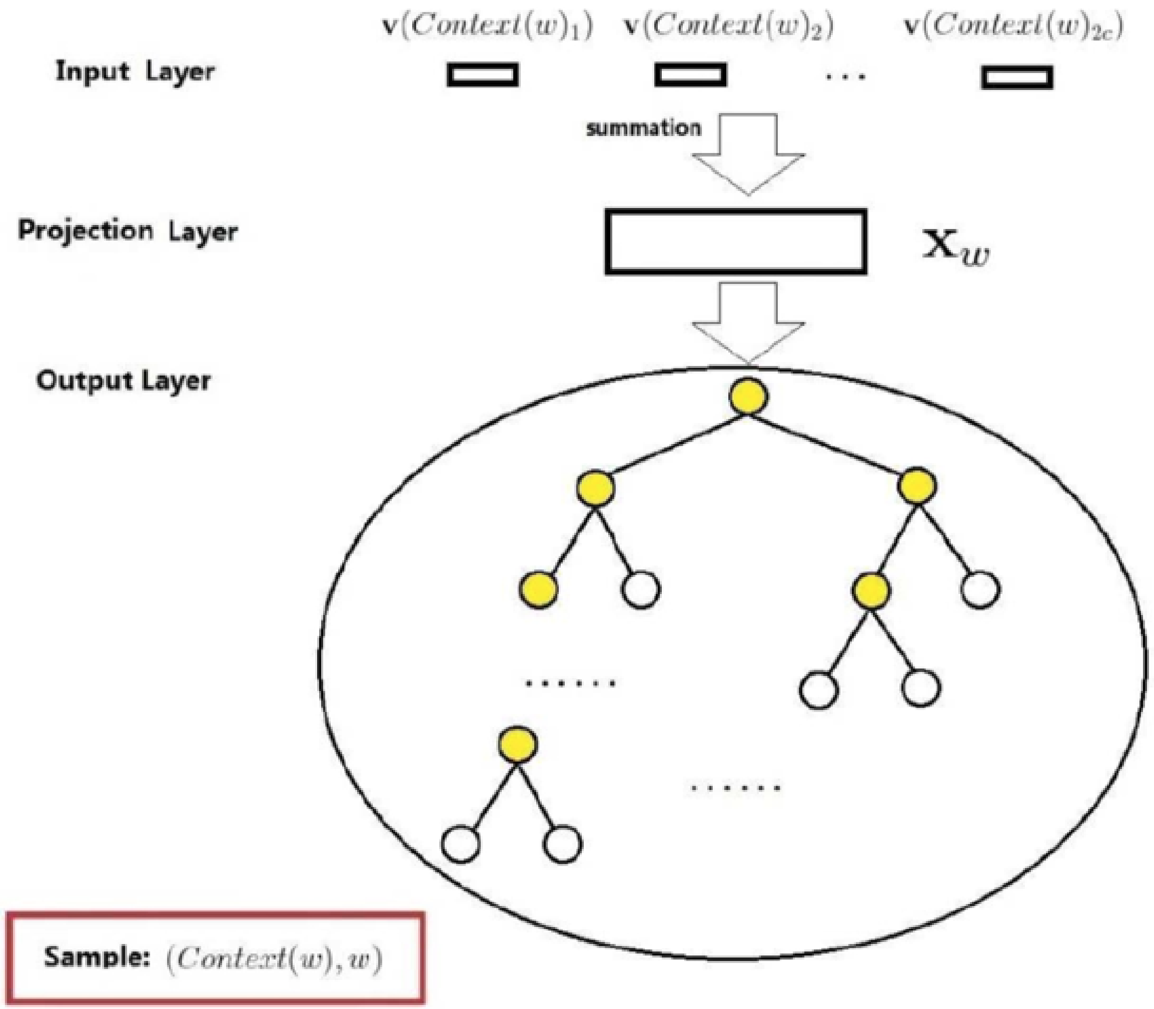


图 10 CBOW 模型的网络结构示意图

Negative Sampling

基本思路

- 将目标函数（似然函数）改为：
  - 对于在上下文中出现的单词，概率最大
  - 对于随机采样生成的不在上下文中的单词，概率最小

对于每一个训练样本：

- 我爱北京天安→门

生成一系列负样本：

- 我爱北京天安→地，瓜，……

Negative Sampling

基本思路

- 将目标函数（似然函数）改为：
  - 对于在上下文中出现的单词，概率最大
  - 对于随机采样生成的不在上下文中的单词，概率最小

$$L = -\log P_{\text{门}}$$

Negative Sampling

基本思路

- 将目标函数（似然函数）改为：
  - 对于在上下文中出现的单词，概率最大
  - 对于随机采样生成的不在上下文中的单词，概率最小

$$L = -(\log P_{门} + \log(1 - P_{地}) + \log(1 - P_{瓜}))$$

$$\log \sigma({v'_{w_O}}^\top v_{w_I}) + \sum_{i=1}^{k} \mathbb{E}_{w_i \sim P_n(w)} \left[\log \sigma(-{v'_{w_i}}^\top v_{w_I}) \right]$$

$$P(w_i) = \frac{f(w_i)^{0.75}}{\sum_{j=0}^{n} (f(w_j))^{0.75}}$$

Word Vectors

GloVe: Global Vectors for Word Representation

Jeffrey Pennington, Richard Socher, Christopher D. Manning

Introduction

GloVe is an unsupervised learning algorithm for obtaining vector representations for words. Training is performed on aggregated global word-word co-occurrence statistics from a corpus, and the resulting representations showcase interesting linear substructures of the word vector space.

Getting started (Code download)

- Download the latest latest code (licensed under the Apache License, Version 2.0). Look for "Clone or download"
- Unpack the files: unzip master.zip
- Compile the source: cd GloVe-master && make
- Run the demo script: ./demo.sh
- Consult the included README for further usage details, or ask a question

Download pre-trained word vectors

- Pre-trained word vectors. This data is made available under the Public Domain Dedication and License v1.0 whose full text can be found at: http://www.opendatacommons.org/licenses/pddl/1.0/.
  - Wikipedia 2014 + Gigaword 5 (6B tokens, 400K vocab, uncased, 50d, 100d, 200d, & 300d vectors, 822 MB download): glove.6B.zip
  - Common Crawl (42B tokens, 1.9M vocab, uncased, 300d vectors, 1.75 GB download): glove.42B.300d.zip
  - Common Crawl (840B tokens, 2.2M vocab, cased, 300d vectors, 2.03 GB download): glove.840B.300d.zip
  - Twitter (2B tweets, 27B tokens, 1.2M vocab, uncased, 25d, 50d, 100d, & 200d vectors, 1.42 GB download): glove.twitter.27B.zip
- Ruby script for preprocessing Twitter data

Citing GloVe

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. [pdf] [bib]

https://nlp.stanford.edu/projects/glove/

Resources

- English word vectors
- Word vectors for 157 languages
- Wiki word vectors
- Aligned word vectors
- Supervised models
- Language identification
- Datasets

English word vectors

This page gathers several pre-trained word vectors trained using fastText.

Download pre-trained word vectors

Pre-trained word vectors learned on different sources can be downloaded below:

1. wiki-news-300d-1M.vec.zip: 1 million word vectors trained on Wikipedia 2017, UMBC webbase tokens).
2. wiki-news-300d-1M-subword.vec.zip: 1 million word vectors trained with subword infomation c statmt.org news dataset (16B tokens).
3. crawl-300d-2M.vec.zip: 2 million word vectors trained on Common Crawl (600B tokens).
4. crawl-300d-2M-subword.zip: 2 million word vectors trained with subword information on Comi

Format

The first line of the file contains the number of words in the vocabulary and the size of the vectors. like in the default fastText text format. Each value is space separated. Words are ordered by descer loaded in Python using the following code:

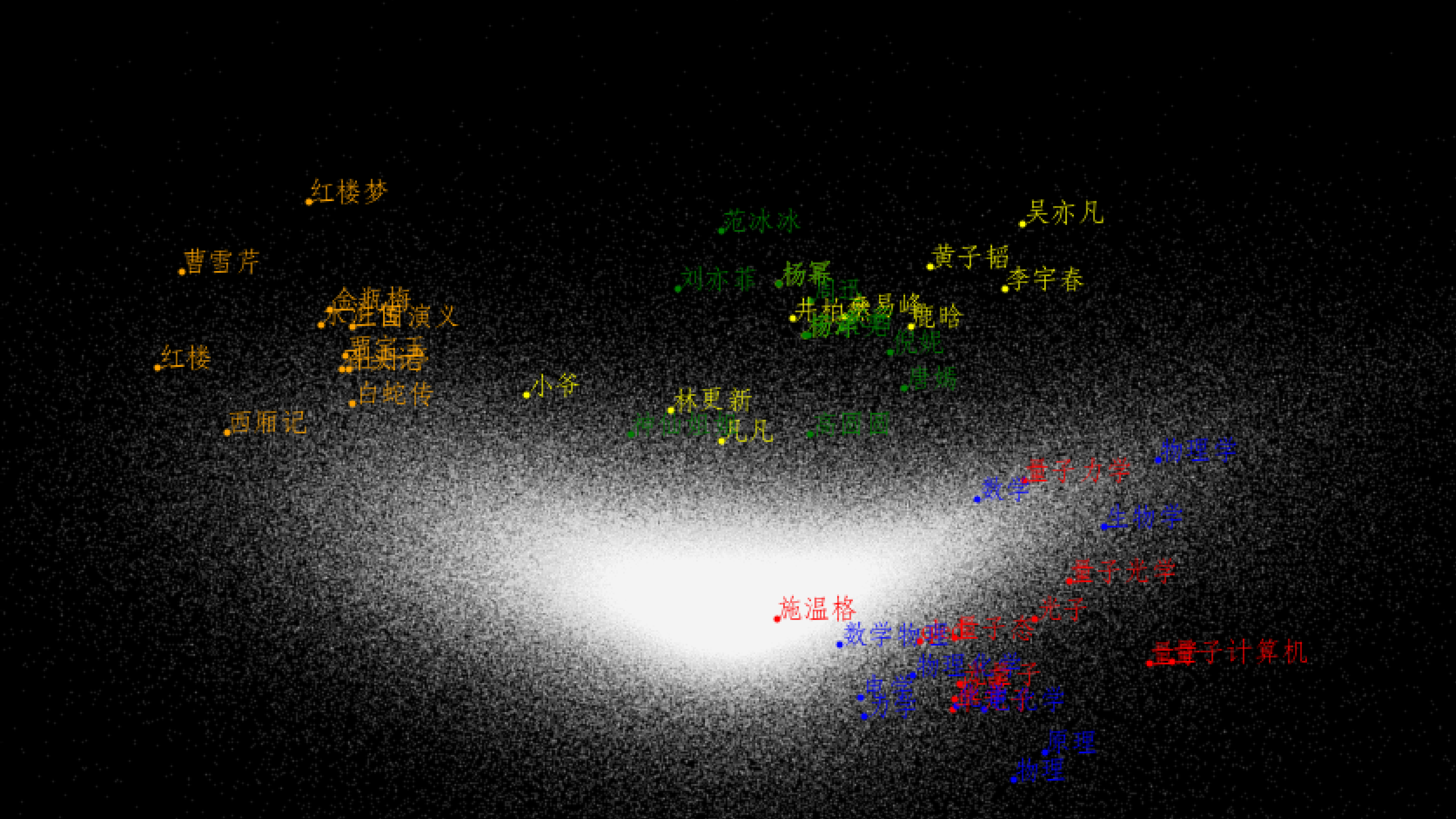
```
import io

def load_vectors(fname):
    fin = io.open(fname, 'r', encoding='utf-8', newline='\n', errors='ignore')
    n, d = map(int, fin.readline().split())
    data = {}
    for line in fin:
        tokens = line.rstrip().split(' ')
        data[tokens[0]] = map(float, tokens[1:])
    return data
```

中文词向量

Corpus	Context Features			
	Word	Word + Ngram	Word + Character	Word + Character + Ngram
Baidu Encyclopedia 百度百科	300d	300d	300d	300d
Wikipedia_zh 中文维基百科	300d	300d	300d	300d
People's Daily News 人民日报	300d	300d	300d	300d
Sogou News 搜狗新闻	300d	300d	300d	300d
Financial News 金融新闻	300d	300d	300d	300d
Zhihu_QA 知乎问答	300d	300d	300d	300d
Weibo 微博	300d	300d	300d	300d
Literature 文学作品	300d	300d	300d	300d
Complete Library in Four Sections 四库全书*	300d	300d	NAN	NAN
Mixed-large 综合	300d	300d	300d	300d

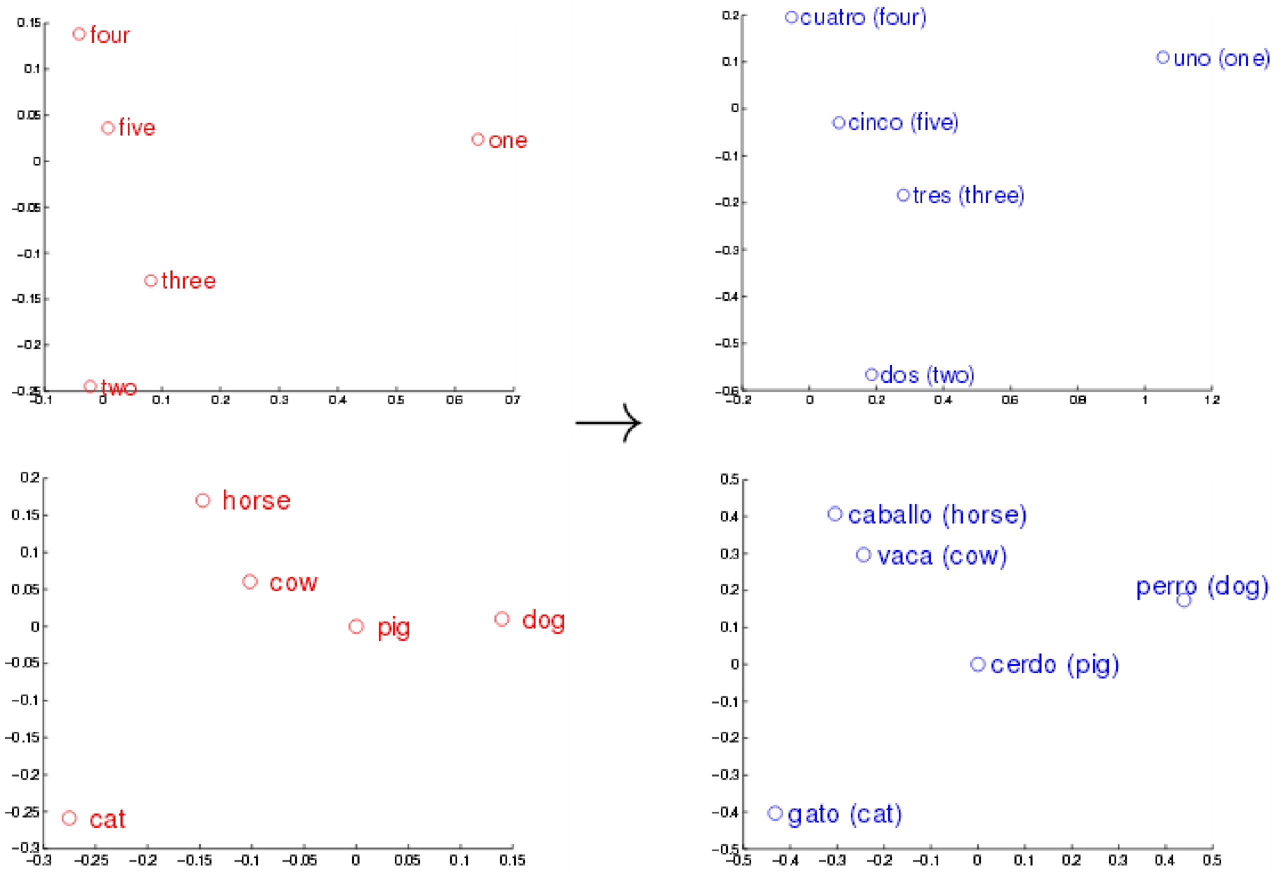
Word2ec / Skip-Gram with Negative Sampling (SGNS)

https://zhuanlan.zhihu.com/p/36872508



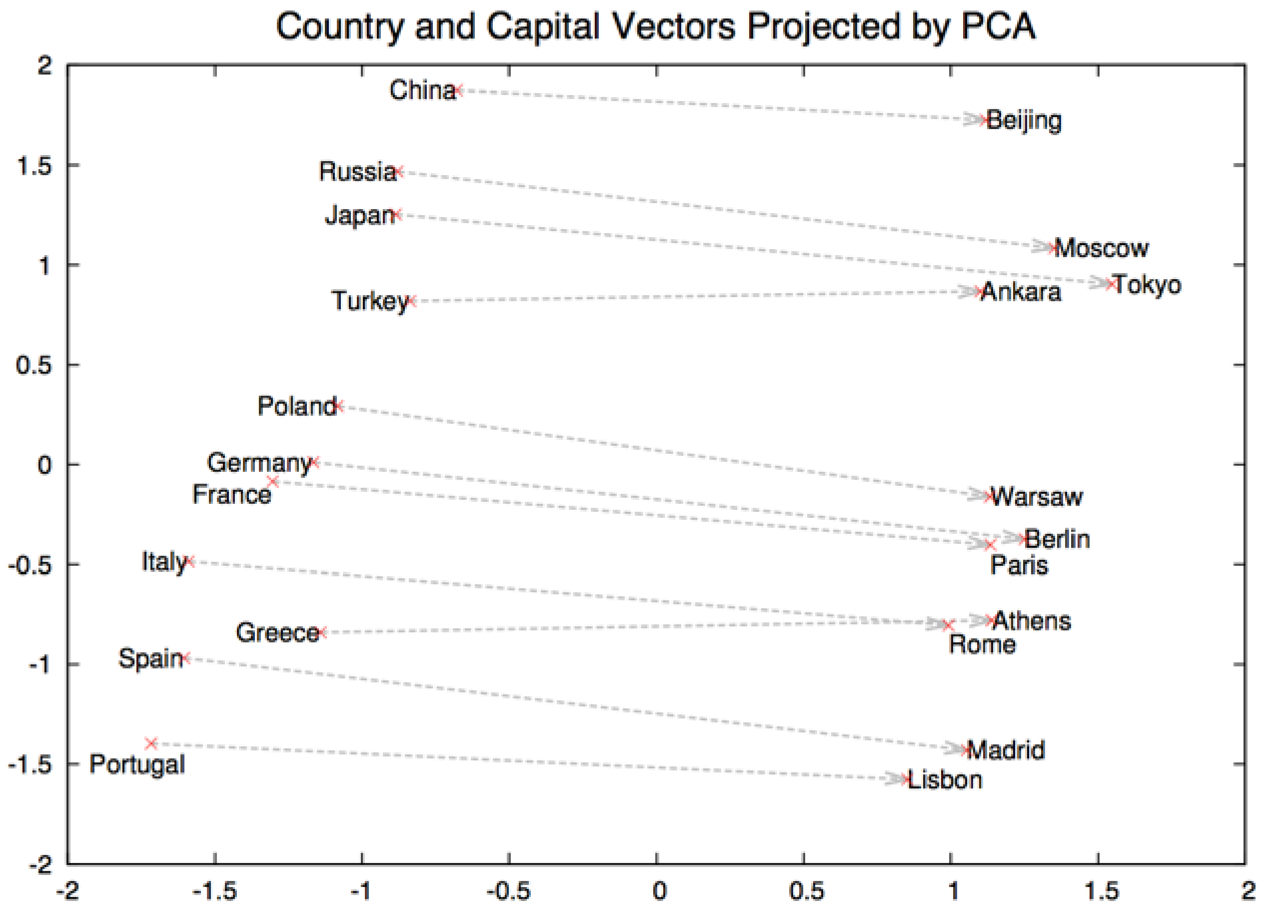
红楼梦
曹雪芹
金瓶梅
红楼
白蛇传
西厢记
小爷
范冰冰
刘亦菲
杨幂
吴亦凡
黄子韬
李宇春
鹿晗
唐嫣
林更新
凡凡
高圆圆
物理学
量子力学
数学
生物学
量子光学
施温格
光子
量子态
数学物理
量子计算机
原理
物理
力学

Difference of Vectors and Relationship

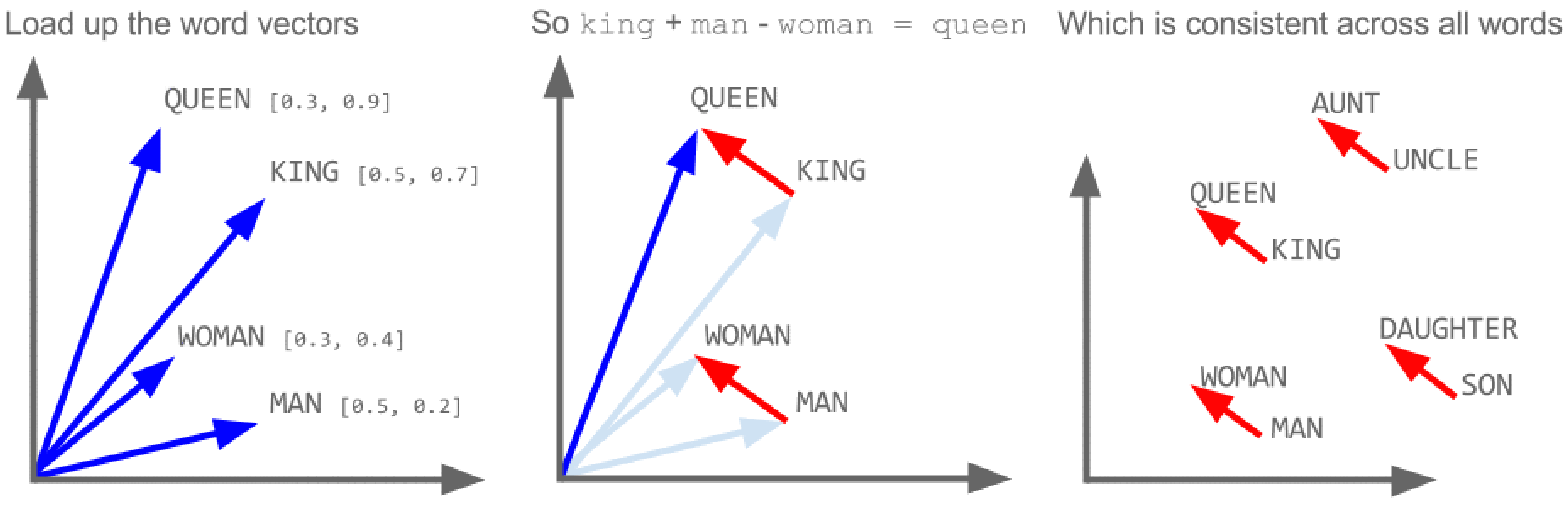


https://arxiv.org/pdf/1310.4546.pdf

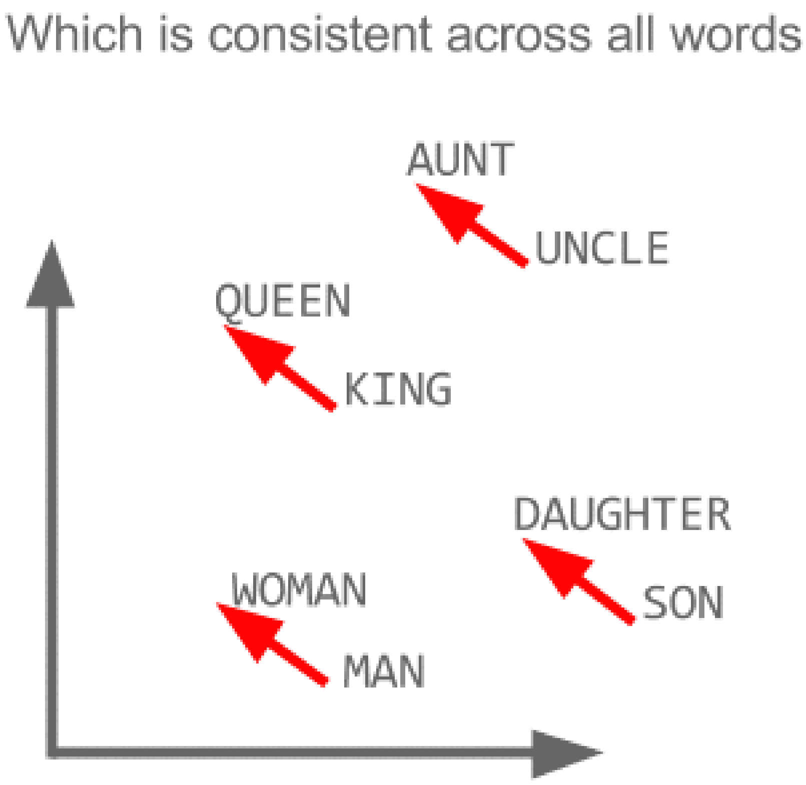
Difference of Vectors and Relationship

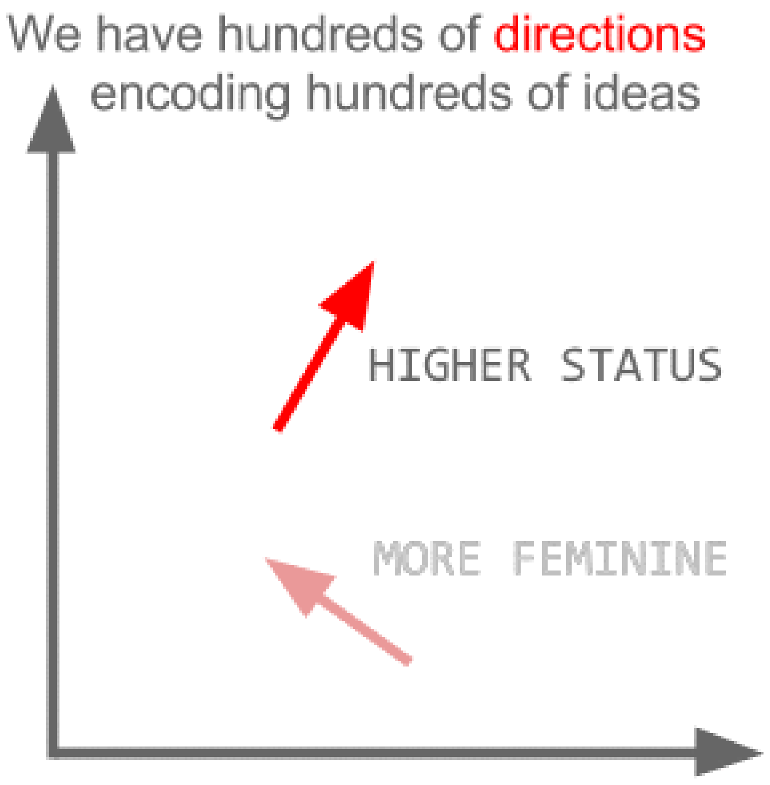


Analogy


Load up the word vectors
QUEEN [0.3, 0.9]
KING [0.5, 0.7]
WOMAN [0.3, 0.4]
MAN [0.5, 0.2]
So king + man - woman = queen
QUEEN
KING
WOMAN
MAN
Which is consistent across all words
AUNT
UNCLE
QUEEN
KING
DAUGHTER
SON
WOMAN
MAN

Orthogonal Directions





Gender & Class

American Sociological Review

Impact Factor: **12.444** / 5-Year Impact Factor: **13.153** | JOUR

Available access | Research article | First published online September 25, 2019

The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings

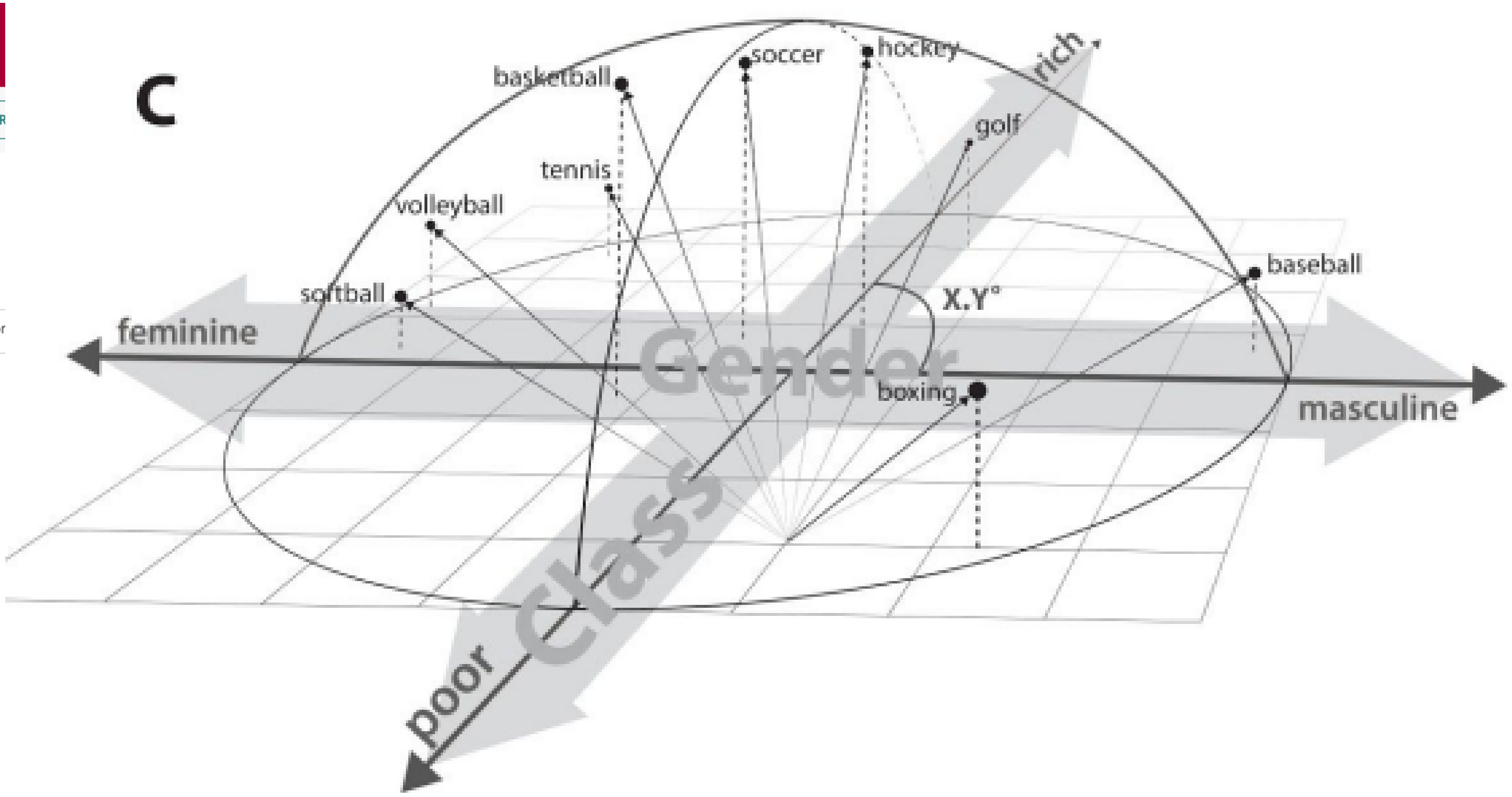
Austin C. Kozlowski, Matt Taddy, and James A. Evans | View all authors and affiliations

Volume 84, Issue 5 | https://doi.org/10.1177/0003122419877135 | View correction

PDF / ePub | Cite article | Share options | Information, rights and permissions | Metrics and citation

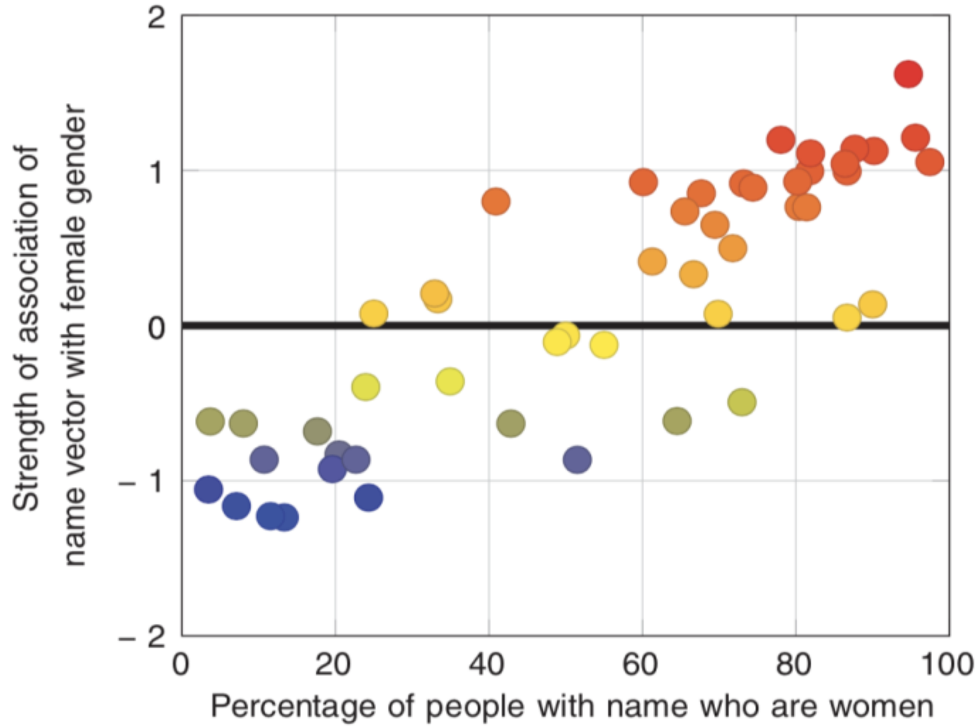
Abstract

We argue word embedding models are a useful tool for the study of culture using a historical analysis of shared understandings of social class as an empirical case. Word embeddings represent semantic relations between words as relationships between vectors in a high-dimensional space, specifying a relational model of meaning consistent with contemporary theories of culture. Dimensions induced by word differences (*rich – poor*) in these spaces correspond to dimensions of cultural meaning, and the projection of words onto these dimensions reflects widely shared associations, which we validate with surveys. Analyzing text from millions of books published over 100 years, we show that the markers of class continuously shifted amidst the economic transformations of the twentieth century, yet the basic cultural dimensions of class remained remarkably stable. The notable exception is education, which became tightly linked to affluence independent of its association with cultivated taste.



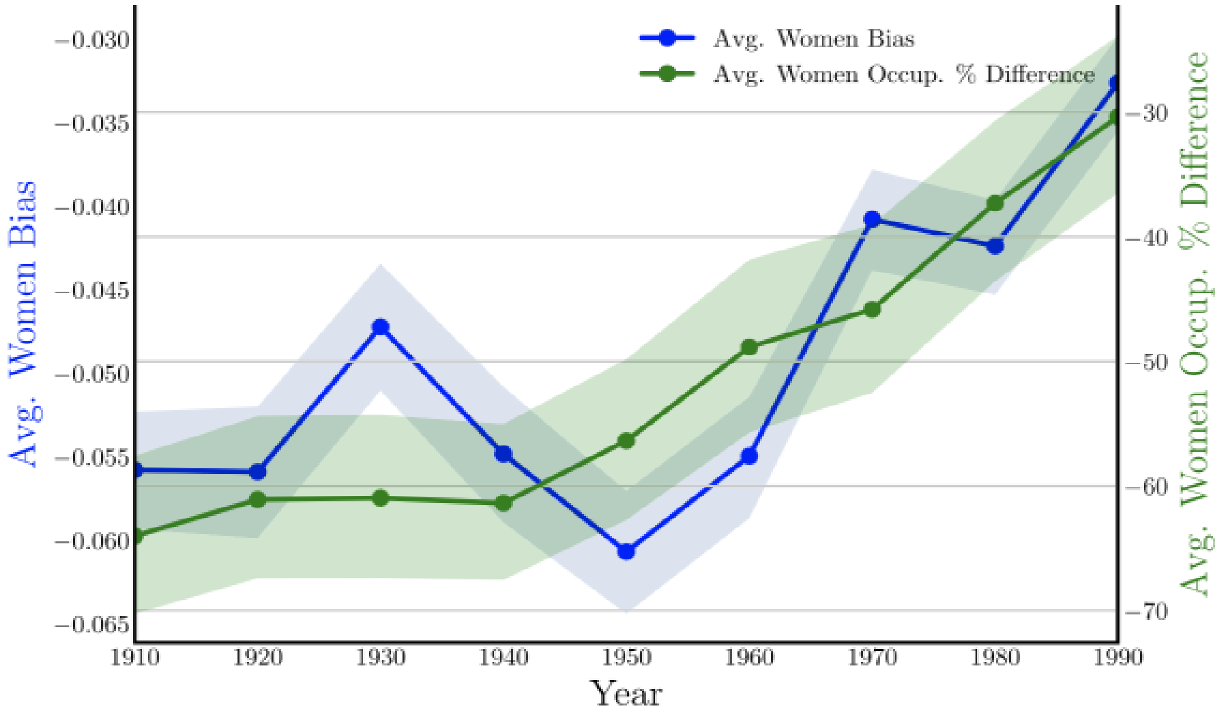
https://arxiv.org/abs/1803.09288

Word & Culture



Caliskan et al., Science 356, 183–186 (2017)

Word & Culture



Nikhil Garg et al.: PNAS vol 115, no. 16, E3635-E3644, 2018

Why it works?

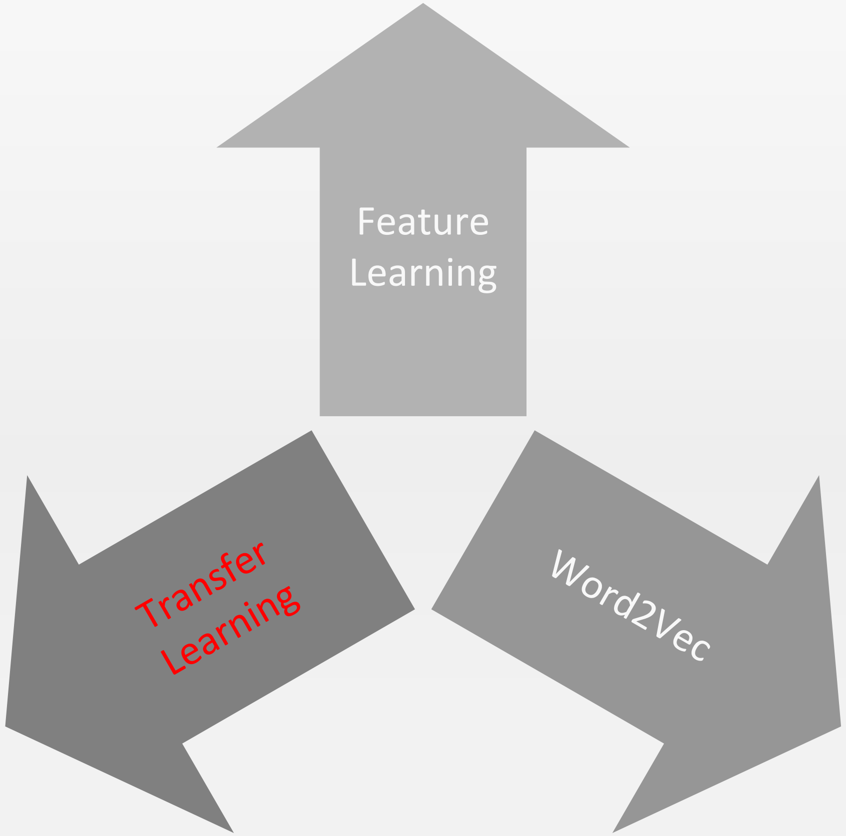
相似的上下文会给出相似的预测

如果两个单词经常出现在相似的上下文，它们就相似

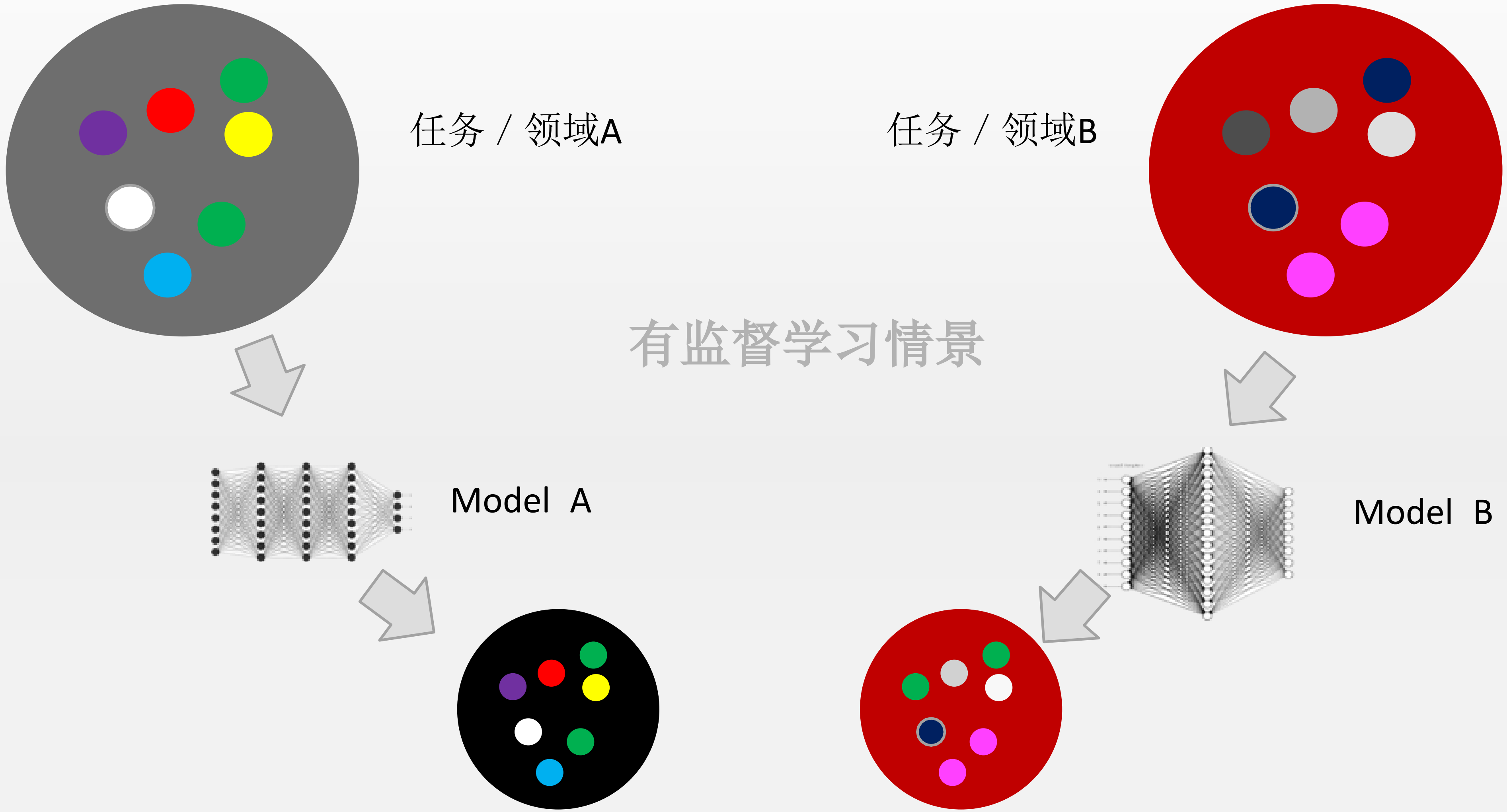
例如：

- 火星是太阳系中的一颗行星
- 水星是太阳系中有一颗行星

Outline



Transfer Learning



Transfer Learning

AI Fights Poverty

RESEARCH ARTICLES

ECONOMICS

Combining satellite imagery and machine learning to predict poverty

Neal Jean,[1,2]* Marshall Burke,[3,4,5]*† Michael Xie,[1] W. Matthew Davis,[4] David B. Lobell,[3,4] Stefano Ermon[1]

Reliable data on economic livelihoods remain scarce in the developing world, hampering efforts to study these outcomes and to design policies that improve them. Here we demonstrate an accurate, inexpensive, and scalable method for estimating consumption expenditure and asset wealth from high-resolution satellite imagery. Using survey and satellite data from five African countries—Nigeria, Tanzania, Uganda, Malawi, and Rwanda—we show how a convolutional neural network can be trained to identify image features that can explain up to 75% of the variation in local-level economic outcomes. Our method, which requires only publicly available data, could transform efforts to track and target poverty in developing countries. It also demonstrates how powerful machine learning techniques can be applied in a setting with limited training data, suggesting broad potential application across many scientific domains.

Accurate measurements of the economic characteristics of populations critically influence both research and policy. Such measurements shape decisions by individual governments about how to allocate scarce resources and provide the foundation for global efforts to understand and track progress toward improving human livelihoods. Al

tries had no DHS asset-based surveys taken, and an additional 19 had only one. These shortcomings have prompted calls for a "data revolution" to sharply scale up data collection efforts within Africa and elsewhere (1). But closing these data gaps with more frequent household surveys is likely to be both prohibitively costly—perhaps costing hundreds of billions of U.S. dollars to

ECONOMICS

Fighting poverty with data

Machine learning algorithms measure and target poverty

By **Joshua Evan Blumenstock**

Policy-makers in the world's poorest countries are often forced to make decisions based on limited data. Consider Angola, which recently conducted its first postcolonial census. In the 44 years that elapsed between the prior census and the recent one, the country's population grew from 5.6 million to 24.3 million, and the country experienced a protracted civil war that displaced millions of citizens. In situations where reliable survey data are missing or out of date, a novel line of research offers promising alternatives. On page 790 of this issue, Jean *et al.* (1) apply recent advances in machine learning to high-resolution satellite imagery to accurately measure regional poverty in Africa.

Traditionally, wealth and poverty are mea-

the economy of North Korea (7), where official statistics are dubious.

A series of studies in wealthy nations explore how data from the internet and social media can provide proxies for economic activity (8, 9). Mining the tweets and search queries of millions of individuals promises real-time alternatives to more traditional methods of data collection. However, these approaches are less relevant to remote and developing regions, where internet infrastructure is limited and few people use social media.

In developing countries, researchers have found ways to measure wealth and poverty using the digital footprints left behind in the transaction logs of mobile phones, which are increasingly ubiquitous even in very poor regions. Regional patterns of mobile phone use correlate with the regional distribution of wealth (10). This relationship persists at

Poverty

非洲贫困国家：尼日利亚、坦桑尼亚、乌甘达、马拉维、卢旺达

每天1.9美元的贫困线，该地区极贫人口比例预计在2015年将降至35%

从1990年的56%到2012年降至43%

捐助甚至会加重贫富差距

精准获取数据是解决非洲贫困的首要问题

Poverty

非洲贫困国家：尼日利亚、坦桑尼亚、乌甘达、马拉维、卢旺达

每天1.9美元的贫困线，该地区极贫人口比例预计在2015年将降至35%

从1990年的56%到2012年降至43%

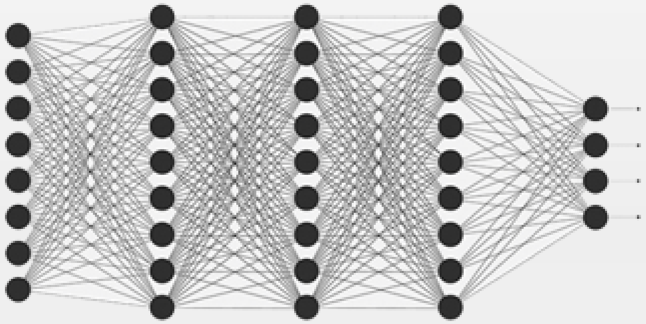
捐助甚至会加重贫富差距

精准获取数据是解决非洲贫困的首要问题

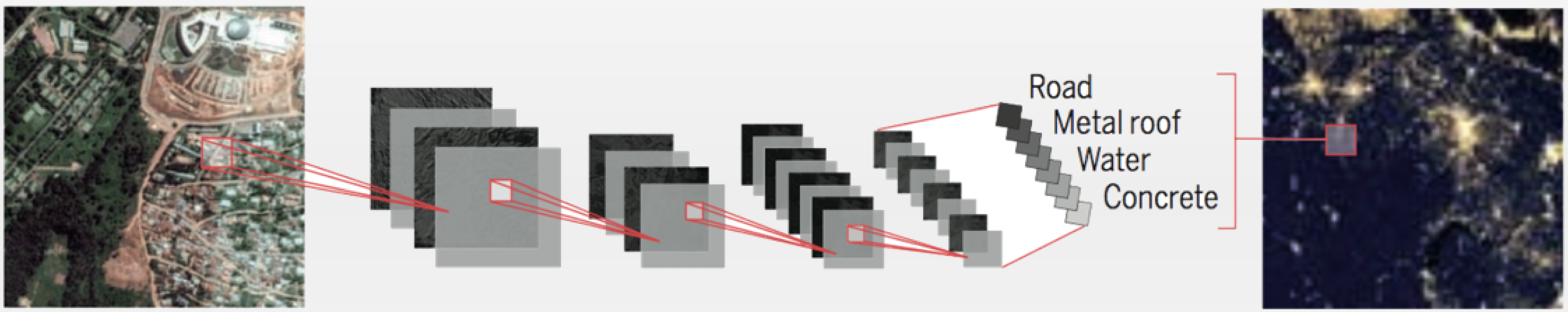
数据？

黄海
瀬戸内海
Awaji-shima
Shikoku
Cheju-do
Hirado-shima
Fukue-jima
Kyushu
Tanega-shima
东海
湖南省
Okinawa-jima
台北市
Miyako-jima
Ishigaki-shima
台湾
壮族自治区
© 2011 Europa Technologies
© 2011 Google
© 2011 ZENRIN
© 2011 Mapabc.com
香港
香港
澳门
28°56'27.44" 北 122°35'27.89" 东 海拔 80 米
©2010 Google
菲律宾
CLASSIC023.COM
视角海拔高度 2268.91 公里

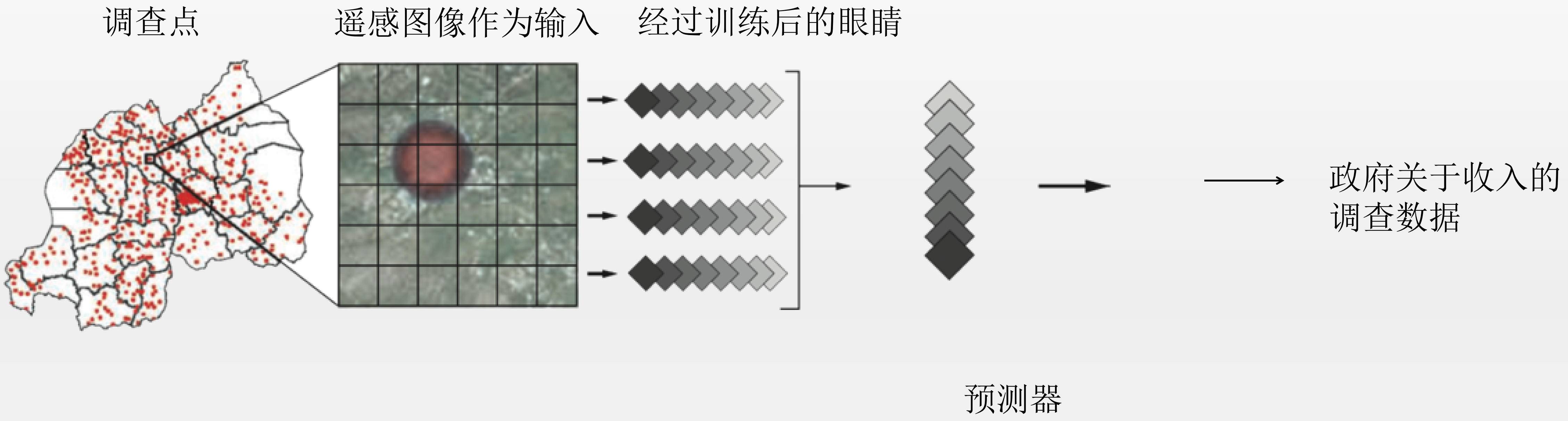
Difficulty in Prediction



Step 1: Predicting night lights



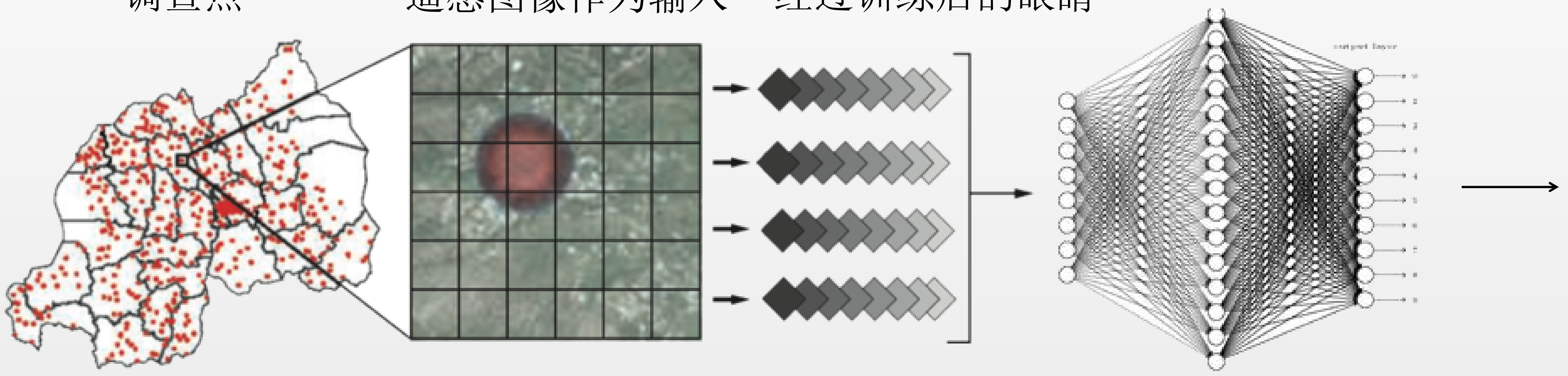
Step 2: Transfer Learning



运用数百个调查数据训练预测器

让熟悉卫星遥感图片的眼睛来指导预测

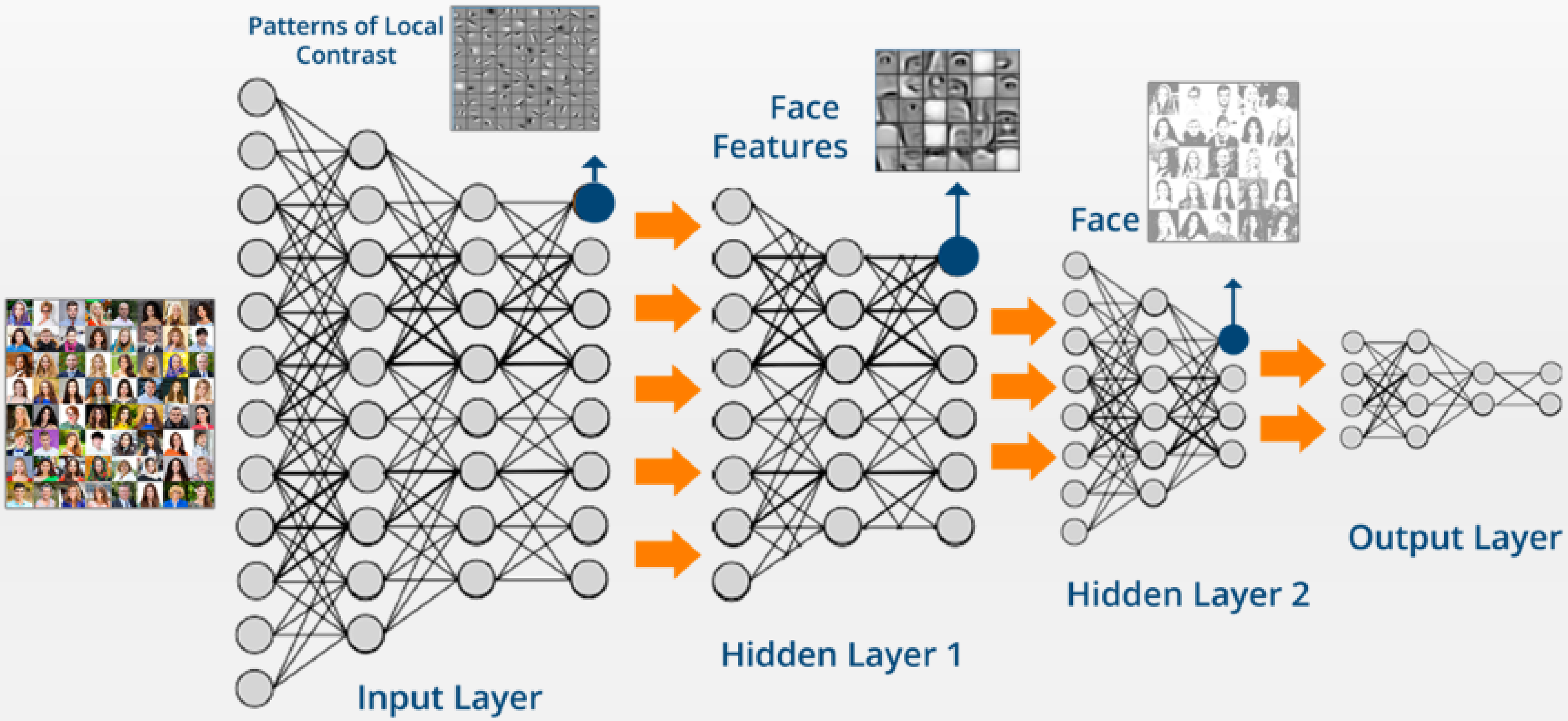
Step 3: Prediction



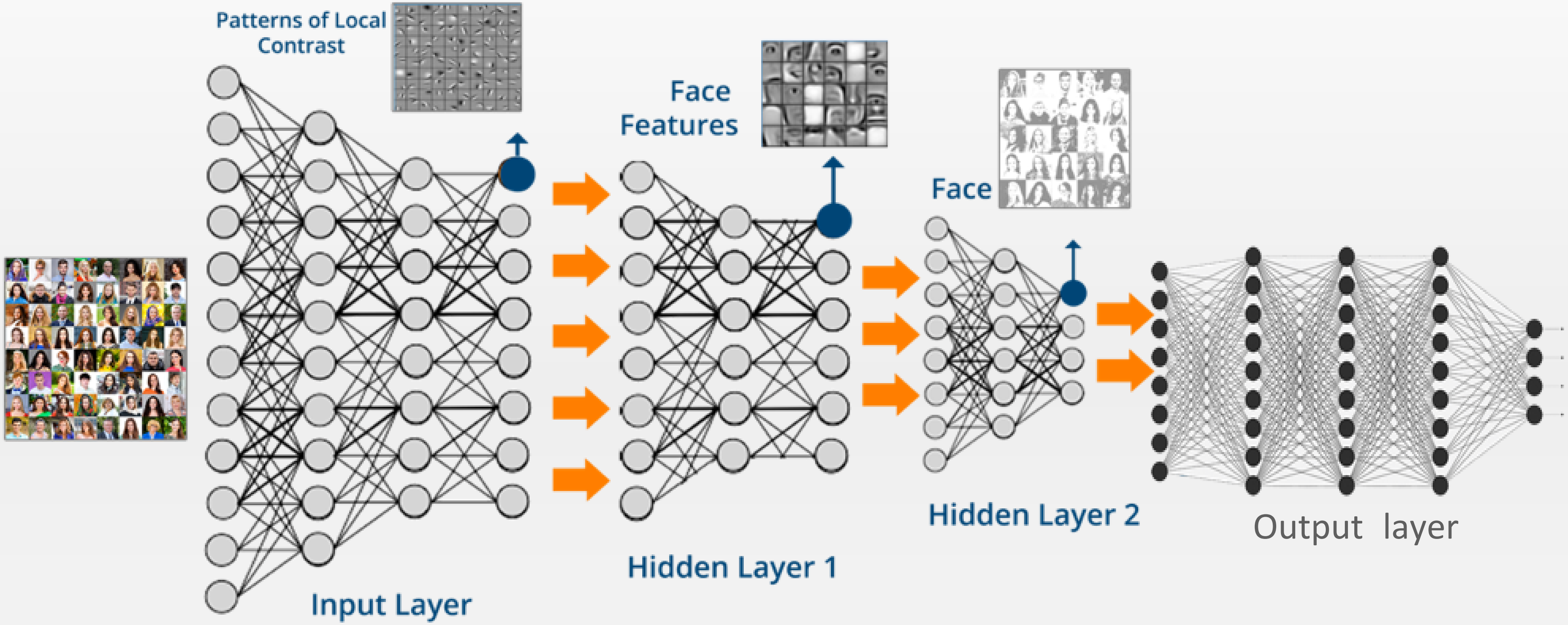
运用不同地点的图像，就可以预测该地点的贫困程度。

让熟悉卫星遥感图片的眼睛来指导预测

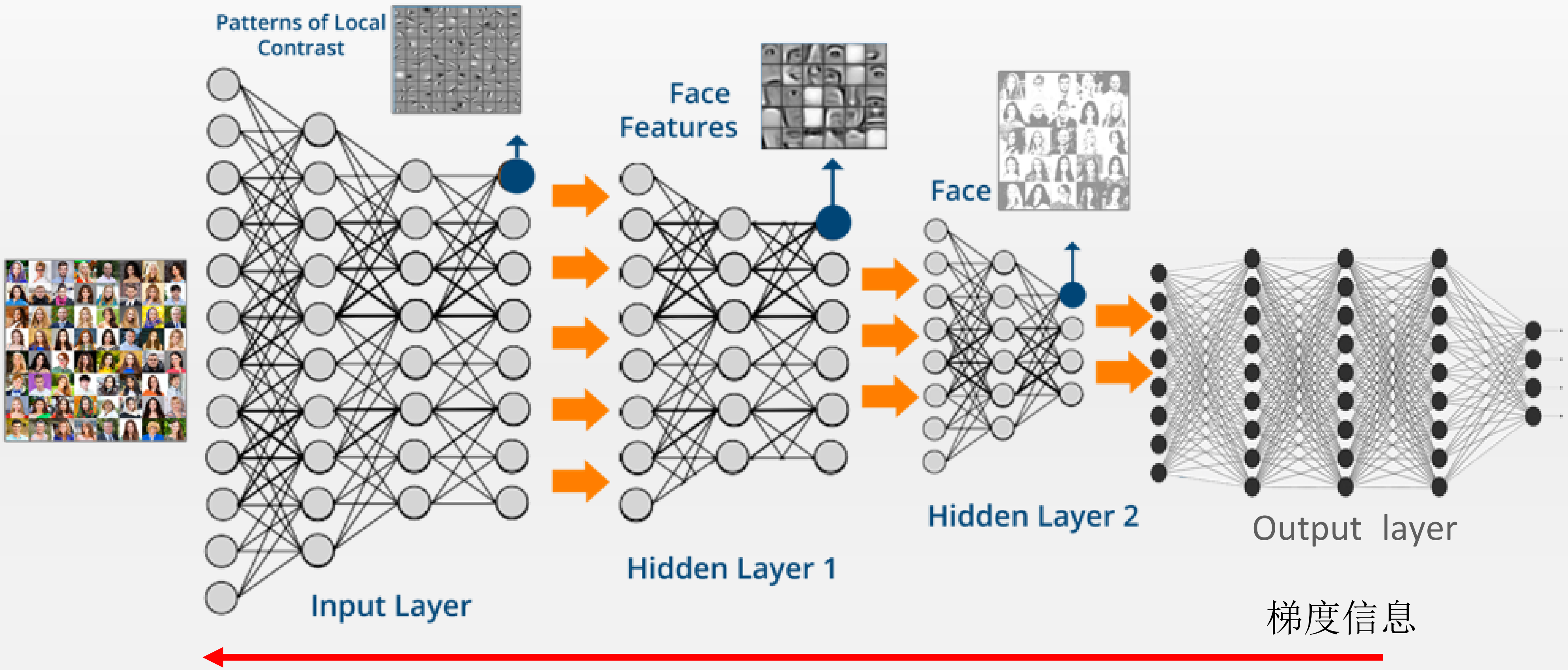
Transfer Learning in CNN



Transfer Learning in CNN

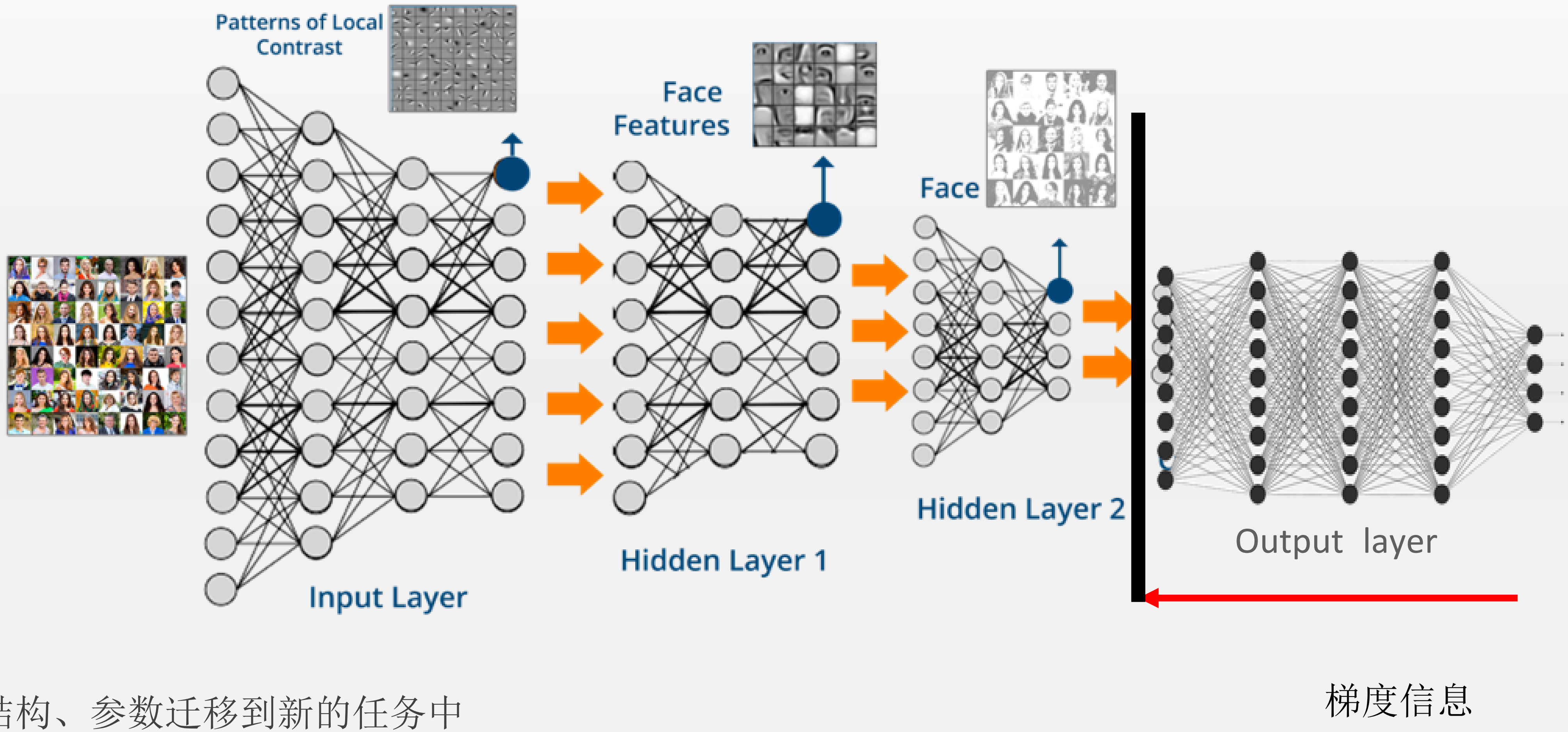


Transfer Learning in CNN



预训练：
- 将网络的结构、参数迁移到新的任务中
- 参数的初始化取值设定为原网络的数值
- 所有的参数重新训练

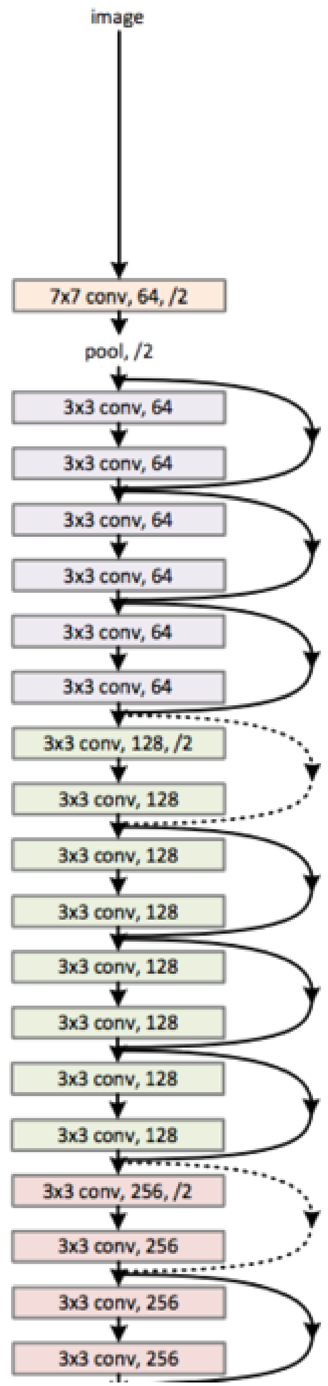
Transfer Learning in CNN



预训练：

- 将网络的结构、参数迁移到新的任务中
- 参数的初始化取值设定为原网络的数值
- 所有的参数重新训练

Download Large Model

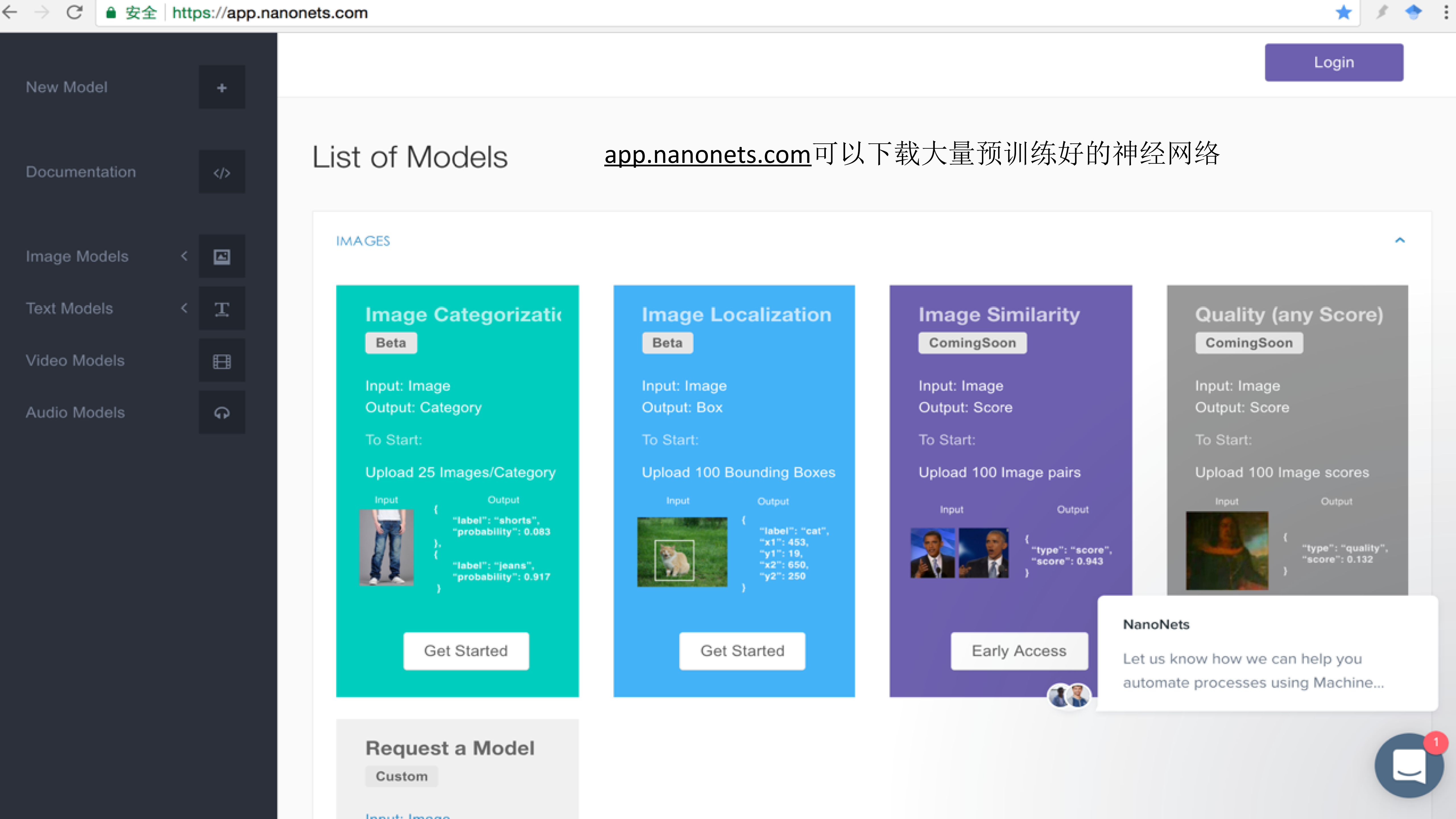


ResNet (

```
(conv1): Conv2d(3, 64, kernel_size=(7, 7), stride=(2, 2), padding=(3, 3), bias=False)
(bn1): BatchNorm2d(64, eps=1e-05, momentum=0.1, affine=True)
(relu): ReLU (inplace)
(maxpool): MaxPool2d (size=(3, 3), stride=(2, 2), padding=(1, 1), dilation=(1, 1))
(layer1): Sequential (
  (0): BasicBlock (
     (conv1): Conv2d(64, 64, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1), bias=False)
     (bn1): BatchNorm2d(64, eps=1e-05, momentum=0.1, affine=True)
     (relu): ReLU (inplace)
```

```
net = models.resnet18(pretrained=True)
```

```
     (conv2): Conv2d(64, 64, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1), bias=False)
     (bn2): BatchNorm2d(64, eps=1e-05, momentum=0.1, affine=True) ) )
(layer2): Sequential (
  (0): BasicBlock (
     (conv1): Conv2d(64, 128, kernel_size=(3, 3), stride=(2, 2), padding=(1, 1), bias=False)
     (bn1): BatchNorm2d(128, eps=1e-05, momentum=0.1, affine=True)
     (relu): ReLU (inplace)
     (conv2): Conv2d(128, 128, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1), bias=False)
```

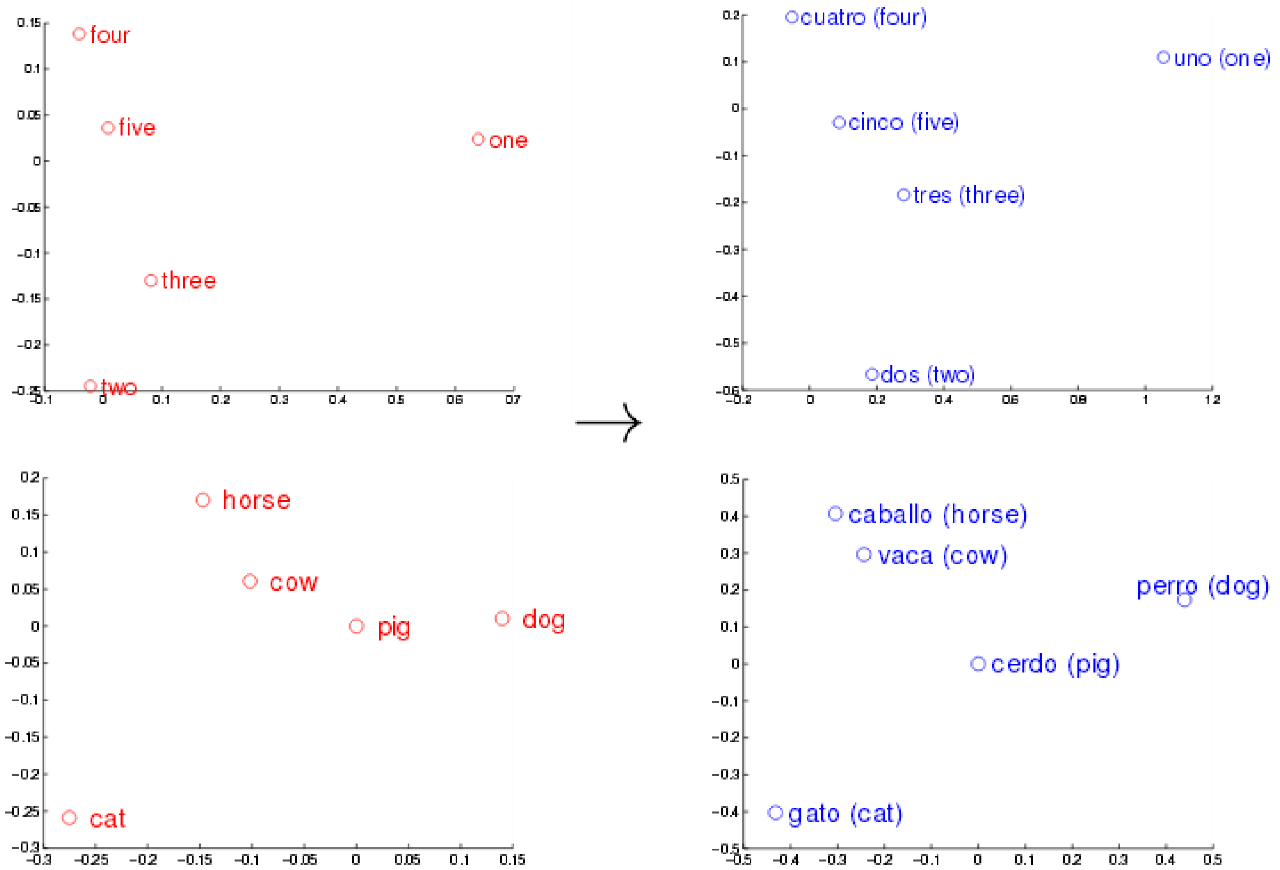


安全 https://app.nanonets.com
Login
New Model
Documentation
Image Models
Text Models
Video Models
Audio Models
List of Models
app.nanonets.com可以下载大量预训练好的神经网络
IMAGES
Image Categorizati
Beta
Input: Image
Output: Category
To Start:
Upload 25 Images/Category
Input
Output
"label": "shorts", "probability": 0.083
"label": "jeans", "probability": 0.917
Get Started
Image Localization
Beta
Input: Image
Output: Box
To Start:
Upload 100 Bounding Boxes
Input
Output
"label": "cat", "x1": 453, "y1": 19, "x2": 650, "y2": 250
Get Started
Image Similarity
ComingSoon
Input: Image
Output: Score
To Start:
Upload 100 Image pairs
Input
Output
"type": "score", "score": 0.943
Early Access
Quality (any Score)
ComingSoon
Input: Image
Output: Score
To Start:
Upload 100 Image scores
Input
Output
"type": "quality", "score": 0.132
NanoNets
Let us know how we can help you automate processes using Machine...
Request a Model
Custom
Input: Image

Recaps

- Representation Learning
  - Auto-encoder
- Word2Vec
  - Language Model
  - CBOW, Skipgram
  - Negative Sampling
- Transfer Learning

作 ： 向量翻 器



- 运用Google训练好的英文词向量：
    - 验证：man-woman=king-queen, man-woman=son-daughter, 等
- 分别将中文词向量中的一、二、三、……和英文词向量中的one,two,three,……画在二维平面上，并比较二者
- 分别将中文词向量中的马、牛、驴、……和英文词向量中的horse,cow,donkey,……等动物名字画在二维平面上，并比较二者
- 比较更多的中英文对译词对
- 训练一个神经网络，做到自动将输入的英文词翻译为中文

https://arxiv.org/pdf/1310.4546.pdf