6 why deep

2020年5月24日 星期日 下午2:27

Deep → Modularization → Less training data?

Modularization → classifiers → output

Image (input) → Modularization

— Modularization ~ speech

hierarchical structure of human languages

1. speech recognition

classification: Input → acoustic feature, output → state

... acoustic features

states a b a b b c

使用 GMM (Gaussian Mixture Model) → P(x|"t-d+uw1")

DNN:

Input

$\begin{bmatrix} x_1 \\ \vdots \\ x_i \end{bmatrix}$ → DNN → o o o o ...

$P(a|x_i)$ $P(b|x_i)$ $P(c|x_i)$ ...

Univerality theorem: $f: R^N \to R^M$ 所有function只需一层,足够多神经元

然而使用: deep structure 更有效率

— Analogy

— End-to-End Learning

— Semi-supervised learning:

A set of unlabeled data: $U \gg R$

Transductive learning: unlabeled data is the testing data

Inductive learning: unlabeled data is not ~ ~ ~

— Generative Model

1. Initialize: $\theta = \{P(C_1)\ P(C_2)\ \mu^1, \mu^2, \Sigma\}$

1.1 compute $P_\theta(C_1|x)$ Depending on model θ

1.2 Update:

$$P(C_1) = \frac{N_1 + \sum_{x^u} P(C_1|x^u)}{N}$$

N: total examples

N_1: Numbers to C_1

$$\mu^1 = \frac{1}{N} \sum_{x^r \in C_1} x^r + \frac{1}{\sum_{x^u} P(C_1|x^u)} \sum_{x^u} P(C_1|x^u) x^u$$

$$MLE: \log L(\theta) = \sum_{x^r} \log P_\theta(x^r, \hat{y}^r) + \sum_{x^u} \log P_\theta(x^u)$$

— self training

entropy-based regularization