Deep Neural Network for Speech Emotion Recognition

—A Study of Deep Learning—

Zhuowei Han

Institut für Signalverarbeitung
und Systemtheorie

Universität Stuttgart

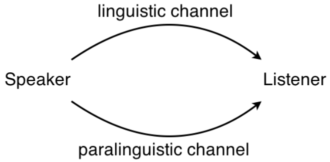
05.06.2014

Motivation

Why speech emotion recognition

- Most current work focuses on speech processing based on linguistic information, e.g.: Skype Translator
- More natural human-machine interaction requires paralinguistic information such as age, gender, emotion.
- Speech Recognition / Speeker Identification / Emotion

Recognition



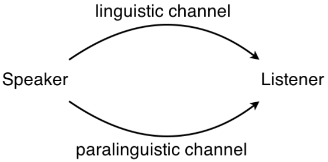
Deep Network Applications

- Handwriting Digit Recognition
- Image Recognition

Why speech emotion recognition

- Most current work focuses on speech processing based on linguistic information, e.g.: Skype Translator
- More natural human-machine interaction requires paralinguistic information such as age, gender, emotion.
- Speech Recognition / Speeker Identification / Emotion

Recognition



Deep Network Applications

- Handwriting Digit Recognition
- Image Recognition

Table of Contents

Foundations

Deep Neural Networks
- Concept
- Problems

Table of Contents

Foundations

Deep Neural Networks
- Concept
- Problems

Mel Frequency Cepstral Features

- short-term power spectrum
- mel-scale approximate human perception
- widely-used in speech recognition tasks

Emotion Recognition Approaches

Traditional Approaches

- pre-selected features
- supervised training
- low-level features not appropriate for claasification
- shallow structure of classifiers

Deep Learning Approaches

- learning representations from high-dim data
- extracting appropriate features without hand-crafting
- low-level features are used to build high-level features as network gets deeper
-

Table of Contents

Foundations

Deep Neural Networks

Concept

Problems

Computing net-activation

$$\begin{aligned}\underline{z}_k^{(l+1)} &= \mathbf{W}^{(l)}\underline{a}_k^{(l)} + \underline{b}^{(l)} \\ \underline{a}_k^{(l+1)} &= \underline{\mathit{\Phi}}\left(\underline{z}_k^{(l+1)}\right) \\ \hat{\underline{y}}_k &= \underline{a}_k^{(ol)}\end{aligned}$$

- Arbitrary non-linear mapping from $\underline{x}_k$ to $\hat{\underline{y}}_k$ possible
- Relation $N \Leftrightarrow$ Complexity
- Deep Architectures $(l \uparrow)$ more efficient than shallow ones $(l \downarrow, N_l \uparrow)$

Determining the parameters

Training objective

$$J(\mathbf{W}, \underline{b}) = \sum_{\forall k} \frac{1}{2} ||\underline{y}_k - \hat{\underline{y}}_k||^2 + \frac{\lambda}{2} \sum_{\forall l} ||\mathbf{W}^{(l)}||_F^2 \qquad (1)$$

$$\mathbf{W}, \underline{b} = \arg\min_{\mathbf{W}, \underline{b}} J(\mathbf{W}, \underline{b}) \qquad (2)$$

Numerical minimization

- Gradient calculation with Backpropagation
- Stochastic gradient descent
- **L**imited memory **B**royden-**F**letcher-**G**oldfarb-**S**hanno (L-BFGS)

Problems

- Optimization problem non-convex
  $\Rightarrow$ getting stuck in poor local minima
- Diffusion of gradients
- Large p small n problem $\Rightarrow$ overfitting