2017年6月16日　統計数理研究所　オープンハウス

統計的機械学習による音声認識研究

松井　知子　　モデリング研究系　教授

【概要】

本研究室では統計的学習機械を用いて、音声/音楽/画像/SNSなどを処理する方法について研究しています。具体的にはカーネルマシン、ブースティング、協調フィルタリングの手法を用いて、

1. 音声・話者認識
2. 音楽情報処理
3. 画像識別
4. SNS解析
5. WEBユーザビリティ評価
6. 都市インテリジェンス　など

の研究課題に取り組んでいます。

都市インテリジェンス

SNS

カーネルマシン
ブースティング
協調フィルタリング

画像情報処理

音情報処理

本研究室では統計的機械学習とその応用研究に興味のある学生さんを募集しています！

【統計的機械学習】

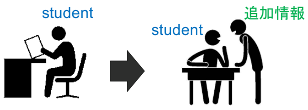
- 統計科学を用いて、
  - データから、内在する数学的な構造を発見する。
  - その数学的な構造に基づいて、予測や判別などの情報処理を行う。
- 帰納的アプローチ

  v.s.
- 自然科学でよく見られる演繹的アプローチ
  - 仮説をたて、推論し、実験的または理論的に検証する。

- カーネルマシン
  - 自動的な特徴(/モデル)選択機構を含む。
  - 非線形の扱いに優れている。
  - サポートベクターマシン(SVM)、罰金付ロジスティック回帰マシン
- いろいろな確率モデルによる方法
  - 混合ガウス分布モデル
  - 隠れマルコフモデル
- ガウス過程状態空間モデル　など

【Generalized distillation frameworkを用いた音声認識】

Generalized Distillation



先生がいる or 追加情報を使う方が学習が進む！

[D. D. Lopez-Paz, L. Bottou, B. Scho¨lkopf, and V. Vapnik, "Unifying distillation and privileged information," ICLR, 2016]

Hinton's Distillation

- Training many different models on same training data:
  - Improves the performance, but
  - Makes the whole model big and unsuitable in practice.
- How to train single, small model with simlar performance?
  - Use the output of the big model as "soft" targets for the small model – *Model Compression* (Caruana, 2006).
- When references, i.e. "hard" targets, are available (Hinton, 2015):
  - Combine the "soft" and "hard" targets and control the **softness** of the "soft" targets.
  - The big model is called *teacher* and the small one – *student*.
- Given the c-class classification task with training data $\{(x_i, y_i)\}_{i=1}^n \sim P^n(x,y)$, $x_i \in \mathbb{R}^d, y_i \in \mathbb{Q}^c$, where $\mathbb{Q}^c$ is a space of c-dimensional probability vectors, teacher training is to find:

$$f_t = \arg\min_{f \in \mathcal{F}_t} \frac{1}{n} \sum_{i=1}^n l\left(y_i, \sigma(f(x_i))\right) + \Omega(\|f\|)$$

where $\sigma()$ is a softmax, $l()$ is the loss, and $\Omega()$ is a regularizer.
- Then, for the student we have:

$$f_s = \arg\min_{f \in \mathcal{F}_s} \frac{1}{n} \sum_{i=1}^n \left[(1-\lambda) l\left(y_i, \sigma(f(x_i))\right) + \lambda l\left(s_i, \sigma(f(x_i))\right)\right]$$

where $s_i = \sigma\left(\frac{f_t(x_i)}{T}\right) \in \mathbb{Q}^c$ and $T > 0$ controls the smoothness.

Vapnik's Privileged Information

- Often during training some additional information is available which is **not** accessible during test time. Given training data

$$\{(x_i, x_i^*, y_i)\}_{i=1}^n \sim P^n(x, x_i^*, y)$$

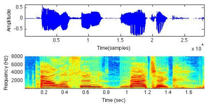
- How to leverage this information to make better model?
  - The naïve way – estimate the mapping $x \xrightarrow{f} x^*$ and generate x^* during testing.
  - Vapnik's way (restricted to SVMs):
    - *Similarity Control* (Vapnik, 2009). Implemented in SVM+ objective.
    - *Knowledge transfer* (Vapnik, 2015). Train f_t on $\{(x_i^*, y_i)\}_{i=1}^n$ and use it during the training of f_s on $\{(x_i, y_i)\}_{i=1}^n$.

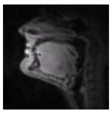
Generalized Distillation

- *Combination* of Hinton's distillation and Vapnik's privileged information approaches (Lopez-Pas, 2016).
- Three step process. Given training data $\{(x_i, x_i^*, y_i)\}_{i=1}^n$
  1. Learn teacher $f_t \in \mathcal{F}_t$ using $\{(x_i^*, y_i)\}_{i=1}^n$;
  2. Compute teacher "soft" labels $s_i = \sigma\left(\frac{f_t(x_i^*)}{T}\right)$ for some temperature T;
  3. Learn student $f_s \in \mathcal{F}_s$ using both $\{(x_i, y_i)\}_{i=1}^n$ and $\{(x_i, s_i)\}_{i=1}^n$, distillation objective and imitation parameter $\lambda \in [0,1]$.
- Generalized distillation reduces to:
  - Hinton's distillation when $x_i = x_i^*$.
  - Vapnik's method when x_i^* is privileged description of x_i.

Application in Speech Recognition

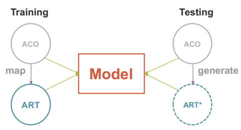
- Features for ASR:
  - Spectrum based – MFCC, FBANK, etc.
    - Main features, widely used.
    - Easy to obtain.
    - Highly variable.
    - Affected by noise, etc.
  - Articulatory movements based.
    - Not affected by noise.
    - Less variable.
    - Difficult to obtain – EMA, X-rays, MRI.
    - Impractical for real time ASR.



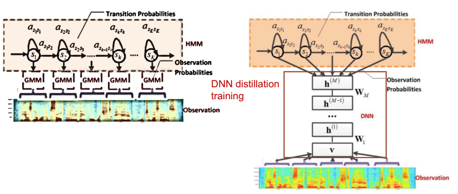


Articulatory and Spectrum Feature Integration

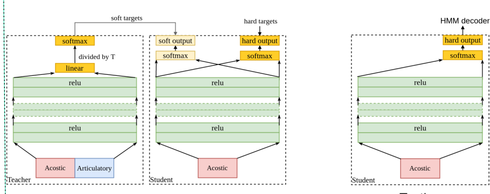
- **Feature based.**
  - Articulatory Inversion.
  - Most popular approach.
  - Mapping with various regression techniques.
- **Model based.**
  - More difficult.
  - HMM/BN (Markov, 2006)
  - Generalized Distillation (this work).



GMM-HMM versus DNN-HMM AMs



DNN Distillation Training and Testing

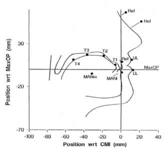


Training　　　Testing

共同研究者：Konstantin Markov (会津大学)

Experiments

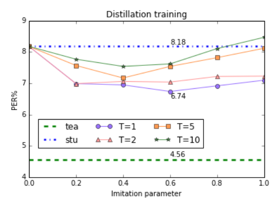
- Database.
  - University of Wisconsin X-ray micro-beam database (XRMB).
  - Consists of **simultaneously** recorded acoustic and articulatory measurements from 47 American English speakers.



- Features
  - Acoustic – MFCC (39 dim.)
  - Articulatory – Displacement of 8 articulatory points (16 dim.)
  - All feature vectors normalized and synchronized.
- Training procedure
  1. Train conventional GMM-HMM model using both acoustic and articulatory features.
  2. Perform forced alignment to obtain DNN "hard" targets.
  3. Train teacher DNN using both acoustic and articulatory features.
  4. Train student DNN using acoustic features <u>only</u> and guided by the "teacher".
- Testing procedure
  1. Use student DNN with acoustic features <u>only</u> to obtain HMM state probabilities.
  2. Use standard HMM decoding (Viterbi) to obtain recognition result.
- Evaluation metric – Phoneme Error Rate (PER)

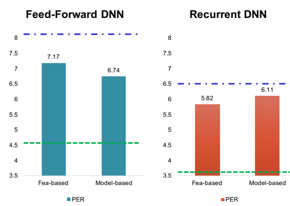
Results

- DNN parameters:
  - Feed Forward.
  - Input widow – 17 frames.
  - Activation – ReLU.
  - Dropout – 40%.
  - Teacher DNN
    - 5 layers
    - 3073 nodes.
  - Student DNN
    - 4 layers
    - 2048 nodes.



By DNN type:
Feed Forward
vs.
Recurrent

By integration type:
Feature based
vs.
Model based



Conclusions

- **Generalized distillation:**
  - Is an effective method for model based integration of information unavailable at testing time.
  - Allows smaller student models (4 layers / 2048 nodes) to reach performance close to bigger teacher models (5 layers / 3072 nodes).
- **DNN structure:**
  - Recurrent DNNs outperform Feed-Forward DNNs in the ASR task since they better model long-term temporal dependencies.
  - Time complexity of Recurrent DNNs is higher than Feed-Forward DNNs.
- **Integration approach:**
  - Model based and Feature based integration achieve comparable results.
  - Feature based integration requires higher computational power.