人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-043-01 (11/12)

音声・画像・映像における Deep Learning を用いたパターン認識

Pattern Recognition using Deep Learning for Speech, Image and Video

篠田浩一
Koichi SHINODA
東京工業大学
Tokyo Institute of Technology
shinoda@cs.titech.ac.jp

Abstract

近年、マルチメディア分野では、Deep Learning(深層学習) が盛んに研究されている。特に、音声認識や画像における一般物体認識では、従来法から大幅にエラーを削減し、すでに標準的な技術として商用にも使われている。本稿では、まず、マルチメディア分野における深層学習のこれまでの研究を概観した上で、現段階における課題とそれに対するアプローチを解説する。研究の進展は急であり、そろそろできることとできないことがはっきりしてきた。最後に、今後、深層学習を用いたパターン認識の研究がどのような方向に進んでいくかを議論したい。

Empirical evidence: Summary

(Dahl, Yu, Deng, Acero 2012, Seide, Li, Yu 2011 + new result)

- Voice Search SER (24 hours training)

AM	Setup	Test
GMM-HMM	MPE	36.2%
DNN-HMM	5 layers x 2048	30.1% (-17%)

- Switch Board WER (309 hours training)

AM	Setup	Hub5'00-SWB	RT03S-FSH
GMM-HMM	BMMI (9K 40-mixture)	23.6%	27.4%
DNN-HMM	7 x 2048	15.8% (-33%)	18.5% (-33%)

- Switch Board WER (2000 hours training)

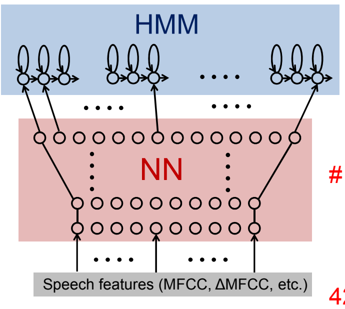
AM	Setup	Hub5'00-SWB	RT03S-FSH
GMM-HMM (A)	BMMI (18K 72-mixture)	21.7%	23.0%
GMM-HMM (B)	BMMI + fMPE	19.6%	20.5%
DNN-HMM	7 x 3076	14.4% (A: -34% B: -27%)	15.6% (A: -32% B: -24%)

(Dong Yu, 2012)

Neural network based speech recognition

1989: Time-Delay Neural Network (TDNN)
1994: Hybrid approach of NN and HMM
2000: Tandem connectionist features
2009: DNN phone recognition
2010: Recurrent NN (RNN) for language model
2011: DNN for LVCSR
(large vocabulary continuous speech recognition)
← The same as Hybrid approach (1994)

1994: Hybrid approach of NN and HMM



Use DNN outputs as HMM output prob.

$$p(x|c) = \frac{p(c|x)p(x)}{p(c)}$$

x: data, c: class

# of nodes:

132 x 50 x 5 (1994)

429 x 2048^5 x 761 (2011)

Bourlard and Morgan, "Connectionist Speech Recognition: A Hybrid Approach", The Springer International Series in Engineering and Computer Science, vol. 247, 1994

Replace GMM with DNN

- GMM (Gaussian Mixture Model) is mixture of experts (MoE), DNN is product of experts (PoE).
  - For GMM, it is difficult deal with multiple events in one window
  - GMM parameter estimation is easier to be parallelized
- DNN can get more info from multiple frames
  - GMM often use diagonal covariance and ignore correlation among them

Hinton et al., "Deep neural networks for acoustic modeling in speech recognition", IEEE Signal Processing Magazine, Nov. 2012.

Deep Learning (DL) in ICASSP2014

Already *de facto* standard

- 84 of 304 (28%) papers deals with DL
- Four sessions titled "DL" or "NN"
- DL penetrates into most speech sub-areas

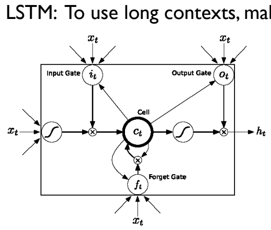
  Robustness (14), ASR systems (8), Features (7), Language model (5), Speaker recognition (5), Spoken term detection (3), Speech understanding (2), Emotion recognition (2)....

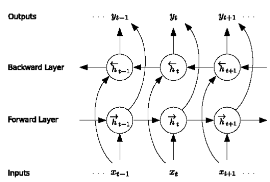
These trends continued in ICASSP2015

For high accuracy:

LSTM+Bi-directional RNN

Use LSTM (long-short-term memory) in RNN (Recurrent NN)
RNN: Effectively represents time sequence data
Bidirectional: Use info not only past but also future
LSTM: To use long contexts, make a cell which consists of 4 nodes





Graves et al., "Speech recognition with deep recurrent networks", ICASSP 2013.

For data sparsity:

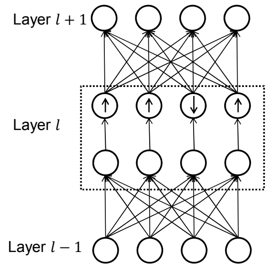
Speaker adaptation

To avoid overtraining, utilize prior knowledge about speakers

1. Regularization in parameter estimation (Bayesian approach)
2. Linear combination of speaker-cluster NNs
3. Add "speaker code" to NN inputs
4. Estimate activation function parameters

For data sparsity:

Estimate a new parameter of each node



Output of layer l

$$h^l = a(r^l) \circ \phi(W^{l\mathrm{T}} h^{l-1})$$

$\circ$: Element-wise multiplication

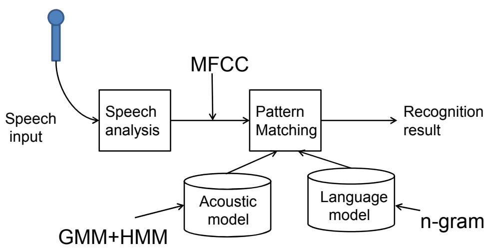
$a(r^l)$: Estimate for each speaker

free parameters $\simeq$ # nodes

P. Swietojanski and S. Renals, "Learning hidden unit contribution for unsupervised speaker adaptation of neural network acoustic models", IEEE SLT 2014.

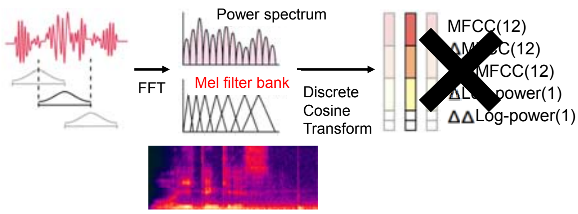
For end-to-end

Speech Recognition System - Before



For end-to-end

MFCC is no more needed



Mel filter bank features reduced 5-10% errors from MFCCs

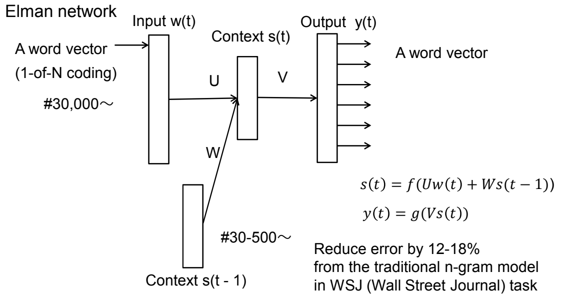
- MFCC was used to de-correlate the Mel filter bank features
- In DNN, such de-correlation process is not needed

Mohamed et al. "Acoustic modeling using deep belief network", IEEE Trans. ASLP, vol. 20, no. 1, 2012.

For end-to-end

2010: Recurrent NN for language model

Elman network



$$s(t) = f(Uw(t) + Ws(t-1))$$

$$y(t) = g(Vs(t))$$

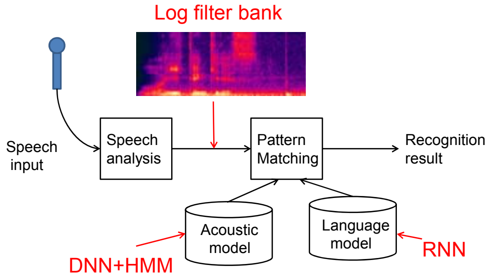
Reduce error by 12-18% from the traditional n-gram model in WSJ (Wall Street Journal) task

Mikolov et al. "Recurrent neural network based language model", INTERSPEECH2010

For end-to-end

Speech Recognition System - After

Log filter bank

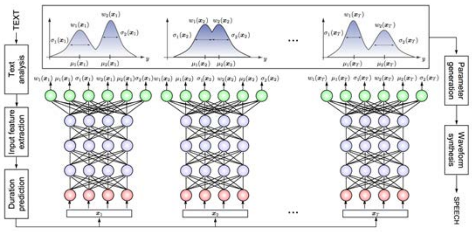


Mohamed et al. "Acoustic modeling using deep belief network", IEEE Trans. ASLP, vol. 20, no. 1, 2012.
Arisoy et al. "Deep neural network language models", NAACL-HLT 2012 workshop

Various applications

DNN for speech synthesis

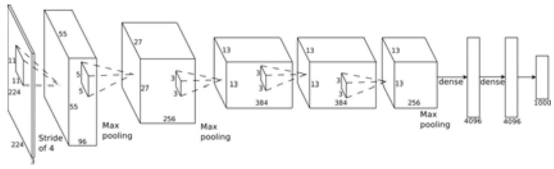
- Use DNN in reverse - input: label, output: data
- Output GMM parameters, mean and variance



Zen et al., Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis", ICASSP2014

ImageNet Challenge: ILSVRC 2012

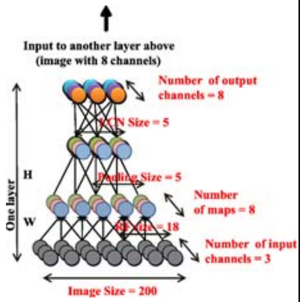
- Detect images of 1000 categories
- 1.2 million training samples
- Error 16% !



Krizhevsky et al. "ImageNet Classification with Deep Convolutional Neural Networks", NIPS2012.

Cat  Human face 

- Unsupervised learning
- 10 billion images from YouTube videos, each 200x200 pixels
- Sparse autoencoder with 9 layers, 1 billion nodes



Le et al. "Building high-level features using large scale unsupervised learning", ICML2012

TRECVID

(TREC Video Retrieval Evaluation)

Spinned out from Text REtrieval Conference (TREC) in 2001,
Organized by NIST(National Institute of Standard and Technology)
Aim : Promote research on video contents analysis and search
International, Competitive, Closed
Homepage: http://trecvid.nist.gov

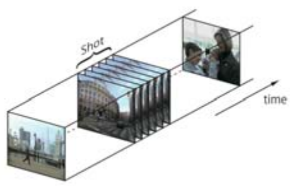
TokyoTech participated from 2006 (9 years)

2014 TRECVID task

- Semantic INdexing (SIN)
  Detect generic objects, scenes, actions
- Surveillance Event Detection (SED)
  Detect specific actions from surveillance video
- INstance Search (INS)
  Given a still image of an object, search video clips including it
- Multimedia Event Detection (MED)
  Detect complex "event"
- Multimedia Event Recounting (MER) (Pilot)
  Explain "event" detected

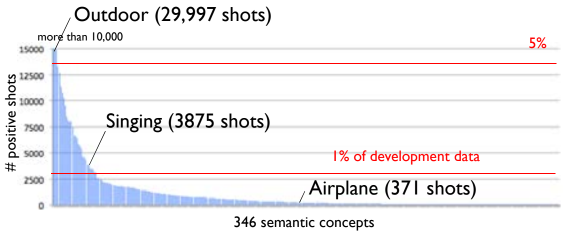
Semantic Indexing

Detect concepts from a set of video shots
Shot: The minimum unit of video
No. Concepts: 60
Training set: 549,434 shots, 800 hours
Test set: 332,751 shots, 400 hours



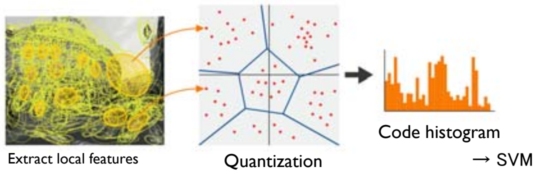
Frequency of Appearance (2011 task)

Number of positive samples in 264,673 training video shots



Bag of Visual Words

1. Quantize local features (e.g., SIFT) by using a codebook
   (Code word: Visual Word)
2. Use a code histogram as an input to SVM



Quantization Error!

Recent Trend

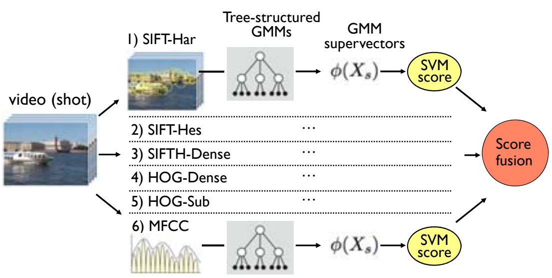
Tackle the data sparseness problem

- More features
  SIFT, Color SIFT, SURF, HOG, GIST, Dense features
- Multi-modal
  Use Audio : Singing, Dance, Car, etc.
- Multi-frame
  Not only key frames
- Soft clustering
  Reduce quantization errors. GMM etc.

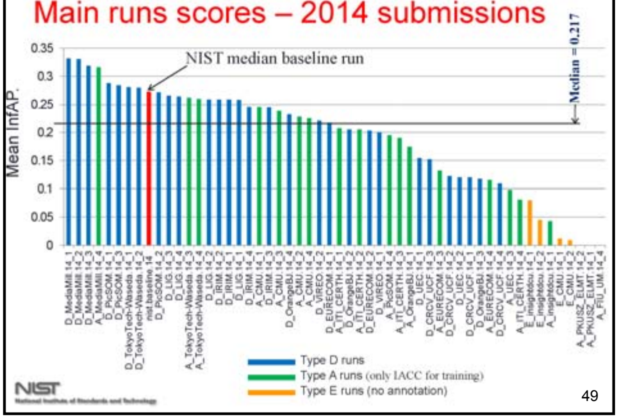
Less effective than expected

- Global features such as color histogram
  Local features are enough (no complementary info)
- Speech recognition, OCR
  Do not have performance high enough to contribute
- Object location
  Fail to detect. Many concepts do not have "location"
- Context between concepts
  Too Little data

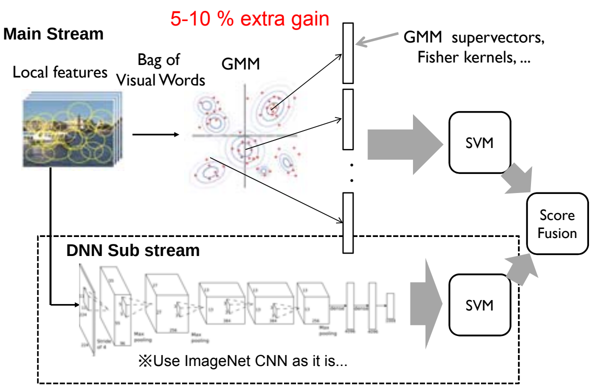
TokyoTech Framework



Main runs scores – 2014 submissions



Deep Learning



BoF is also deep learning!

Fisher Kernel based method is 5-layer DNN

stage	operation	type
SVM prediction	sign $f(X)$ $f(X) = \langle w, \phi(X) \rangle$	non-linear linear
per image vector, $\psi(X)$	square root, normalize (3) compute average of $\psi(x_i)$	non-linear linear
per descriptor vector, $\psi(x_i)$	multiply by γ_k in (1)/(2) bracket $(\cdot)$ in (1)/(2)	non-linear linear
preprocessing	L^2-normalization PCA projection	non-linear linear
SIFT	local pooling gradient filter	non-linear linear
image (as multiple overlapping regions)		

Table 1. Schematic description of a Fisher kernel SVM as a 5-layer feed-forward architecture (from bottom to top).

Sydorov et al., "Deep Fisher Kernels. End to End Learning of the Fisher Kernel GMM Parameters", CVPR2014

TRECVID Multimedia Event Detection (MED) task

- Extract "complex event" from many video clips (shot sequences)
  e.g. "Batting a run in", "Making a cake"
- Database : Home video 2000 hours
- Sponsored by IAPRA (The Intelligence Advanced Research Projects Activity)

Deep Learning at present

- Can be better than human in "well-defined" tasks with large data

MED task

- Multimedia
  Visual features, audio features, speech recognition, OCR
- Dynamic nature
- Training data for each event may be very small

Problems of Deep Learning

- How to deal with more complex problems such as MED?
- Only for "end-to-end" problems
  - Do we really need to solve them?
  - What is "semantics"?
- How to combine many modes in multimedia application
  - Combinatorial explosion
  - Time sequence

What we can do...

- Time Sequence
- Segmentation and Recognition
- Signal and symbol processing

Summary

- Deep learning is already de-facto in speech recognition
- Now, we are busy with replace "traditional" units by "DNN" units in a speech recognition system
  - What I explained today is only a small part of them
- Still ad-hoc, not enough theoretical background
  - How to optimize structures?
  - Why is Deep learning better?
  - How to combine acoustic and language models?

Speech is "lighter" compared with the other media.
Good test bed for exploring Deep learning!