Deep Activation Mixture Model for Speech Recognition

C. Wu and M.J.F. Gales {cw564,mjfg}@eng.cam.ac.uk

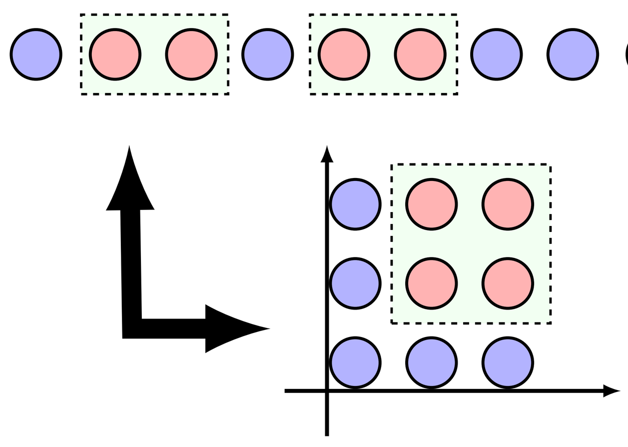
UNIVERSITY OF CAMBRIDGE

Department of Engineering, University of Cambridge

1. Introduction

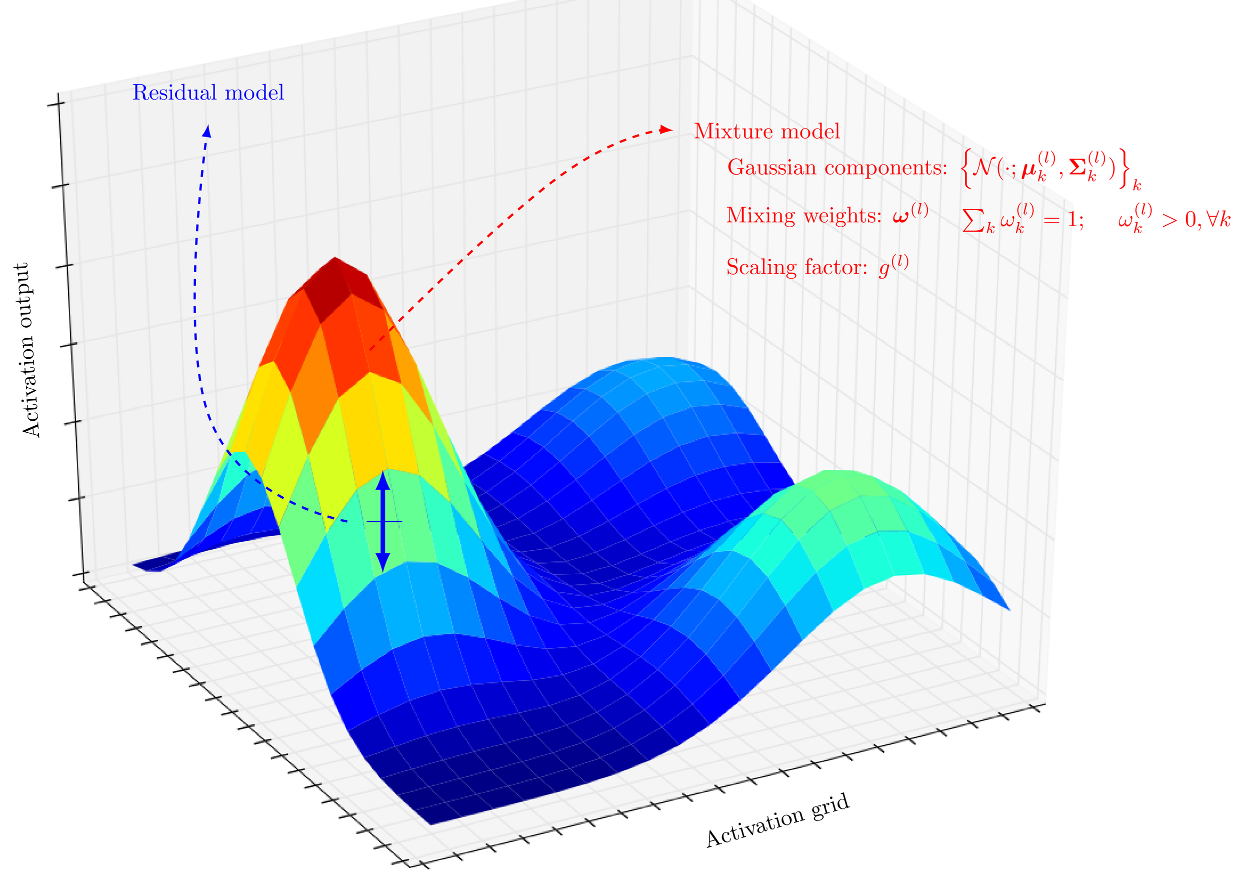
- DNNs are commonly treated as "black-box" models
  - Hard to interpret and group DNN parameters
  - Make regularisation and adaptation challenging
- Deep activation mixture models (DAMMs) are proposed to
  - Improve network regularisation and interpretation
  - Allow novel network adaptation schemes

2. Hidden Unit Reorganisation



- Reorganise units of each hidden layer to form a grid
  - Avoid the arbitrary ordering of hidden units
  - Enable activation functions to be related in regions of the network
- Each unit i is represented as a point $\boldsymbol{s}_i$ in the grid space

3. Network Topology



- How to relate different activation funcitons?
$$\boldsymbol{h}^{(l)} = \boldsymbol{h}^{(l)}_{\text{mix}} + \boldsymbol{h}^{(l)}_{\text{res}}$$
- Mixture model — activation function contour
$$h^{(l)}_{\text{mix},i} = g^{(l)} \sum_{k=1}^{K} \omega_k^{(l)} \mathcal{N}\left(\boldsymbol{s}_i; \boldsymbol{\mu}_k^{(l)}, \boldsymbol{\Sigma}_k^{(l)}\right)$$
  - Dynamic scaling factor $g^{(l)}$ and mixing weights $\boldsymbol{\omega}^{(l)}$
$$g^{(l)} \text{ — sigmoid} \qquad \boldsymbol{\omega}^{(l)} \text{ — softmax}$$
  - Interpret Gaussian component as phoneme or ...
    - E.g. set $\{\boldsymbol{\mu}_k^{(l)}\}$ as 2D projection of phoneme average features
  - Perform adaptation on Gaussian mean vector and covariance matrix
- Residual model — fluctuations on contour
$$\boldsymbol{h}^{(l)}_{\text{res}} = \tanh\left(\boldsymbol{W}^{(l)\text{T}}\boldsymbol{h}^{(l-1)} + \boldsymbol{c}^{(l)}\right)$$
  - Add small "noise" to GMM contour

4. Training

- Highly restricted mixture model *v.s.* powerful residual model
- How to train mixture model to the maximal extent?
$$\boldsymbol{h}^{(l)} \simeq \boldsymbol{h}^{(l)}_{\text{mix}} \quad \Rightarrow \quad \boldsymbol{h}^{(l)}_{\text{res}} \simeq \boldsymbol{0}$$
  - Training criterion with residual-model regularisation
$$\mathcal{F}(\boldsymbol{\theta}_{\text{mix}}, \boldsymbol{\theta}_{\text{res}}) = \mathcal{L}(\boldsymbol{\theta}_{\text{mix}}, \boldsymbol{\theta}_{\text{res}}) + \eta||\boldsymbol{\theta}_{\text{res}}||_2^2$$
  - Isolating training mode

Algorithm 1 Isolating Training Mode of DAMM.

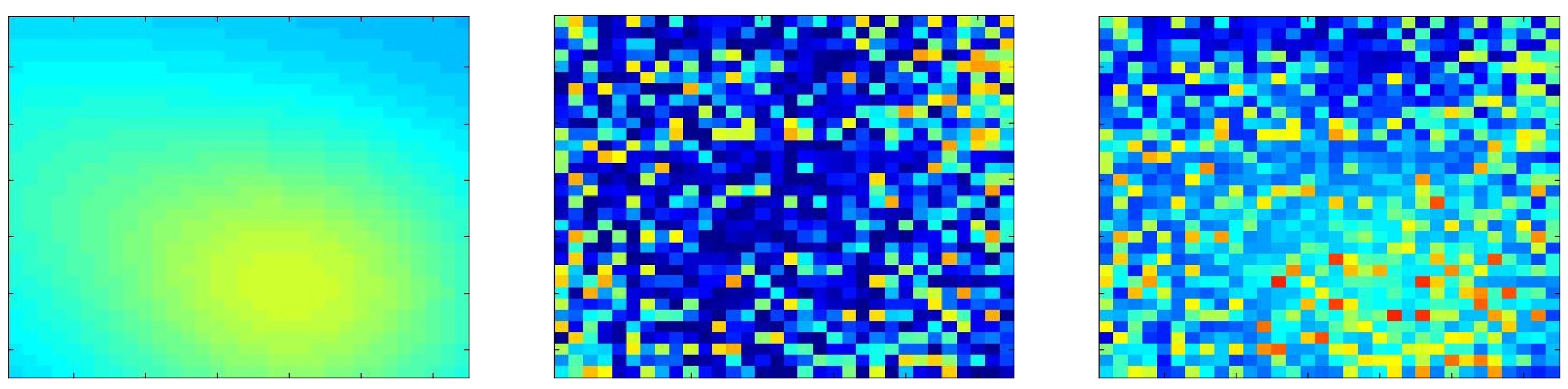
```
for l := 1 to L do
    initialise θ_res^(l) = 0, θ_mix^(l)
    update θ_mix
    update θ_res
end for
finetune θ_res
```

5. Adaptation

- Adaptation on Gaussian components of mixture model
$$h^{(ls)}_{\text{mix},i} = g^{(l)} \sum_{k=1}^{K} \omega_k^{(l)} \mathcal{N}\left(\boldsymbol{s}_i; \boldsymbol{\mu}_k^{(ls)}, \boldsymbol{\Sigma}_k^{(ls)}\right)$$
  - A compact set of parameters to adapt $\{\boldsymbol{\mu}_k^{(ls)}, \boldsymbol{\Sigma}_k^{(ls)}\}_{1\le l\le L}$
  - Allow robust and rapid adaptation
- $\boldsymbol{\Sigma}_k^{(l)}$ can be rewritten using unit variance and correlation coefficient
$$\boldsymbol{\Sigma}_k^{(l)} = \begin{vmatrix} \sigma_{k1}^{(l)2} & \rho_k^{(l)}\sigma_{k1}^{(l)}\sigma_{k2}^{(l)} \\ \rho_k^{(l)}\sigma_{k1}^{(l)}\sigma_{k2}^{(l)} & \sigma_{k2}^{(l)2} \end{vmatrix}$$
  - Parametrise $\boldsymbol{\sigma}_k^{(l)}$ and $\rho_k^{(l)}$ to satisfy the positive-definite constraint
$$\boldsymbol{\sigma}_k^{(l)} = \textbf{exp}\left(\tilde{\boldsymbol{\sigma}}_k^{(l)}\right), \ \rho_k^{(l)} = \tanh\left(\tilde{\rho}_k^{(l)}\right)$$

6. Experiment

- Data and setup
  - 144-hour English broadcast news dataset (`LDC97S44`, `LDC98S71`)
  - DNN-HMM hybrid ASR framework
  - 5 hidden layers with 1024 units for both DNN and DAMM systems
  - 46 Gaussian components to form DAMM mixture models
  - Unsupervised utterance-level adaptation
- Grid-output example of mixture and residual models



(a) Mixture (b) Residual (c) Mixture+residual

- CE performance comparison [WER (%)]

System	Dev	Eval
DNN(tanh)	12.8	11.0
DAMM	**12.3**	**10.6**
DNN(sigmoid)	12.4	10.8

- Adaptation of mean, variance and correlation coefficient on CE DAMM

System	Adapt			Dev	Eval
	mean	variance	correlation		
SI	✗	✗	✗	12.3	10.6
SD	✓	✗	✗	12.2	10.6
	✗	✓	✓	12.1	10.5
	✓	✓	✓	**12.0**	**10.4**

  - More effective to adapt covariance matrix than mean vector
- MPE performance comparison

System	Adapt			Dev	Eval
	mean	variance	correlation		
DNN(sigmoid)	–			11.4	10.1
DAMM	✗	✗	✗	11.4	10.0
	✓	✓	✓	**11.1**	**9.8**

- Up to 3% rel. WER reduction by adaptation

7. Conclusions

- Propose DAMM for network regularisation and adaptation
- Extend L2 regularisation to approach a dynamic surface, not zero
- Novel adaptation scheme to modify the dynamic surface