Simplifying very deep convolutional neural network architectures for robust speech recognition

Joanna Rownicka, Steve Renals, Peter Bell

The Centre for Speech Technology Research, University of Edinburgh, United Kingdom

THE UNIVERSITY of EDINBURGH

Objectives

- analyse to what extent 2D convolutions are suitable for robust speech recognition task;
- determine which components of very deep CNN (VDCNN) models are necessary to achieve state-of-the-art results.

Very Deep CNNs

- VDCNNs ⇒ designed for computer vision;
- recently applied also to ASR and other sequence to sequence tasks;
- parameter sharing causes the convolutional layer to have the **equivariance to translation** property;
  - convolutions in time ⇒ equivariance to shifts in time;
  - convolutions in frequency ⇒ equivariance to shifts in frequency;
    - the same word at a different pitch can produce the same representation;
    - if the distortion is more apparent in some bands of the spectrum than in others, representations can be computed from the cleaner parts of the spectrum;
- we simplify the VDCNN models for noise robust speech recognition in terms of layers diversification;
  - we experiment with downsampling and fully-connected layers.

Experimental setup

- static FBANK 11 x 40 feature map input
- minibatch SGD using Adam
- "Xavier" initialization for the weights
- batch normalization

Datasets

- Aurora4 training set (15 h each)
  - clean-condition → equivalent to the SI-84 WSJ
  - multi-condition → clean-condition training set recorded with a mismatched microphone and corrupted using six noise types at different SNR levels
- Aurora4 test sets (9 h)
  - A → clean
  - B → with 6 types of additive noise
  - C → recorded with a mismatched microphone
  - D → with 6 types of additive noise and recorded with a mismatched microphone
- MGB-3 training set → 750 episodes (about 350 hours)
- MGB-3 test sets
  - MGB-3 dev set (5 h)
  - MGB-1 test set (19 h)

Results: Aurora4

Model	A	B	C	D	AVG
DNN	3.47	7.67	7.85	19.73	12.55
CNN	3.33	6.89	6.59	17.92	11.34
VDCNN-max-4FC	2.43	5.92	5.74	16.26	10.09
VDCNN-avg	2.75	5.88	7.27	16.46	10.29
VDCNN-max	2.56	5.78	5.36	15.40	9.64
VDCNN-max-addconv	2.50	6.02	6.95	16.35	10.26
VDCNN-allconv	2.32	5.45	5.38	15.56	**9.55**

WERs [%] for the baseline models (DNN, CNN, VDCNN-max-4FC) and our VDCNN models (avg, max, max-addconv, allconv) trained with alignments generated from multi-condition training set of Aurora4.

Model	A	B	C	D	AVG
DNN/clnali	3.19	6.42	7.04	17.04	10.79
VDCNN-max-4FC/clnali	2.54	5.33	4.61	13.77	8.70
VDCNN-allconv/clnali	2.43	4.43	4.50	12.50	**7.75**

WERs [%] for different models trained with alignments generated from synchronized clean-condition training set of Aurora4.

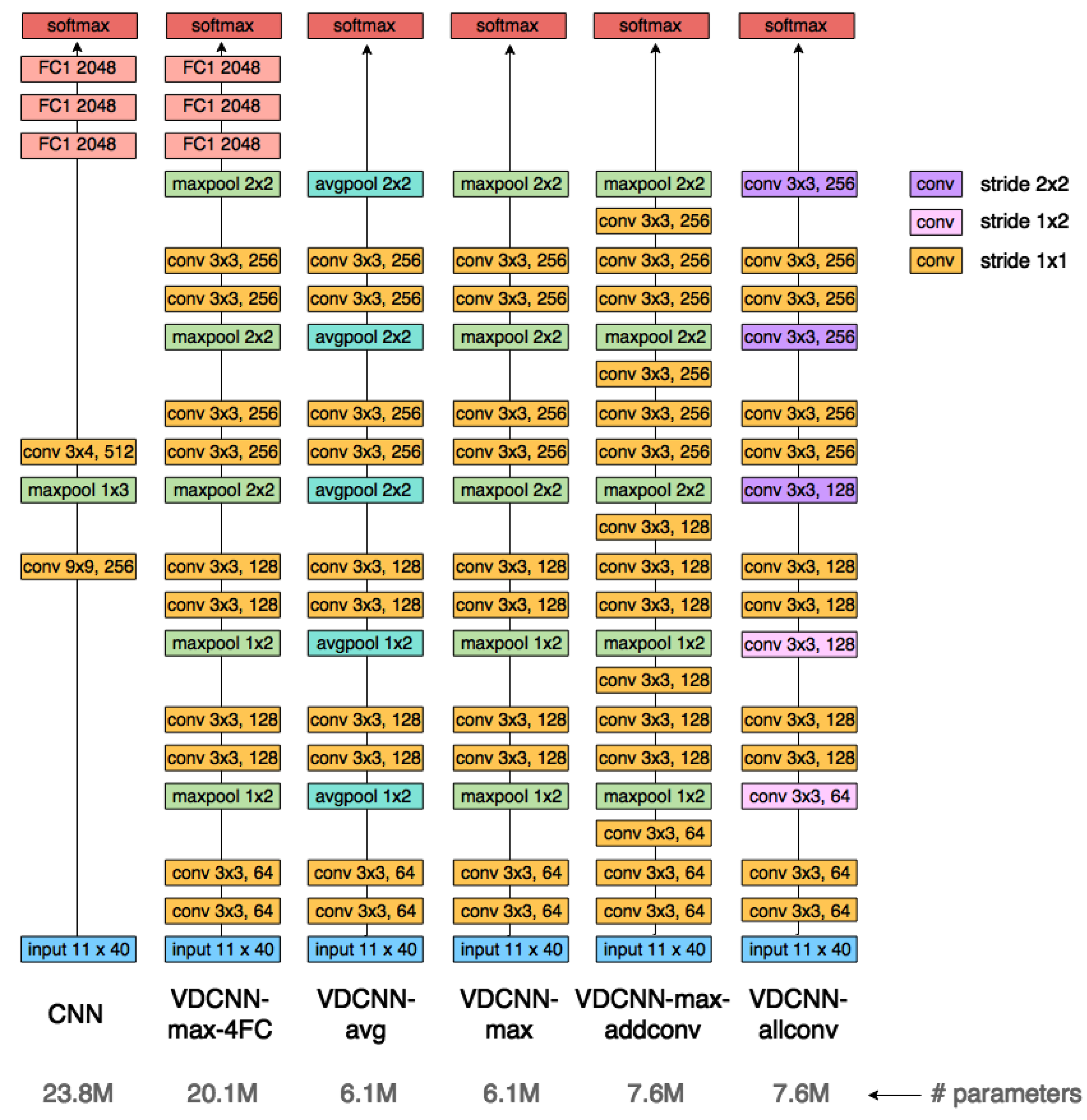
Model	A	B	C	D	AVG
DNN/clntr	2.71	43.00	24.06	58.66	45.48
VDCNN-max-4FC/clntr	2.32	35.99	21.20	52.53	39.62
VDCNN-allconv/clntr	1.98	40.62	20.08	55.57	42.80

WERs [%] for the models trained with clean-condition training set.

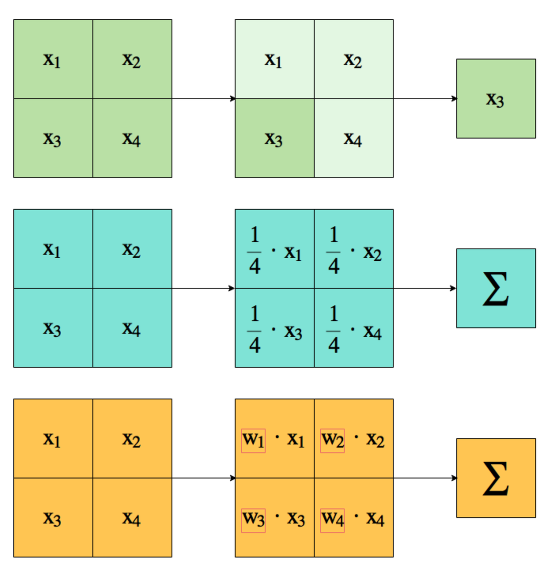
Most performance gains per test set over the baseline

Best model on average

Architectures



Pooling



Using convolutional layers instead of pooling layers to downsample feature maps ⇒ learning the pooling operation rather than fixing it.

Results: MGB-3

Model	WER	
	MGB3-dev	MGB1-test
VDCNN-max-4FC	53.2	42.4
VDCNN-avg	52.4	41.4
VDCNN-max	52.5	41.4
VDCNN-allconv	52.2	41.2
VDCNN-allconv-4G	**50.0**	**38.7**

WERs [%] for MGB-3 dev set and MGB-1 test set for the baseline model (VDCNN-max-4FC) and our VDCNN models (avg, max, allconv). The last VDCNN-allconv-4G model is rescored with a 4-gram LM. All models were trained on MGB-3 training data.

Conclusions

- Pooling layers are not necessary to achieve state-of-the-art results for speech recognition with VDCNNs.
  - Using **conv layers with increased stride** can effectively enable the model to learn the necessary invariances.
- **Removing fully-connected layers** from a VDCNN architecture contributed most to the performance gains in our experiments, especially for noisy test data.
- Our model consisting solely of fifteen 2D conv layers with the same kernels sizes throughout the network and a single softmax classification layer gives the best performance consistently on the Aurora4 and MGB tasks.

Acknowledgements

- This work was supported by a PhD studentship supported by The DataLab Innovation Centre, Ericsson Media Services, and Quorate Technology.

[1] J. T. Springenberg, A. Dosovitskiy, T. Brox, M. Riedmiller, "Striving for Simplicity: The All Convolutional Net," ICLR 2014.
[2] Y. Qian, M. Bi, T. Tan and K. Yu, "Very Deep Convolutional Neural Networks for Noise Robust Speech Recognition," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 12, pp. 2263-2276, Dec. 2016.

Contact: j.m.rownicka@sms.ed.ac.uk