An Overview of Residual Networks

Presented by Peidong Wang

03/30/2017

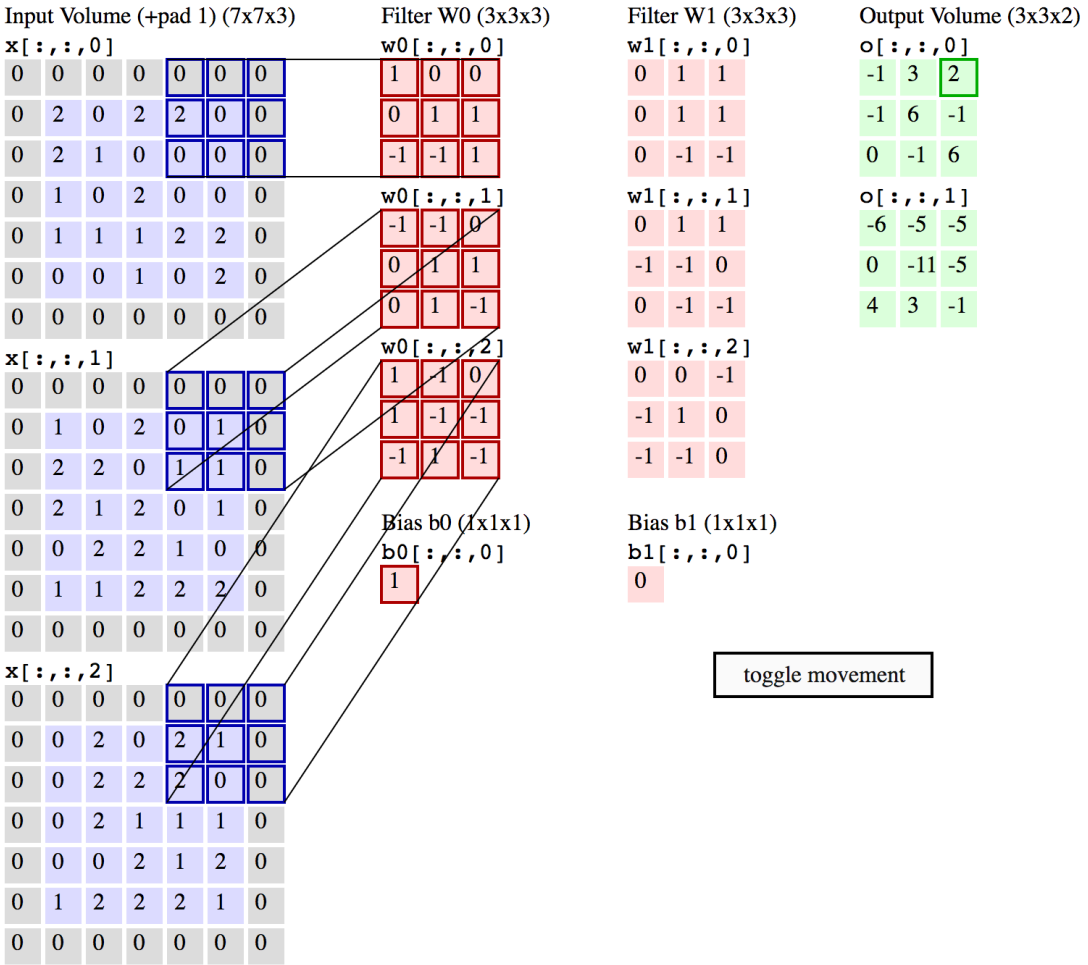
Content

- Review of Convolutional Neural Network (CNN)
- Residual Network
- Wide Residual Network
- Wide Residual Network for Robust Speech Recognition
- Summary

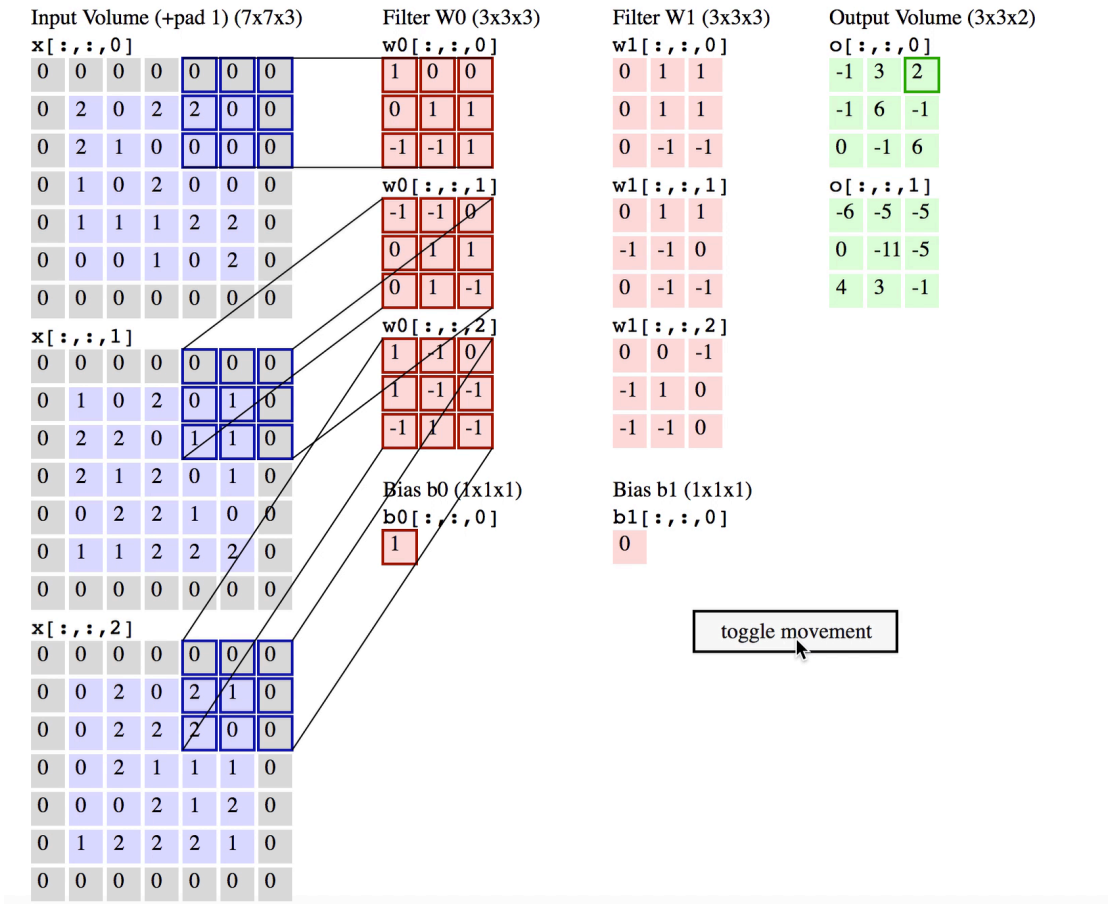
Content

- Review of Convolutional Neural Network (CNN)
  - "CS231n Convolutional Neural Networks for Visual Recognition", *http://cs231n.github.io/convolutional-networks/*, 2017.
- Residual Network
- Wide Residual Network
- Wide Residual Network for Robust Speech Recognition
- Summary

Review of Convolutional Neural Network



Review of Convolutional Neural Network



Review of Convolutional Neural Network

- Deep CNN
  - Evidence shows that network depth is of crucial importance, and the leading results on the challenging ImageNet dataset all exploit “very deep” models.
  - Driven by the significance of depth, a question arises: Is learning better networks as easy as stacking more layers?

Review of Convolutional Neural Network

- Problems of Deep CNN
  - The notorious problem of vanishing/exploding gradients
  - **Description:** the values of the back-propagated gradients gets smaller as it is propagated downwards
  - **Solutions**: normalized initialization and intermediate normalization layers

Review of Convolutional Neural Network

- Problems of Deep CNN (cont'd)
  - Degradation problem of Deep CNNs
  - **Description:** With the network depth increasing, accuracy gets saturated and then degrades rapidly. Such degradation is not caused by overfitting, and adding more layers to a suitably deep model leads to higher *training* error.
  - **Solution**: ?

Content

- Review of Convolutional Neural Network (CNN)
- Residual Network
  - He, Kaiming, et al. "Deep residual learning for image recognition." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.
- Wide Residual Network
- Wide Residual Network for Robust Speech Recognition
- Summary

Residual Network

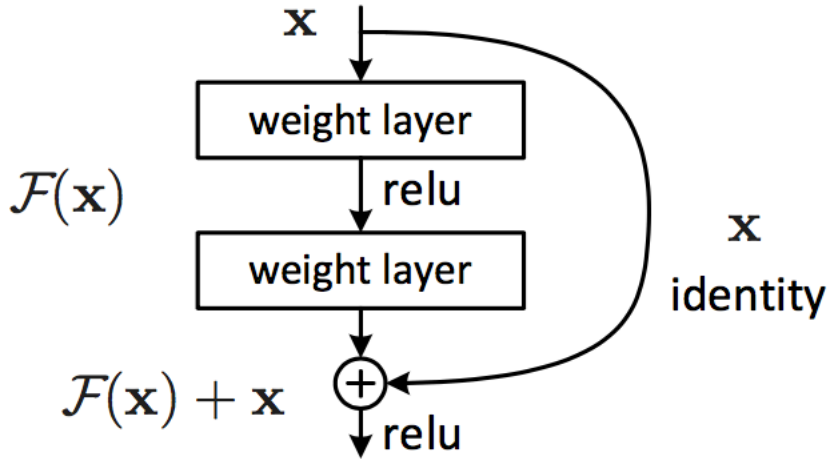
- Idea
  - The degradation indicates that not all systems are similarly easy to optimize.
  - Let us consider a shallower architecture and its deeper counterpart that adds more layers onto it. There exists a solution: the added layers are identity mapping, and the other layers are copied from the learned shallower model.
  - The existence of this constructed solution indicates that a deeper model should produce no higher training error than its shallower counterpart.

Residual Network

- Residual Network
  - Instead of hoping each few stacked layers directly fit a desired underlying mapping, we explicitly let these layers fit a residual mapping.
  - Formally, denoting the desired underlying mapping as $H(x)$, we let the stacked nonlinear layers fit another mapping of $F(x) := H(x) – x$. The original mapping is recast into $F(x) + x$.

Residual Network

- Residual Network (cont'd)
  - We hypothesize that it is easier to optimize the residual mapping than to optimize the original, unreferenced mapping. To the extreme, if an identity mapping were optimal, it would be easier to push the residual to zero than to fit an identity mapping by a stack of nonlinear layers.



Residual Network

- Experiments
  - Dataset: CIFAR-10
  - Dataset Details:
    - 32 x 32 color images
    - drawn from 10 classes
    - split into 50k training images and 10k testing images

- Experimental Results
  - Residual Network won the 1st place in ILSVRC 2015 (dataset: ImageNet) and COCO 2015

Residual Network

- Experimental Results (cont'd)

method			error (%)
Maxout [9]			9.38
NIN [25]			8.81
DSN [24]			8.22
	# layers	# params	
FitNet [34]	19	2.5M	8.39
Highway [41, 42]	19	2.3M	7.54 (7.72±0.16)
Highway [41, 42]	32	1.25M	8.80
ResNet	20	0.27M	8.75
ResNet	32	0.46M	7.51
ResNet	44	0.66M	7.17
ResNet	56	0.85M	6.97
ResNet	110	1.7M	**6.43** (6.61±0.16)
ResNet	1202	19.4M	7.93

Table 6. Classification error on the **CIFAR-10** test set. All methods are with data augmentation. For ResNet-110, we run it 5 times and show "best (mean±std)" as in [42].

Residual Network

- Problem of Residual Network
  - Diminishing feature reuse
  - **Description:** As gradient flows through the network, there is nothing to force it to go through residual block weights and it can avoid learning anything during training, so it is possible that there is either only a few blocks that learn useful representations, or many blocks share very little information with small contribution to the final goal.
  - **Solution:** ?

Content

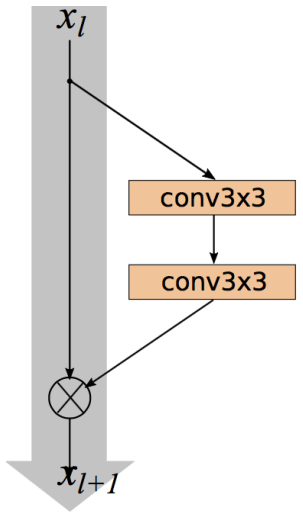
- Review of Convolutional Neural Network (CNN)
- Residual Network
- Wide Residual Network
  - Zagoruyko, Sergey, and Nikos Komodakis. "Wide residual networks." *arXiv preprint arXiv:1605.07146* (2016).
- Wide Residual Network for Robust Speech Recognition
- Summary

Wide Residual Network

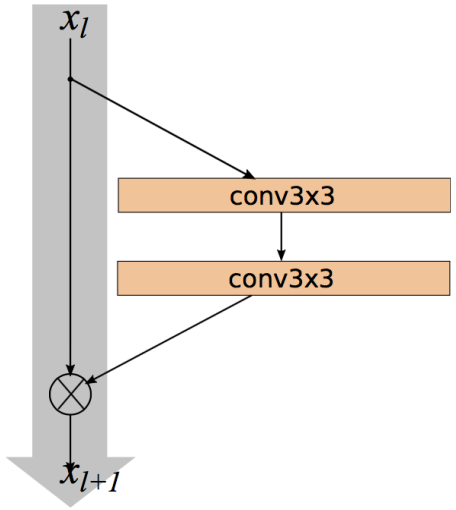
- Idea
  - To reduce depth and increase the representation power of residual blocks, there are three simple ways.
    - To add more convolutional layers per block
    - To widen the convolutional layers by adding more feature planes
    - To increase filter sizes in convolutional layers

Wide Residual Network

- Wide Residual Network
  - “To widen the convolutional layers by adding more feature planes” can be interpreted as to increase the number of filters for each convolution layer. [*]



(a) basic



(c) basic-wide

[*] Added by the Presenter

Wide Residual Network

- Experiments
  - Dataset: CIFAR-10 and CIFAR-100
  - Dataset Details:
    - 32 x 32 color images
    - drawn from 10 classes and 100 classes, respectively
    - split into 50k training images and 10k testing images

Wide Residual Network

- Experimental Results:

	depth-k	# params	CIFAR-10	CIFAR-100
NIN [20]			8.81	35.67
DSN [19]			8.22	34.57
FitNet [24]			8.39	35.04
Highway [28]			7.72	32.39
ELU [5]			6.55	24.28
original-ResNet[11]	110	1.7M	6.43	25.16
	1202	10.2M	7.93	27.82
stoc-depth[14]	110	1.7M	5.23	24.58
	1202	10.2M	4.91	-
pre-act-ResNet[13]	110	1.7M	6.37	-
	164	1.7M	5.46	24.33
	1001	10.2M	4.92(4.64)	22.71
WRN (ours)	40-4	8.9M	4.53	21.18
	16-8	11.0M	4.27	20.43
	28-10	36.5M	**4.00**	**19.25**

Table 5: Test error of different methods on CIFAR-10 and CIFAR-100 with moderate data augmentation (flip/translation) and mean/std normalzation. We don't use dropout for these results. In the second column k is a widening factor. Results for [13] are shown with mini-batch size 128 (as ours), and 64 in parenthesis. Our results were obtained by computing median over 5 runs.

Content

- Review of Convolutional Neural Network (CNN)
- Residual Network
- Wide Residual Network
- Wide Residual Network for Robust Speech Recognition
  - Jahn Heymann, Lukas Drude, and Reinhold Haeb-Umbach. "Wide Residual BLSTM Network with Discriminative Speaker Adaptation for Robust Speech Recognition."
- Summary

Wide Residual Network for Robust ASR

- Wide Residual Network for Robust ASR
  - State-of-the-art model for CHiME4 single channel task with baseline language model
  - Combine wide residual network with Bidirectional LSTM (BLSTM)
  - Wide residual network is used to refine the feature, and BLSTMs are used to incorporate context information. [*]

[*] Added by the Presenter

Wide Residual Network for Robust ASR

- Wide Residual Network for Robust ASR (cont'd)

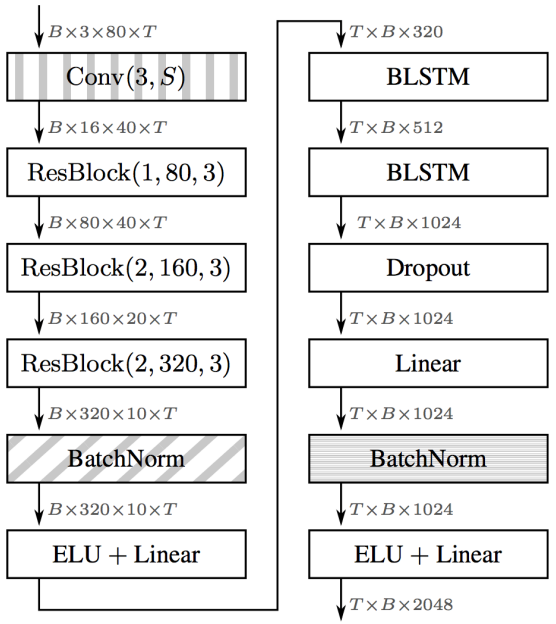


Figure 3: Overview of the back-end structure. The annotations in gray indicate the dimension of the tensors where B is the mini-batch size and T is the number of frames of the largest utterance within the batch. The building blocks are explained in Fig. 1 and Fig. 2. The convolution and the diagonally striped batch normalization is defined as in in Fig. 2. The horizontally striped batch normalization just collects statistics along the time frame axis.

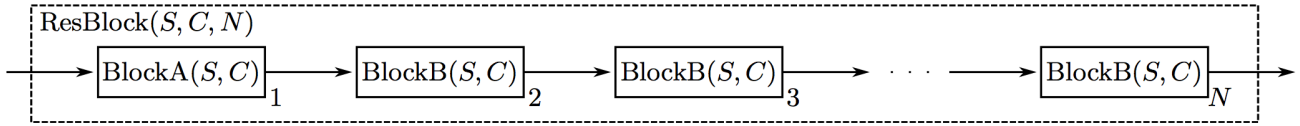


Figure 1: Detailed view of a ResBlock. A ResBlock is parameterized by its striding S, the number of output channels C and the number of inner blocks N. Accordingly, BlockB is repeated $N - 1$ times.

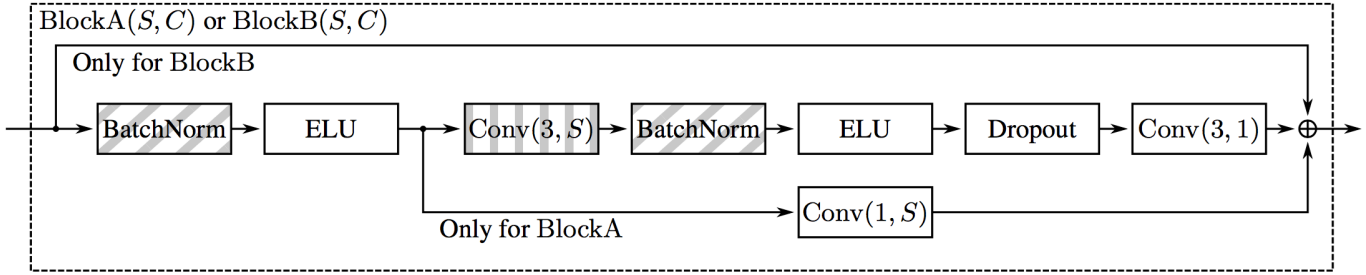


Figure 2: Detailed view of the building blocks BlockA or BlockB. The batch normalization collects statistics along the frequency band axis and along the time frame axis. A convolution block Conv(A, S) is parameterized by the filter size $A \times A$, the zero padding $(A - 1)/2$ in both directions and the consecutive striding S.

Wide Residual Network for Robust ASR

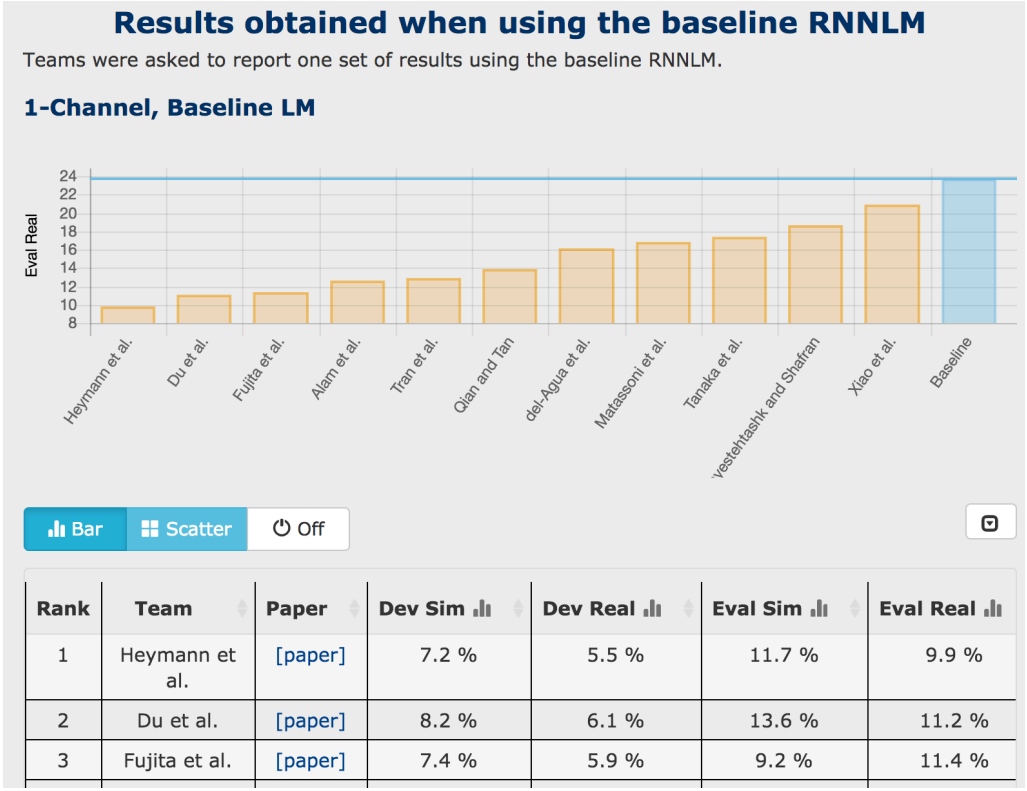
- Experiment
  - Dataset: CHiME4
  - Dataset Details:
    - Training set: 1600 (real) + 7138 (simulated) = 8738 noisy utterances
    - Development set: 410 (real) X 4 (environments) + 410 (simulated) X 4 (environments) = 3280 utterances
    - Test set: 330 (real) X 4 (environments) + 330 (simulated) X 4 (environments) = 2640 utterances
    - Environments: bus (BUS), cafe (CAF), pedestrian area (PED), and street junction (STR)

Wide Residual Network for Robust ASR

- Experimental Results

Table 2: Average WER (%) for the tested systems. Bold results correspond to the officially submitted results. The individual abbreviations mean: ”Kaldi”: baseline backend, ”WRBN”: our WRBN (Section 2.2.1), ”+BN”: with Batch-Normalization (Sec. 2.2.3), ”+SA”: with additional linear speaker adaptation layer (Sec. 2.2.4) ”+NTLM”: with own language model (Sec. 2.2.5), ”+GEV”: with GEV beamformer (Sec. 2.1), ”+BFIT”: with baseline front-end beamformer

Track	System	Dev		Test	
		real	simu	real	simu
1ch	Baseline	11.57	12.98	23.70	20.84
	WRBN	6.64	9.09	11.8	13.78
	+BN	5.69	7.53	10.4	12.67
	+SA	5.5	7.18	9.88	11.68
	+NTLM	**5.19**	**6.69**	**9.34**	**11.11**



Rank	Team	Paper	Dev Sim	Dev Real	Eval Sim	Eval Real
1	Heymann et al.	[paper]	7.2 %	5.5 %	11.7 %	9.9 %
2	Du et al.	[paper]	8.2 %	6.1 %	13.6 %	11.2 %
3	Fujita et al.	[paper]	7.4 %	5.9 %	9.2 %	11.4 %

Content

- Review of Convolutional Neural Network (CNN)
- Residual Network
- Wide Residual Network
- Wide Residual Network for Robust Speech Recognition
- Summary

Summary

- We reviewed the history of an important branch of CNN, Residual Networks.
- We learned not only the network frameworks but also the logic and ideas behind each progress.

Thank You!