Audio data augmentation

Лукьянов Павел Александрович
МГУ им. М. В. Ломоносова

27 октября 2020 г.

Image augmentation

Augmentation - method used to increase the amount of data by adding modified copies of already existing data or created synthetic data from existing data.

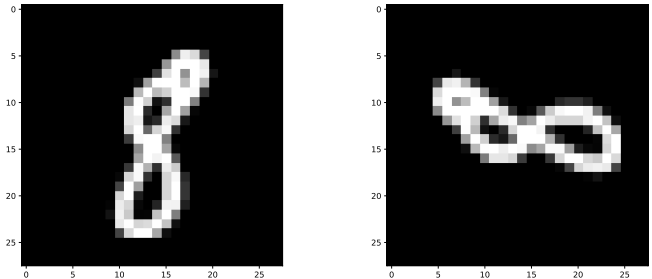
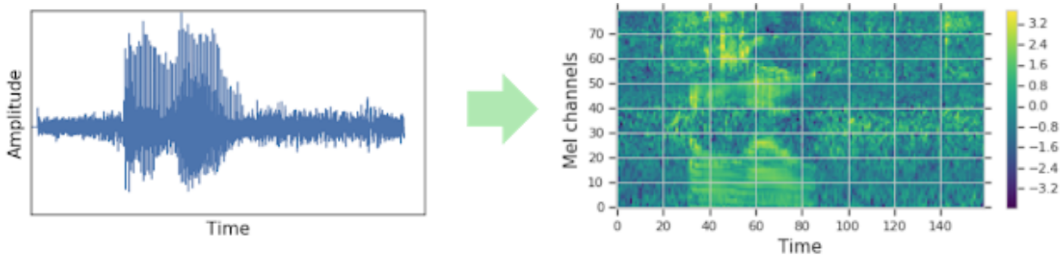


Рис. 1: Example of image augmentation

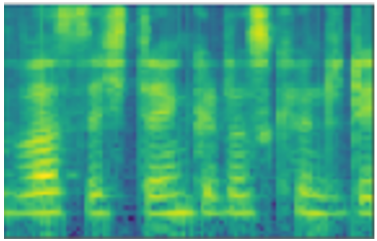
Audio in DL


Amplitude
Time
Mel channels
Time
0
20
40
60
80
100
120
140
0
10
20
30
40
50
60
70
3.2
2.4
1.6
0.8
0.0
−0.8
−1.6
−2.4
−3.2

Speed augmentation

point - $\mathcal{U}(0, T)$, window_length - $\mathcal{U}(0, a)$, speed - $\mathcal{U}(p_1, p_2)$

Original



Speed variation

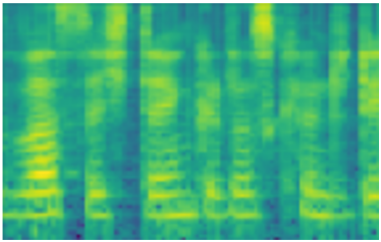
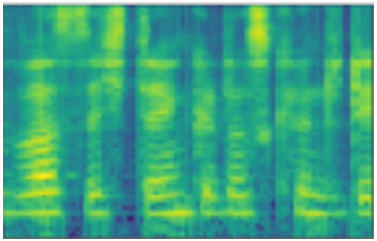


Рис. 2: Example of speed augmentation

Random erasing

Erasing a rectangle region of arbitrary size

Original



Random erasing

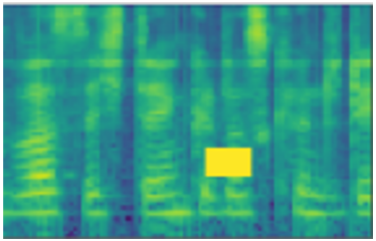
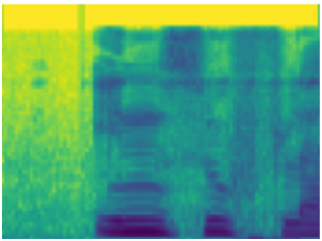


Рис. 3: Example of random erasing

Noise augmentation

Original



Noise augmentation

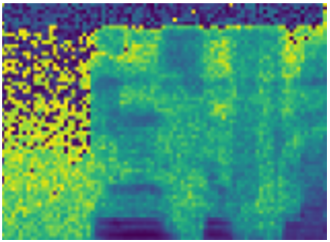


Рис. 4: Example of noise augmentation

Additional augmentations

- Loudness control
  value = minimum + (1 - λ) · (value - minimum)
- Shift augmentation
  shift - $\mathcal{U}(0, \alpha)$
  direction -$Bernoulli(0, 0.5)$

Comparison of augmentations

	ConvClassifier (CC)	SubSpectral Network (SSN)	SubSpectral Classifier (SSC)	Residual Conv-Classifier (RCC)	**Max.**
Raw	64.71 ± 0.30	63.54 ± 0.26	65.82 ± 0.39	64.53 ± 0.53	65.82
Speed	64.71 ± 0.40	63.54 ± 0.91	66.40 ± 0.34	64.07 ± 0.77	66.40
Noise	63.62 ± 0.22	61.35 ± 0.67	64.48 ± 0.46	63.38 ± 0.41	64.48
Loudness	64.31 ± 0.43	62.87 ± 0.64	65.65 ± 0.50	64.14 ± 0.37	65.65
Shift	65.40 ± 0.55	66.31 ± 0.70	$\mathbf{68.20} \pm 0.44$	64.38 ± 0.78	**68.20**
Masking	64.27 ± 0.34	62.27 ± 0.33	65.10 ± 0.32	64.30 ± 0.45	65.10
Combined	65.03 ± 0.41	63.49 ± 0.79	66.35 ± 0.34	64.12 ± 0.42	66.35
Max.	65.40	66.31	**68.20**	64.38	
Model Parameters	633,219	1,801,621	1,533,321	1,095,171	

Рис. 5: Model performance under the influence of a variety of data augmentation strategies

SpecAugment

3 kinds of deformations:

1. Time warping
2. Frequency masking
3. Time masking

Model	Clean (No LM)	Other (No LM)	Clean (LM)	Other (LM)
LAS	4.1	12.5	3.2	9.8
LAS + SpecAugment	2.8	6.8	2.5	5.8

Рис. 6: LibriSpeech 960h WERs (%)

Example of SpecAugment

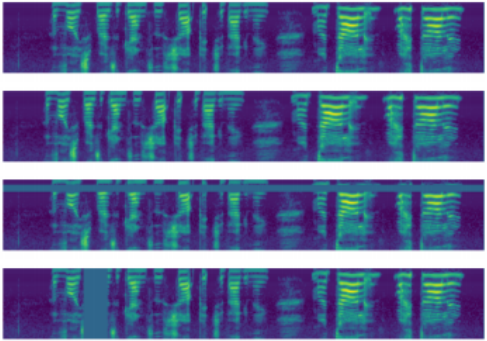


Рис. 7: Original, Time warping, Frequency masking, Time masking

Discussion

Time warping contributes, but is not a major factor in improving performance

Augmentations	test
All 3 augmentations	3.7
Without Time warping	3.8
Without Frequency masking	4.0
Without Time masking	4.1

Рис. 8: WER without LM (%)

Discussions

Augmentation converts an over-fitting problem into an under-fitting problem

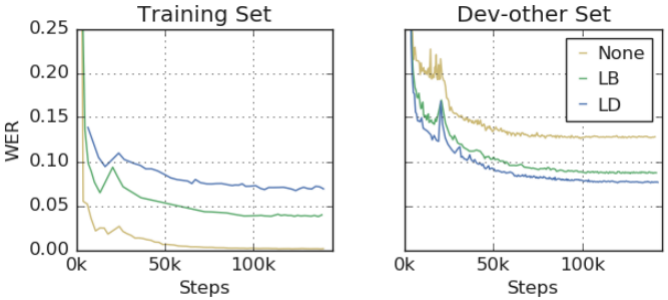
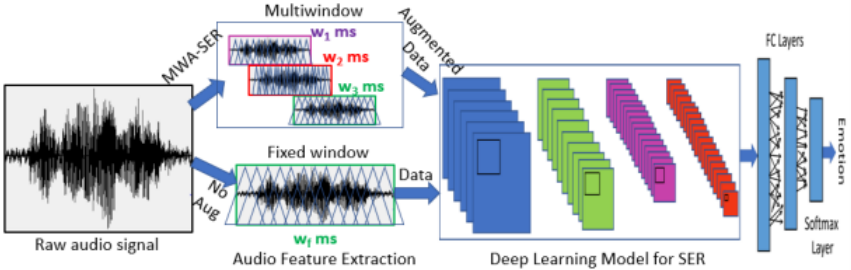


Рис. 9: LAS-6-1280 on LibriSpeech

Multi-Window Data Augmentation



References

- SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition
  Daniel S. Park , William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, Quoc V. Le
  https://arxiv.org/pdf/1904.08779.pdf
- Multi-Window Data Augmentation Approach for Speech Emotion Recognition
  Sarala Padi, Dinesh Manocha, Ram D.Sriram
  https://arxiv.org/pdf/2010.09895.pdf
- Mel-spectrogram augmentation for sequence-to-sequence voice conversion
  Yeongtae Hwang , Hyemin Cho , Hongsun Yang , Dong-Ok Won , Insoo Oh , and Seong-Whan Lee
  https://arxiv.org/pdf/2001.01401.pdf

References

- Surgical Mask Detection with Convolutional Neural Networks and Data Augmentations on Spectrograms
  Steffen Illium, Robert Muller, Andreas Sedlmeier and Claudia Linnhoff-Popien
  https://arxiv.org/pdf/2008.04590.pdf
- Mask Detection and Breath Monitoring from Speech: on Data Augmentation, Feature Representation and Modeling
  Haiwei Wu , Lin Zhang , Lin Yang , Xuyang Wang , Junjie Wang , Dong Zhang , Ming Li
  https://arxiv.org/pdf/2008.05175.pdf