SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition

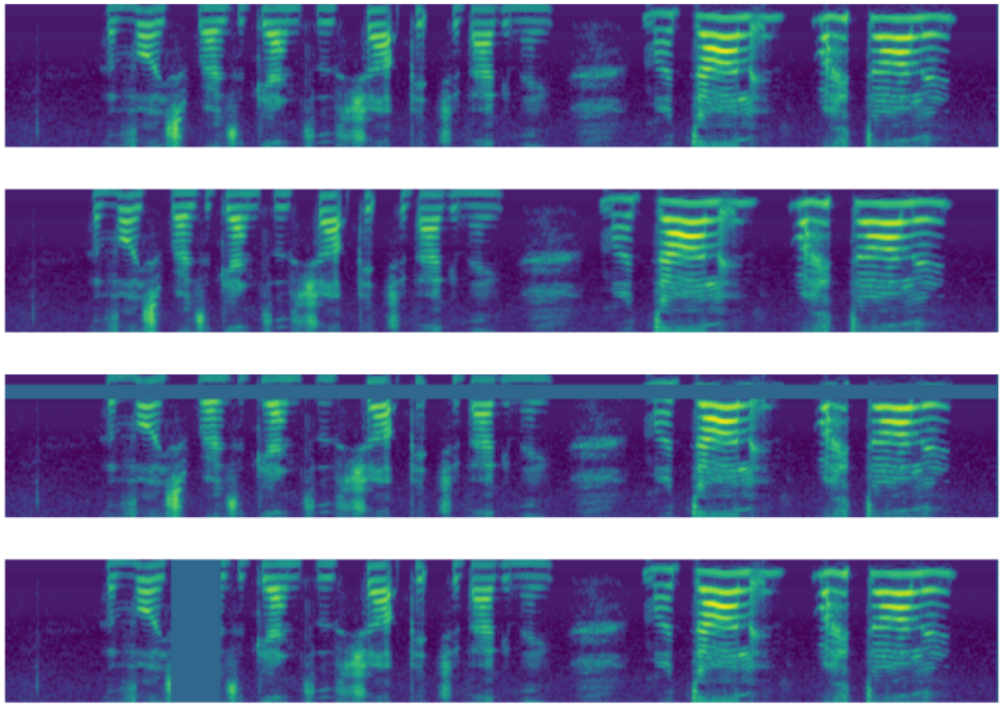
논문을 보기 앞서 spectogram에 대한 개념을 알기 위한 모두연 페이스북 카드 자료입니다.

- 1부 - 복소수 https://www.facebook.com/lab4all/posts/565829610276678
- 2부 - 오일러의 공식 https://www.facebook.com/lab4all/posts/573903102802662
- 3부 - 푸리에 해석 https://www.facebook.com/lab4all/posts/590737297785909

ABSTRACT

이미지의 경우 데이터의 양을 늘리기 위해 다양한 data augmentation 방법들을 사용합니다. 본 논문에서는음성인식을 위한 간단한 data augmentation를 제안하고 이를 SpecAugment라고 명명합니다. SpecAugment는 log mel spectogram을 input으로 Time warping, Frequency masking, Time masking 3가지 방법으로 data augmentation을 합니다. 음성인식 네트워크로 Listen, Attend and Spell을 사용하고, LibriSpeech와 SwitchBoard 데이터 셋을 대상으로 성능 개선을 이루었습니다.

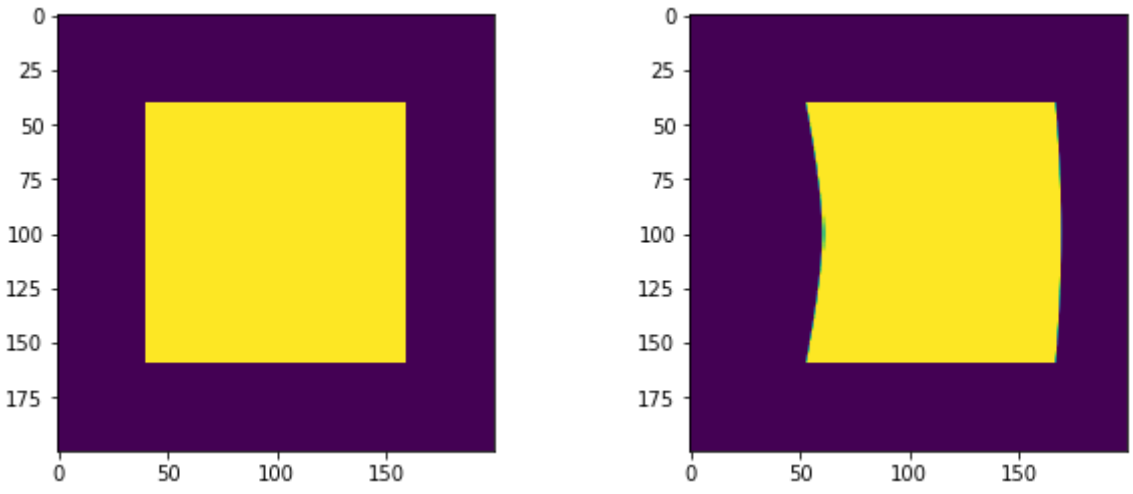
Augmentation Policy



출처 - https://arxiv.org/abs/1904.08779

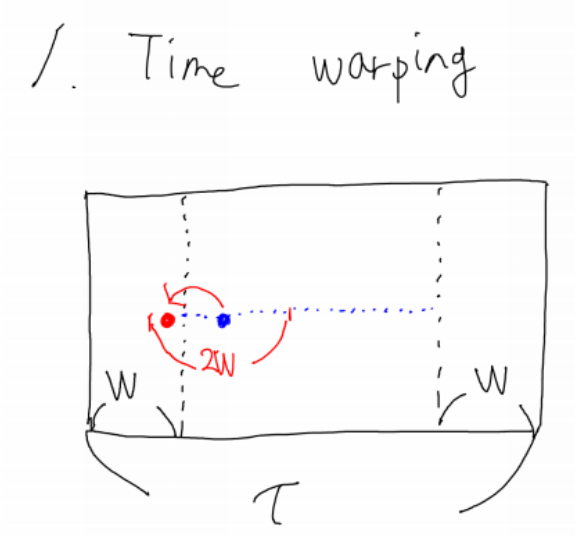
위 그림은 첫번째 원본데이터를 시작으로 각 Time warping, Frequency masking, Time masking 순으로 data augment한 그림입니다. Time warping의 경우 왼쪽 부분은 줄어들고 오른쪽 부분은 늘어난걸 볼 수 있습니다. Frequency maksing과 Time masking의 경우 데이터에 초록색 줄이 나타는 것을 볼 수 있습니다.

Time warping 부터 살펴보겠습니다. 기본적으로 tensorflow에 있는 sparse image warp 함수를 사용합니다. sparse image warp 함수는 하나의 point를 다른 point로 옮기는 함수입니다. 아래 그림에서 사각형의 왼쪽 중심을 오른쪽으로 조금 움직인 예시입니다.

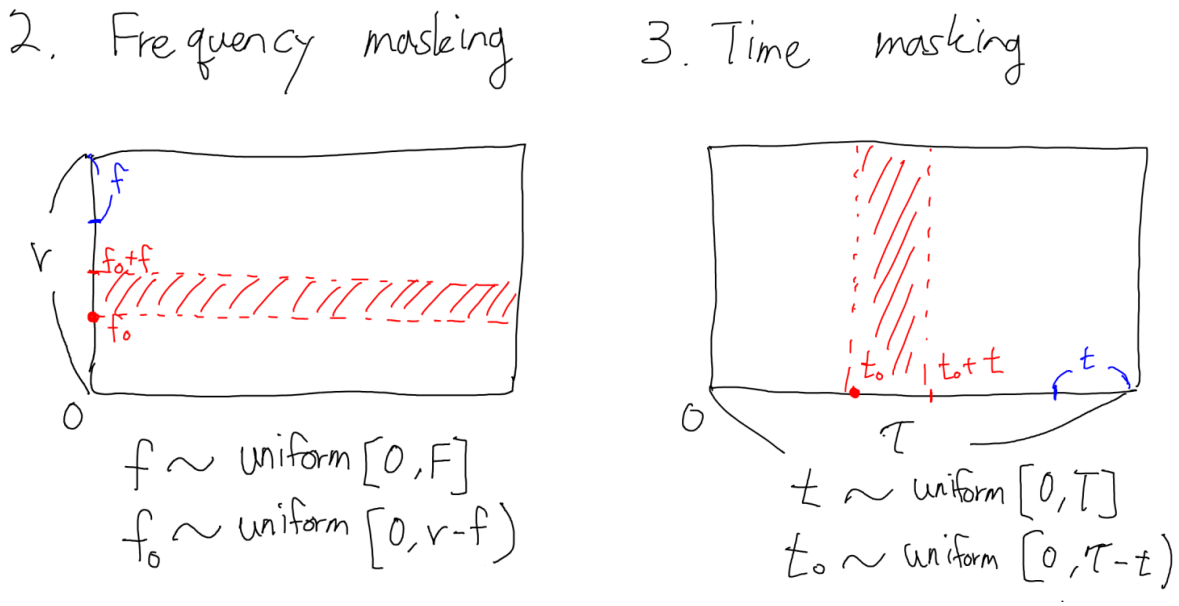


출처 - 이재원님

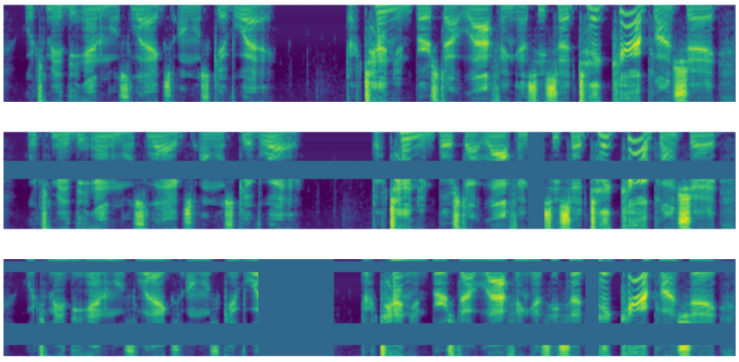
Time warping에서는 하이퍼 파라미터 W가 주어 집니다. 전체 이미지에서 양쪽으로 W만큼 줄인 후 가로 중앙선(파란선)에서 랜덤하게 하나의 점(파란점)을 고릅니다. 파란점이 출발 point가 됩니다. 이후 파란점을 기준으로 2W 범위 내에서 랜덤하게 하나의 점(빨간점)을 고릅니다. 빨간점이 도착 point가 됩니다. 이를 sparse image warp 함수에 파라미터로 사용합니다.



masking의 경우 Frequency, Time 모두 방법은 동일합니다. Frequency로 예를 들면, 먼저 하이퍼 파라미터 F가 주어집니다. 이후 [0,F] 범위에서 랜덤하게 f 값을 뽑습니다. 이 값이 masking할 양이 됩니다. 이후 [0, v-f)에서 랜덤하게 f0 값을 뽑고, [f0, f0+f) 만큼 0으로 masking 합니다. Time masking의 경우도 Frequency와 동일합니다.



데이터에 따라 다른 하이퍼 파라미터를 사용합니다. LibriSpeech basic (LB), LibriSpeech double (LD), Switchboard mild (SM) and Switchboard strong (SS) 입니다.



출처 - https://arxiv.org/abs/1904.08779 LB, LD 예시

Policy	W	F	m_F	T	p	m_T
None	0	0	-	0	-	-
LB	80	27	1	100	1.0	1
LD	80	27	2	100	1.0	2
SM	40	15	2	70	0.2	2
SS	40	27	2	70	0.2	2

왼쪽 표에서 mF와 mT는 masking의 개수를 뜻하고, p는 time masking이 2개 이상일 때 겹칠 수 있는 최대 길이 입니다. 예를 들어 SM과 SS의 경우 겹쳐진 time maksing의 길이가 전체의 0.2를 넘을 수 없습니다.

Model

LAS Network Architectures

SpecAugment로 data augmentation을 한 후, Listen, Attend and Spell(LAS) 네트워크를 사용합니다

- https://arxiv.org/pdf/1508.01211.pdf

Learning Rate Schedules

수월한 학습을 위해 learning rate를 조정합니다. ramp-up, noise, exponentially decay 세 단계를 거칩니다. Sr까지 ramp-up을 하고, Snoise에서 weight에 noise를 주고, [Si, Sf] 구간에서 exponentially decay를 합니다.

$$\text{B(asic): } (s_r, s_{\text{noise}}, s_i, s_f) = (0.5\text{k}, 10\text{k}, 20\text{k}, 80\text{k})$$
$$\text{D(ouble): } (s_r, s_{\text{noise}}, s_i, s_f) = (1\text{k}, 20\text{k}, 40\text{k}, 160\text{k})$$

$$\text{L(ong): } (s_r, s_{\text{noise}}, s_i, s_f) = (1\text{k}, 20\text{k}, 140\text{k}, 320\text{k})$$

Shallow Fusion with Language Models

성능을 높이기 위해 음성인식 뿐 아니라 Language Model도 함께 사용합니다. 아래와 같이 LAS 네트워크 output과 Language Model의 output을 더하여 사용합니다. 논문에서 람다 값은 0.35로 설정합니다.

$$\mathbf{y}^* = \underset{\mathbf{y}}{\text{argmax}} \left(\log P(\mathbf{y}|\mathbf{x}) + \lambda \log P_{LM}(\mathbf{y}) \right) , \qquad (1)$$

작성자 - 정광직