Data Augmentation

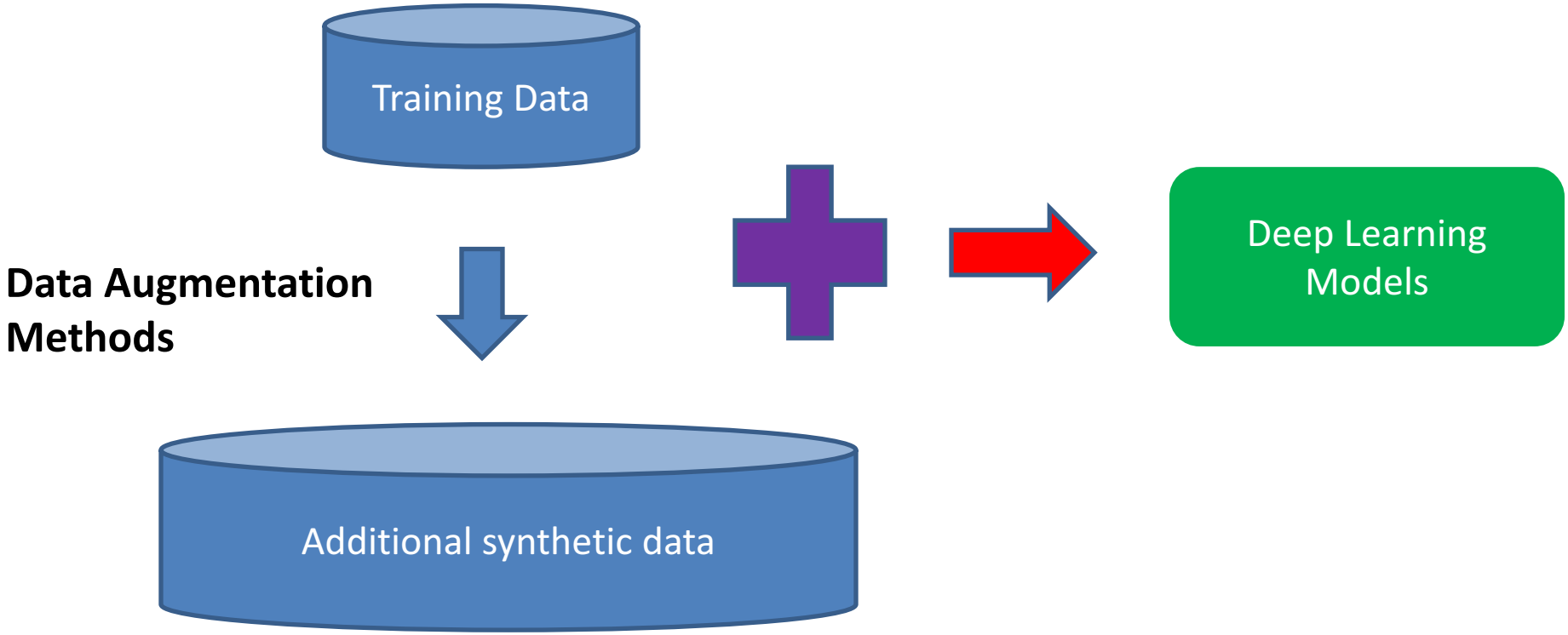
Thang Vu

General Observations

- Deep learning models are powerful
  - But data hungry, esp. annotated data
  - The more data, the better is the performance
- In many applications, we do not have enough data
- One general comment to improve deep learning models
  - Get more data ;-)

Data Augmentation

- Methods that allow you to generate additional synthetic data based on the data which you have



Data Augmentation for Computer Vision

Original	Flip	Rotation	Random crop
			
• Image without any modification	• Flipped with respect to an axis for which the meaning of the image is preserved	• Rotation with a slight angle • Simulates incorrect horizon calibration	• Random focus on one part of the image • Several random crops can be done in a row

Color shift	Noise addition	Information loss	Contrast change
			
• Nuances of RGB is slightly changed • Captures noise that can occur with light exposure	• Addition of noise • More tolerance to quality variation of inputs	• Parts of image ignored • Mimics potential loss of parts of image	• Luminosity changes • Controls difference in exposition due to time of day

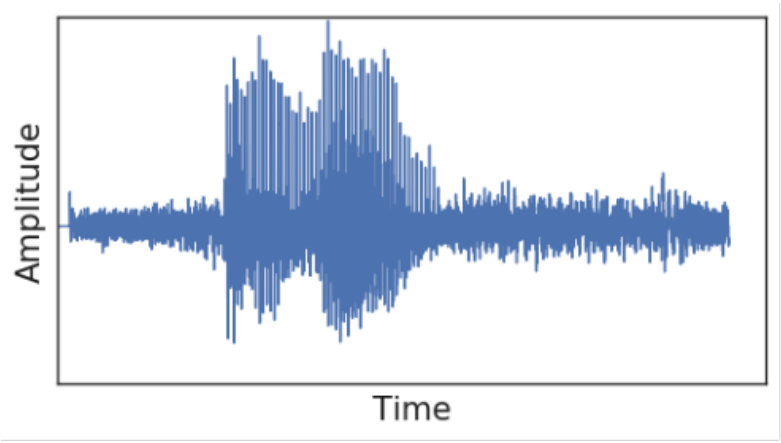
https://stanford.edu/~shervine/teaching/cs-230/cheatsheet-deep-learning-tips-and-tricks

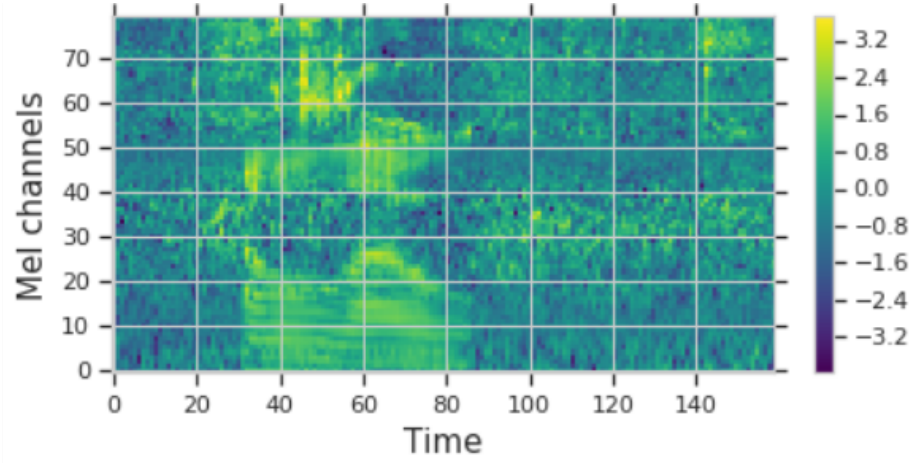
Data Augmentation in Speech & NLP

- SpecAugment: A New Data Augmentation Method for Automatic Speech Recognition
- Back Translation Techniques

SpecAugment

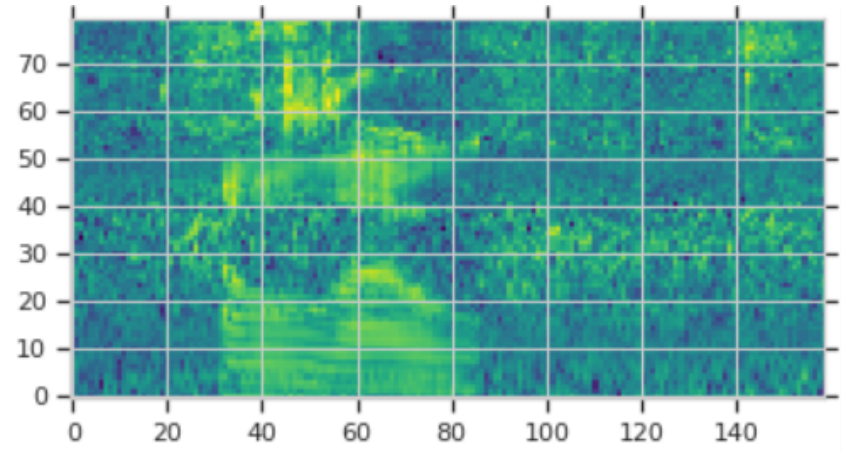
- Traditional augmentation methods:
  - Speeding up or slowing down audio files
  - Adding background noises to the audio files
- The key idea of SpecAugment is to manipulate audio spectrogram

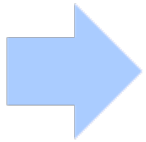


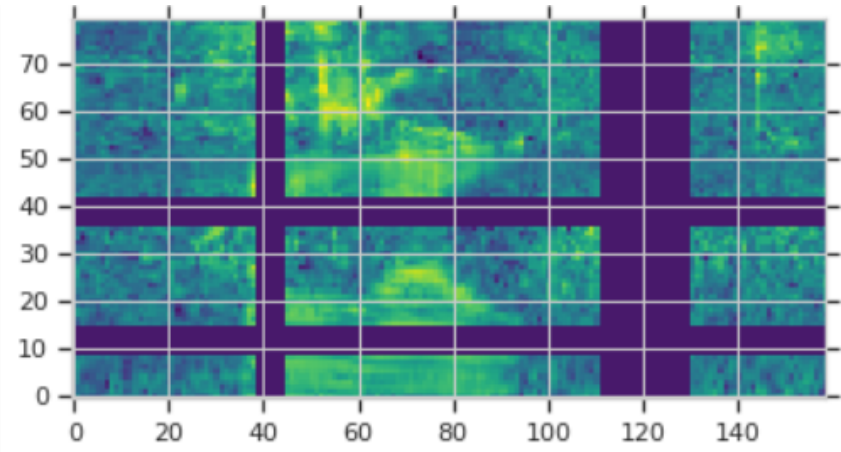


https://ai.googleblog.com/2019/04/specaugment-new-data-augmentation.html

SpecAugment



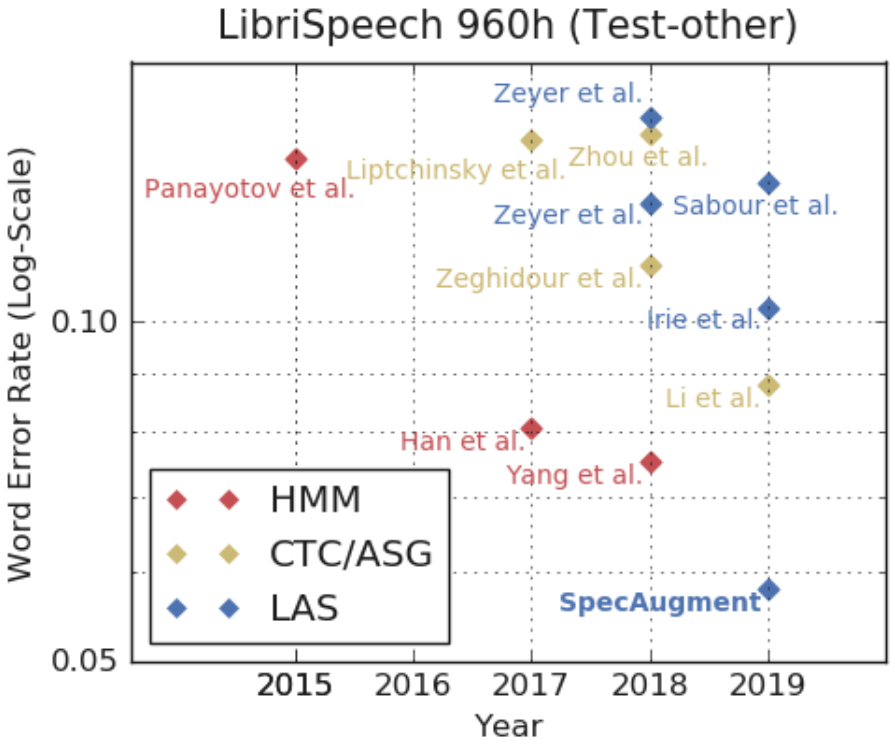


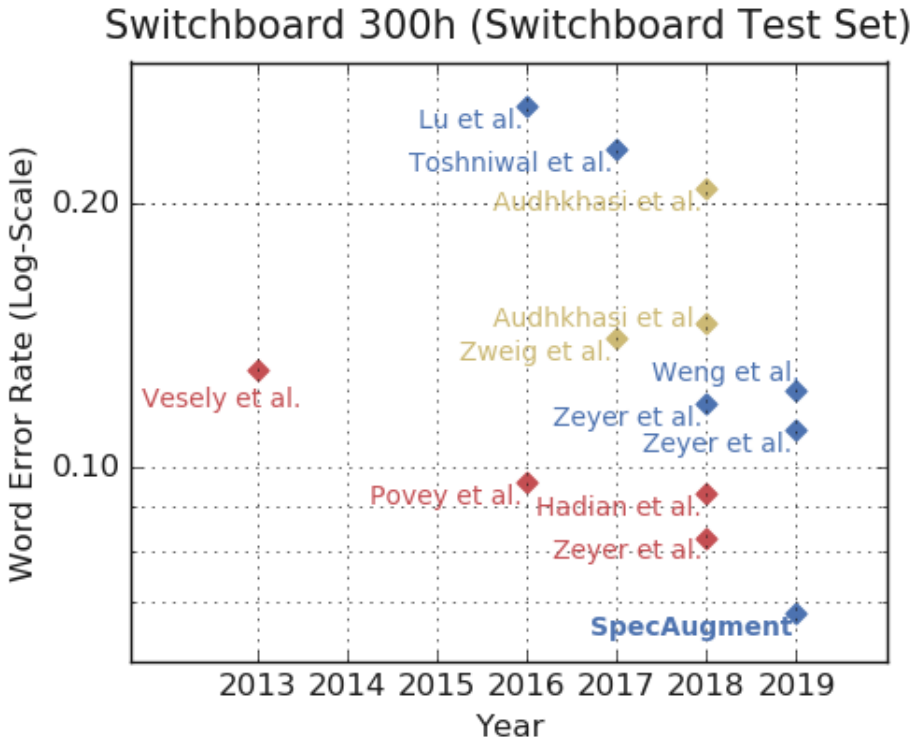


- Warping in the time direction
- Masking multiple blocks of consecutive time steps (vertical masks)
- Masking multiple blocks of mel frequency channels (horizontal masks)

https://ai.googleblog.com/2019/04/specaugment-new-data-augmentation.html

SpechAugment Results





https://ai.googleblog.com/2019/04/specaugment-new-data-augmentation.html

Back Translation

- The key idea is to automatically generate texts using machine translation systems
  - Given an input text in a source language (e.g. English)
  - translate this text to another language (e.g. English to German)
  - translate back the translated text into the source language (e.g. German to English)

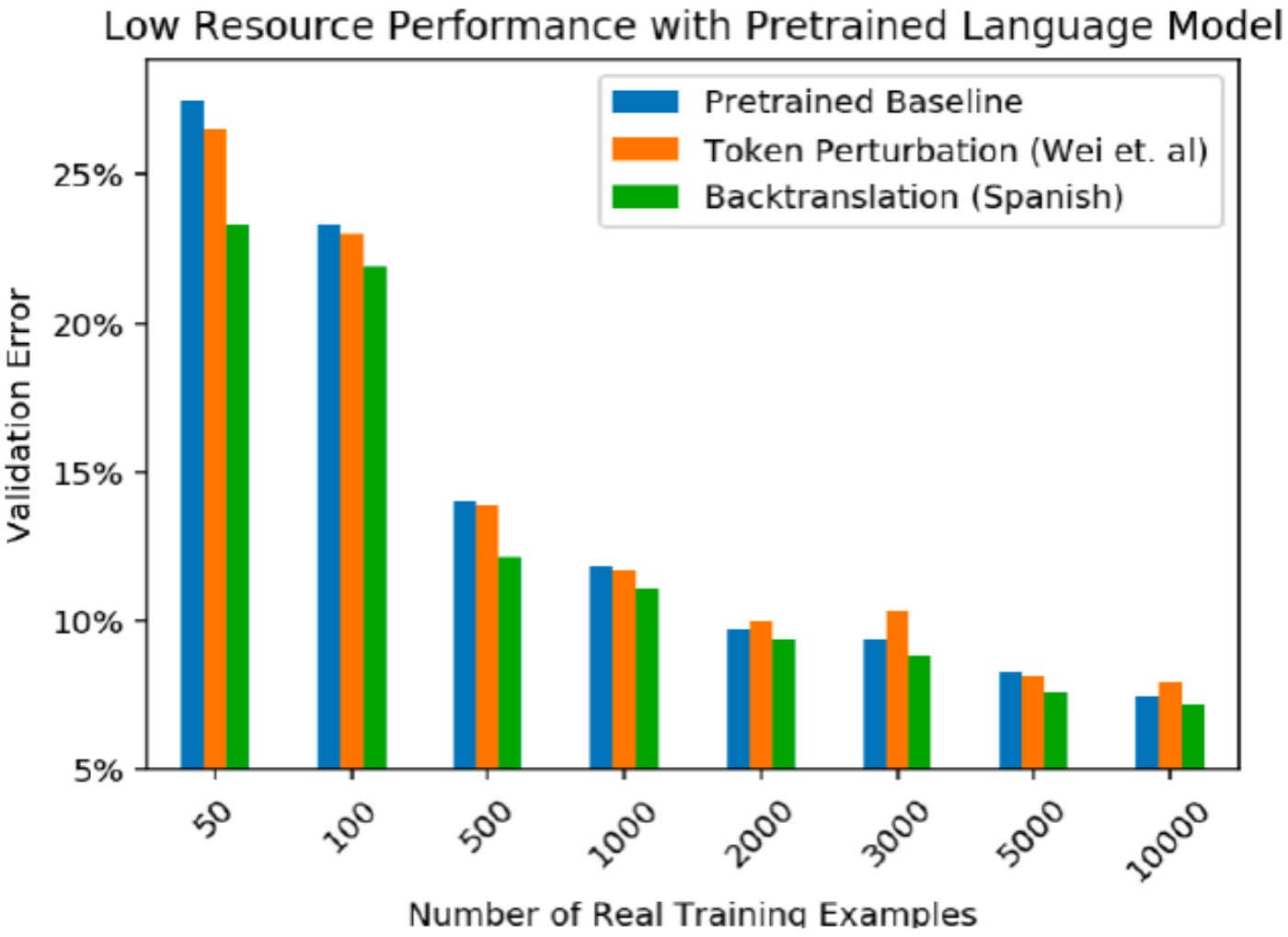
Back Translation for Text Classification

Operation	Sentence
None	A sad human comedy played out on the back roads of life.
BT (Spanish)	A sad human comedy that develops in the secondary roads of life.
BT (Bengali)	A sad man played the street behind comedy life.
Synonmym Replace†	A lamentable human comedy played out on the backward road of life.
Random Insert†	A sad human comedy played out on funniness the back roads of life.
Random Swap†	A sad human comedy played out on roads back the of life.
Random Delete†	A sad human out on the _ roads of life.

Table 1: BT stands for backtranslation. † Token Perturbation techniques from Wei and Zhou, [2019]

https://arxiv.org/pdf/1903.09244.pdf

Back Translation for Text Classification



https://arxiv.org/pdf/1903.09244.pdf