Vienna Deep Learning Meetup

November 20, 2017 @ A1 Telekom Austria

Vienna Deep Learning Meetup

The Organizers:

Thomas Lidy

Jan Schlüter

Alex Schindler

Vienna 14th Deep Learning Meetup

Agenda:

- Welcome by the Organizers
- Introduction by A1 Telekom Austria (Alexander Stock, Director IT)
- **Evolution of Image Search @ Seznam.cz** (Lukáš Vrabel)

30 minutes break

- Hot Topics & Latest News (Tom Lidy, Alex Schindler, Jan Schlüter)
- Discussions and Networking

Vienna Deep Learning Meetup

Vienna
Deep Learning
Meetup

A1

Hot Topics & Latest News

Tom Lidy,
Alexander Schindler,
Jan Schlüter

a short block at every meetup
to briefly present recent papers and news

Send us contributions (tom.lidy@gmail.com)
or come with slides to do a short block yourself!

Detection and Classification of Acoustic Scenes and Events

An IEEE AASP Challenge

DCASE2017 WORKSHOP

16 - 17 November 2017, Munich, Germany

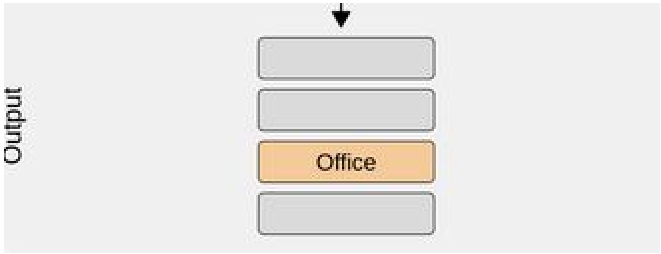
DCASE2017 CHALLENGE

15 March 2017 - 31 July 2017

Acoustic scene classification

Scenes
Task 1

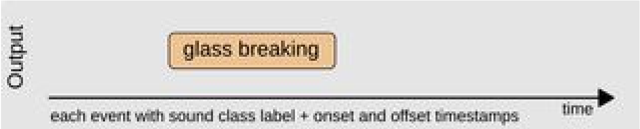
- Classify audio file
- 15 scenes (park, street, office, supermarket, home, etc)
- Multiclass Single Label prediction



Detection of rare sound events

Alarms
Task 2

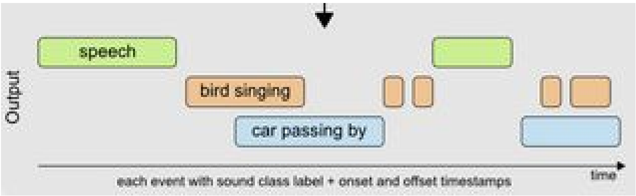
- rare sound events in artificially created mixtures
- 3 classes (Baby Crying, Glass Breaking, Gunshot)
- Sequence labelling (start/stop of event), one class per file


Output
glass breaking
each event with sound class label + onset and offset timestamps
time

Sound event detection in real life audio

Events
Task 3

- Classify audio file
- 6 classes (brakes squeaking, car, children, people speaking, people walking)
- Sequence labelling (start/stop of event), multi-label classification


Output
speech
bird singing
car passing by
each event with sound class label + onset and offset timestamps
time

Large-scale weakly supervised sound event detection for smart cars

Large-scale
Task 4

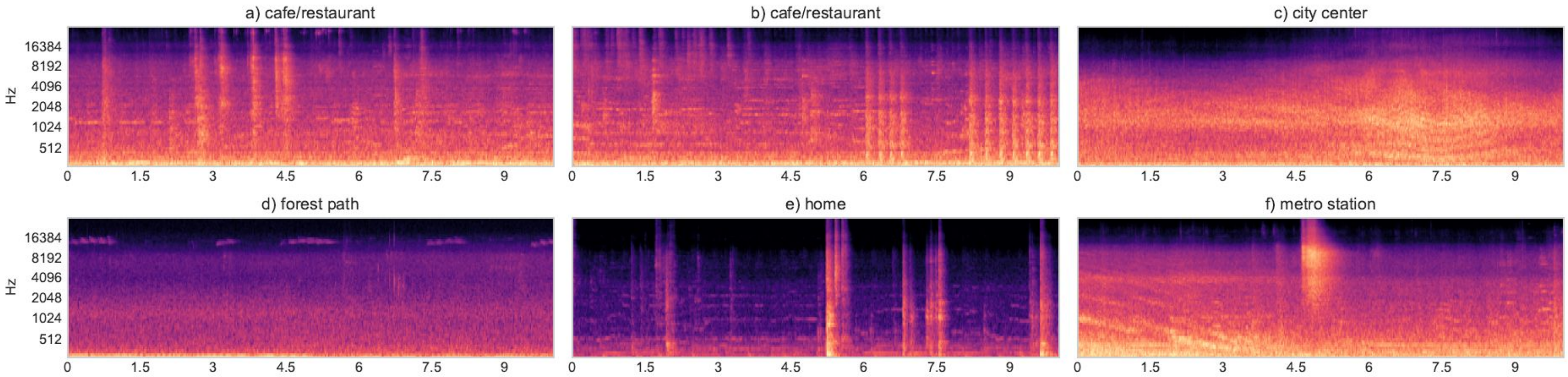
- subset of “AudioSet: An Ontology And Human-Labeled Dataset For Audio Events”
- 17 sound events divided into two categories: “Warning” and “Vehicle”
- Sequence labelling (start/stop of event), multi-label classification

Vienna
Deep Learning
Meetup

Multi-resolution Convolutional Neural Networks

Alexander Schindler, Thomas Lidy and Andreas Rauber

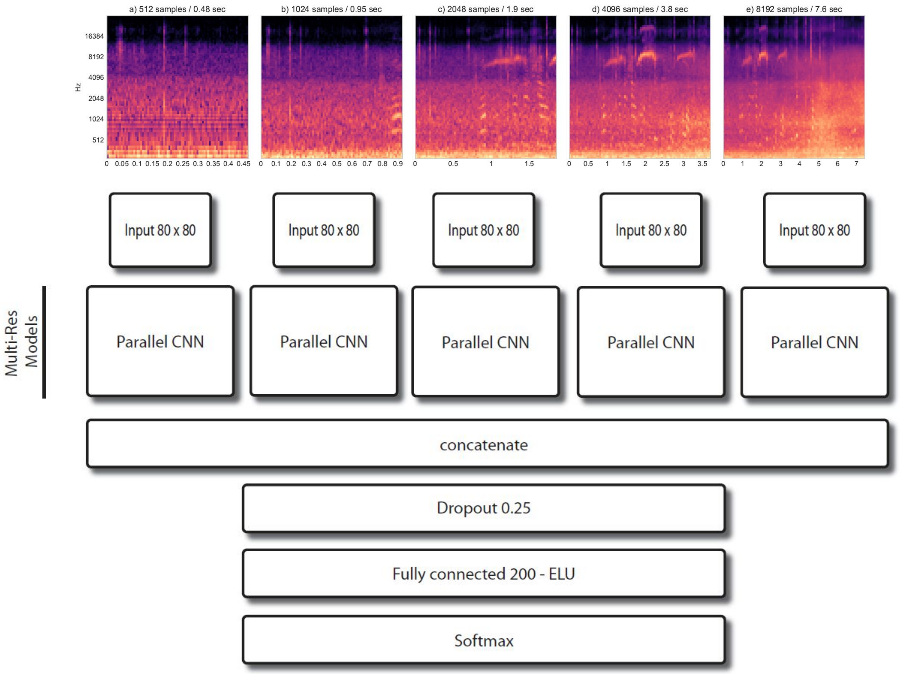
- **Problem:** Observation DCASE 2016
    - Confusion between Train / Tram / Bus / Supermarket
    - Low-Frequent Humming
        - Diesel engine, air-condition, refrigerators
    - Acoustic scene composed of
        - Timbral texture (short-term)
        - Acoustic events (long-term)



Multi-resolution Convolutional Neural Networks

Alexander Schindler, Thomas Lidy and Andreas Rauber

- **Solution:** Train Model on multiple-temporal resolutions



Input 80 x 80 | Input 80 x 80 | Input 80 x 80 | Input 80 x 80 | Input 80 x 80

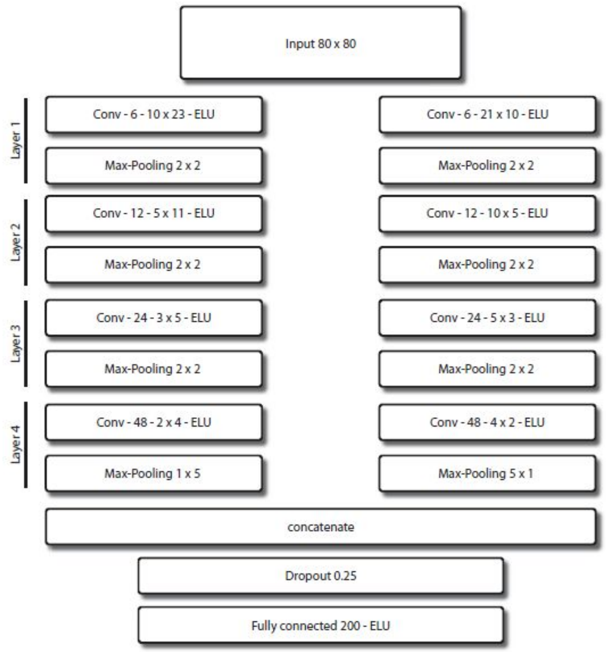
Multi-Res Models: Parallel CNN | Parallel CNN | Parallel CNN | Parallel CNN | Parallel CNN

concatenate

Dropout 0.25

Fully connected 200 - ELU

Softmax



Input 80 x 80

Layer 1
- Conv - 6 - 10 x 23 - ELU | Conv - 6 - 21 x 10 - ELU
- Max-Pooling 2 x 2 | Max-Pooling 2 x 2

Layer 2
- Conv - 12 - 5 x 11 - ELU | Conv - 12 - 10 x 5 - ELU
- Max-Pooling 2 x 2 | Max-Pooling 2 x 2

Layer 3
- Conv - 24 - 3 x 5 - ELU | Conv - 24 - 5 x 3 - ELU
- Max-Pooling 2 x 2 | Max-Pooling 2 x 2

Layer 4
- Conv - 48 - 2 x 4 - ELU | Conv - 48 - 4 x 2 - ELU
- Max-Pooling 1 x 5 | Max-Pooling 5 x 1

concatenate

Dropout 0.25

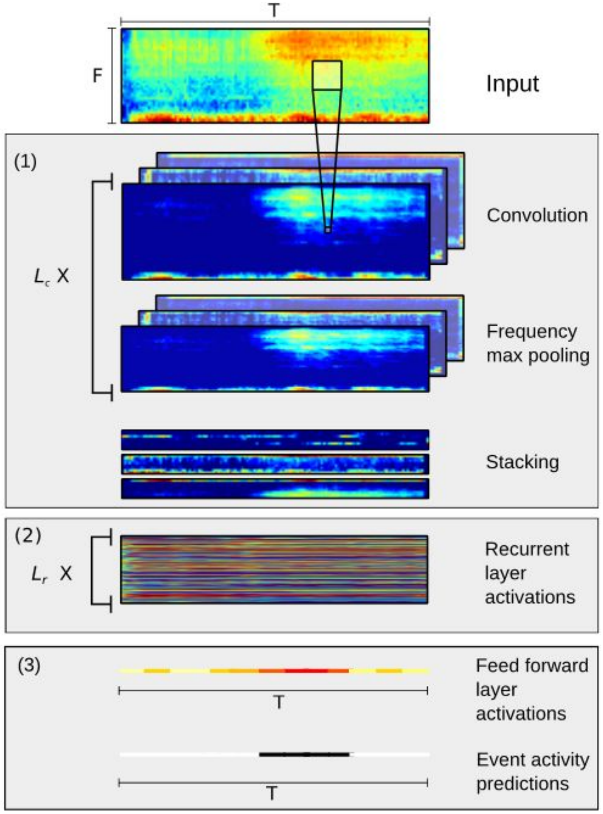
Fully connected 200 - ELU

Vienna Deep Learning Meetup

DCASE Workshop

Hot Topics in
Audio event Detection

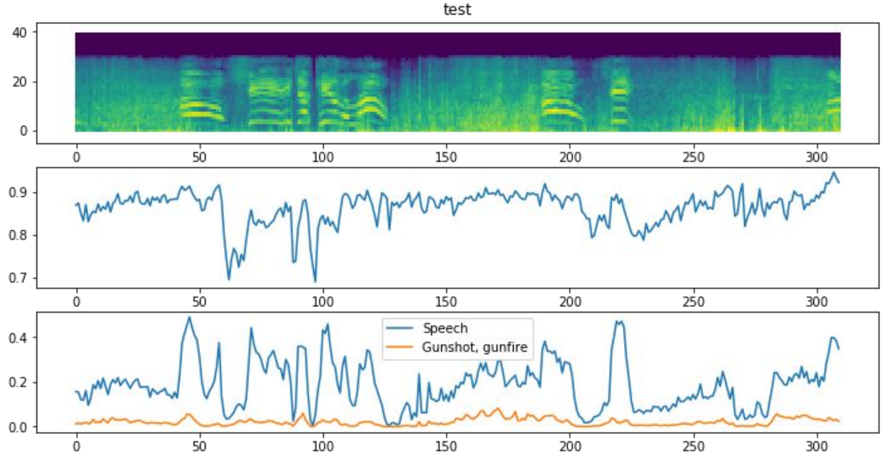
- **Recurrent Convolutional Neural Networks** (RCNN)
  - dominant in audio event detection
  - CNN learns audio representation
  - CNN processes sequential slices of Spectrogram input
  - Stacking CNN outputs
  - Recurrent Layer
  - Sequential ("TimeDistributed") Fully Connected Layers

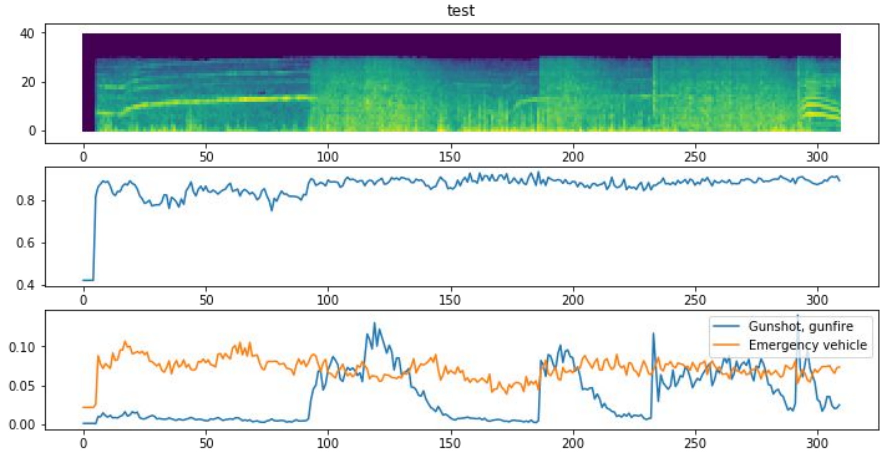

T
F
Input
(1)
L_c X
Convolution
Frequency max pooling
Stacking
(2)
L_r X
Recurrent layer activations
(3)
Feed forward layer activations
T
Event activity predictions
T

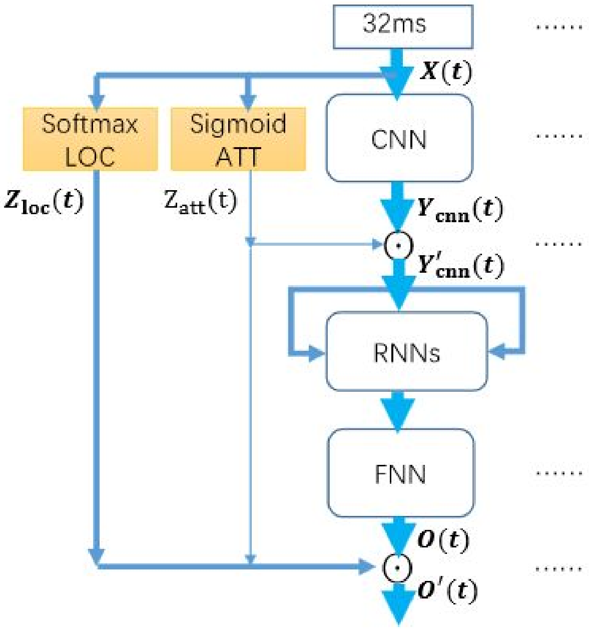
DCASE Workshop

Hot Topics in
Audio event Detection

- **Attention & Localization Layers**

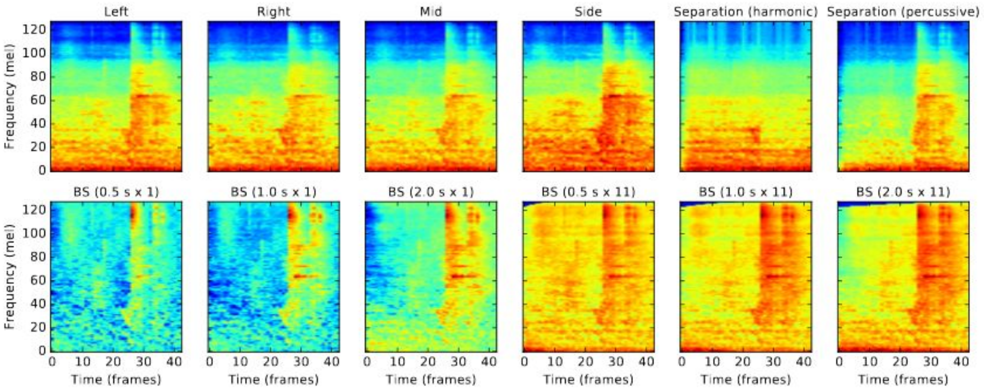


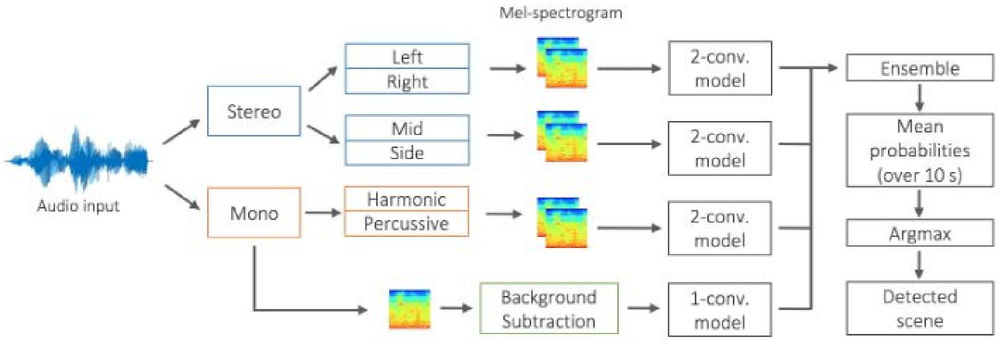




DCASE Workshop
Hot Topics in
Audio event Detection

- **Binaural (stereo) Models**
  - 2nd best model
  - Uses all input channels
    - (usually mono input)

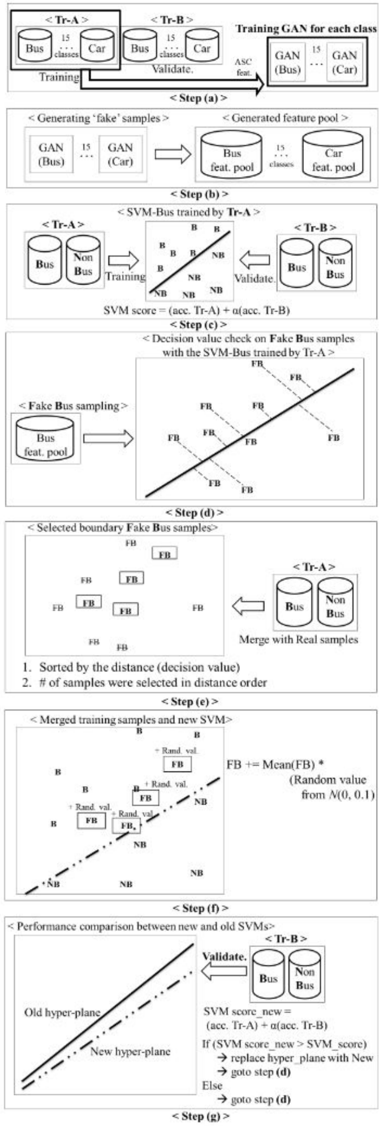




DCASE Workshop

Hot Topics in Audio event Detection

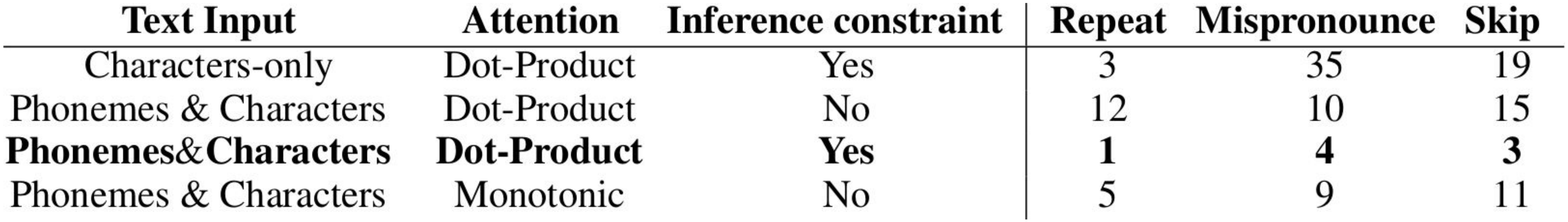
- **Generative Adversarial Network (GAN) based Audio Data Augmentation**
  - Winning Model Task 1
  - Uses Gan to generate training examples
  - Generates feature vectors - not audio



< Tr-A >
Bus
15 classes
Car
Training
< Tr-B >
Bus
15 classes
Car
Validate.
Training GAN for each class
ASC feat.
GAN (Bus)
15
GAN (Car)
< Step (a) >
< Generating 'fake' samples >
GAN (Bus)
15
GAN (Car)
< Generated feature pool >
Bus feat. pool
15 classes
Car feat. pool
< Step (b) >
< SVM-Bus trained by Tr-A >
< Tr-A >
Bus
Non Bus
Training
B
NB
Validate.
< Tr-B >
Bus
Non Bus
SVM score = (acc. Tr-A) + α(acc. Tr-B)
< Step (c) >
< Decision value check on Fake Bus samples with the SVM-Bus trained by Tr-A >
< Fake Bus sampling >
Bus feat. pool
FB
< Step (d) >
< Selected boundary Fake Bus samples>
FB
< Tr-A >
Bus
Non Bus
Merge with Real samples
1. Sorted by the distance (decision value)
2. # of samples were selected in distance order
< Step (e) >
< Merged training samples and new SVM>
B
+ Rand. val.
FB
NB
FB += Mean(FB) * (Random value from N(0, 0.1))
< Step (f) >
< Performance comparison between new and old SVMs>
Old hyper-plane
New hyper-plane
Validate.
< Tr-B >
Bus
Non Bus
SVM score_new = (acc. Tr-A) + α(acc. Tr-B)
If (SVM score_new > SVM_score)
→ replace hyper_plane with New
→ goto step (d)
Else
→ goto step (d)
< Step (g) >

Deep Voice 3: 2000-Speaker Neural Text-to-Speech

- Trained on LibriSpeech dataset (820 hours, 2484 speakers)
- Similar quality to Tacotron, but faster
- No comparison to DeepMind’s new WaveNet (which has no paper yet)
- Interesting evaluation:

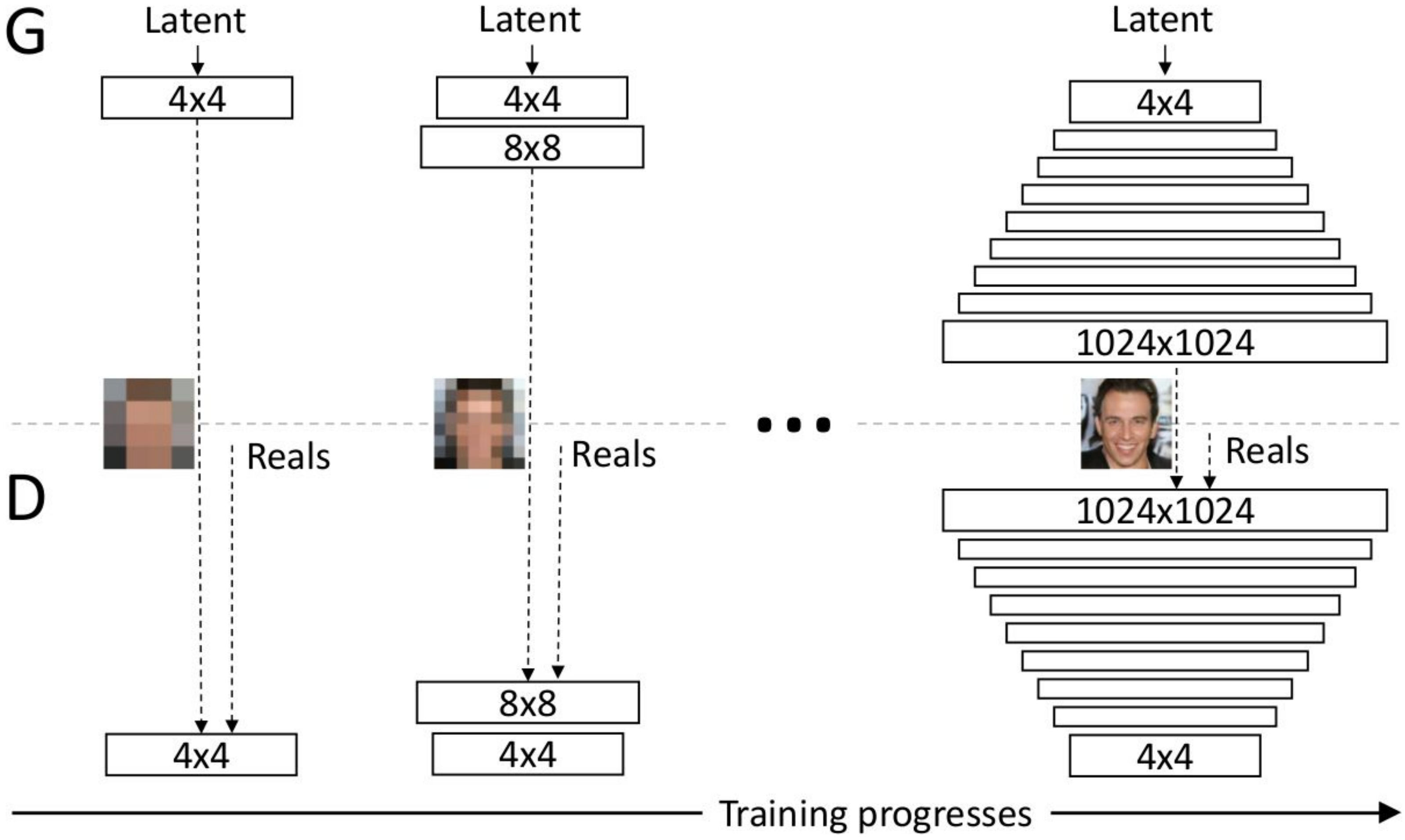


Text Input	Attention	Inference constraint	Repeat	Mispronounce	Skip
Characters-only	Dot-Product	Yes	3	35	19
Phonemes & Characters	Dot-Product	No	12	10	15
Phonemes&Characters	**Dot-Product**	**Yes**	**1**	**4**	**3**
Phonemes & Characters	Monotonic	No	5	9	11

Number of errors made in vocalizing a set of 100 test sentences

http://arxiv.org/abs/1710.07654

Progressive Growing of GANs for Improved Quality, Stability, and Variation

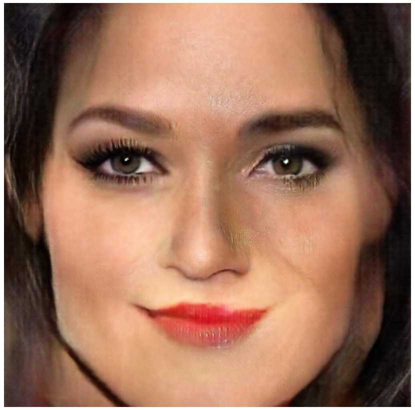


https://github.com/tkarras/progressive_growing_of_gans
http://arxiv.org/abs/1710.10196

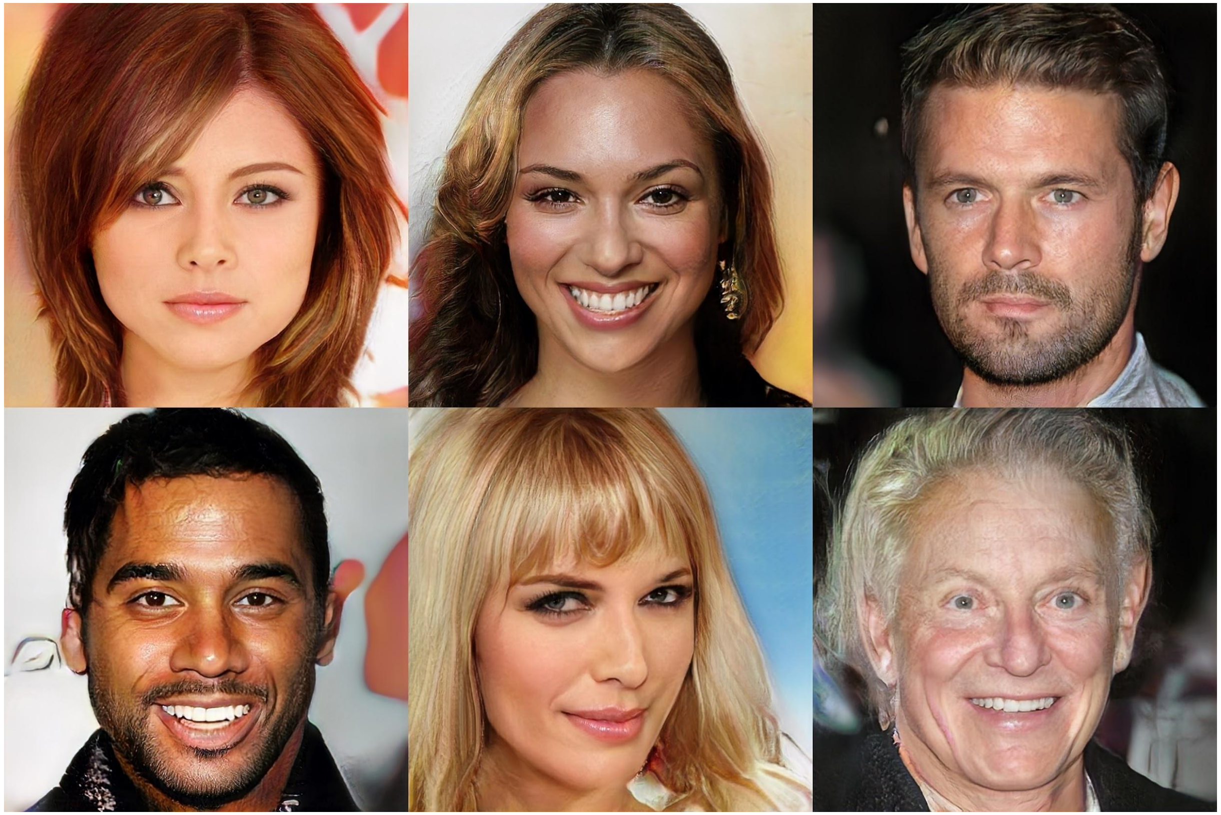
Vienna Deep Learning Meetup

Progressive Growing of GANs for Improved Quality, Stability, and Variation

Existing work:



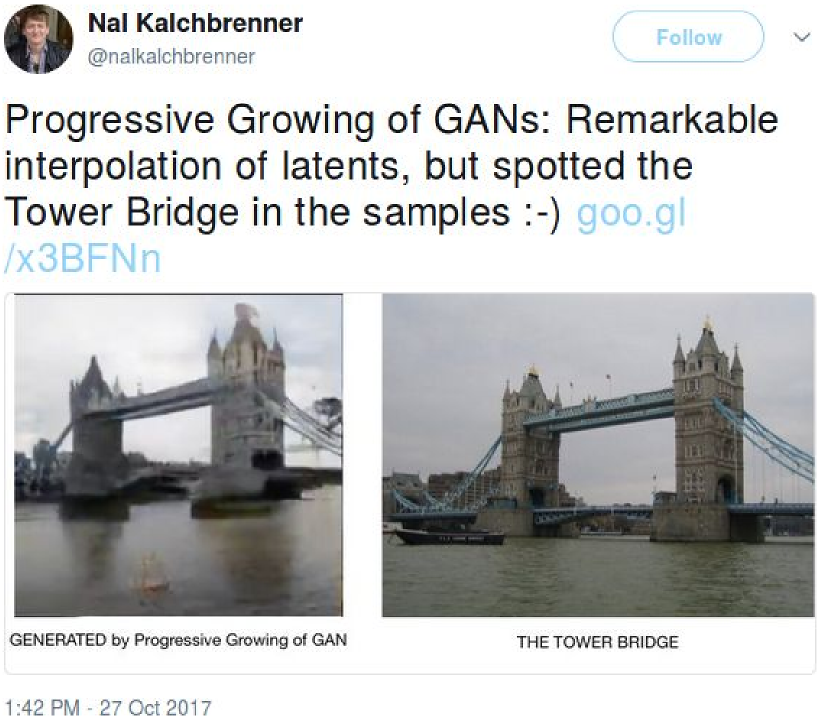
This work:



https://github.com/tkarras/progressive_growing_of_gans
http://arxiv.org/abs/1710.10196

Progressive Growing of GANs for Improved Quality, Stability, and Variation

But:



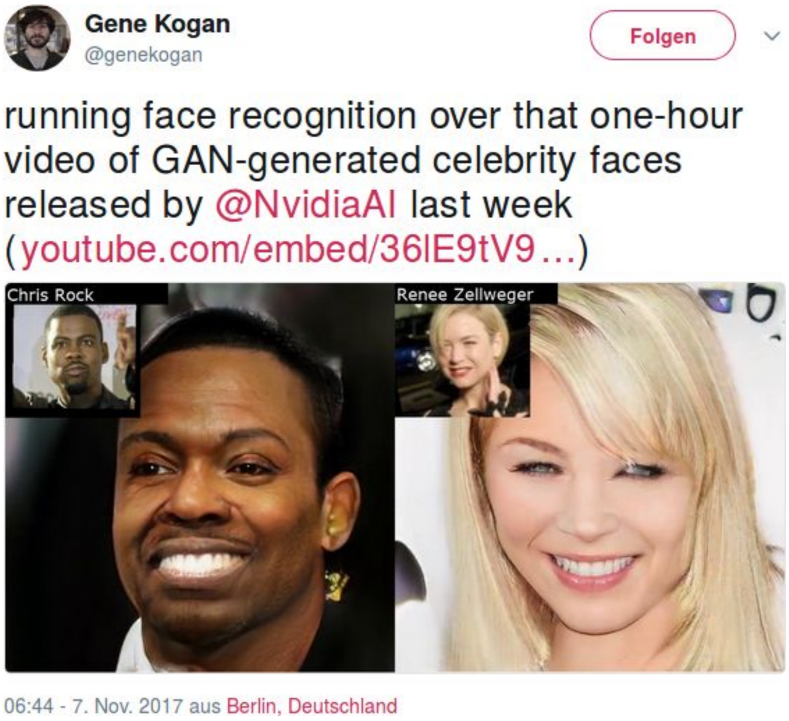
Nal Kalchbrenner
@nalkalchbrenner

Progressive Growing of GANs: Remarkable interpolation of latents, but spotted the Tower Bridge in the samples :-) goo.gl/x3BFNn

GENERATED by Progressive Growing of GAN

THE TOWER BRIDGE

1:42 PM - 27 Oct 2017



Gene Kogan
@genekogan

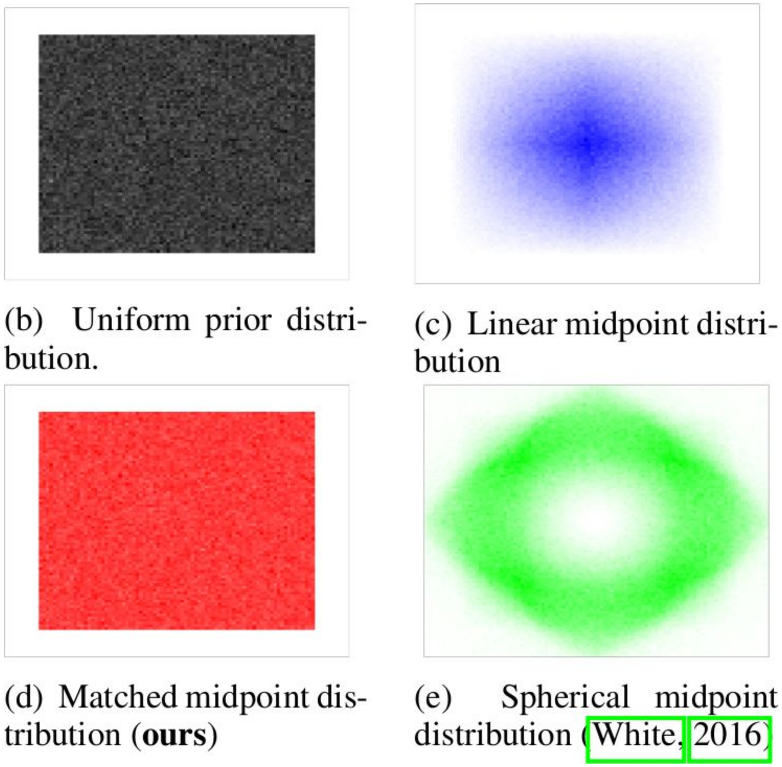
running face recognition over that one-hour video of GAN-generated celebrity faces released by @NvidiaAI last week (youtube.com/embed/36lE9tV9…)

06:44 - 7. Nov. 2017 aus Berlin, Deutschland

https://github.com/tkarras/progressive_growing_of_gans
http://arxiv.org/abs/1710.10196

Optimal transport maps for distribution preserving operations on latent spaces of Generative Models

- GAN learns mapping from random vector in latent space to image
- Typical demonstration in paper: Interpolate in latent space
- **Problem:** in high dimensions, volume of a sphere/cube concentrates near surface/edges ⇒ interpolation traverses space hardly seen in training
- This can be fixed!

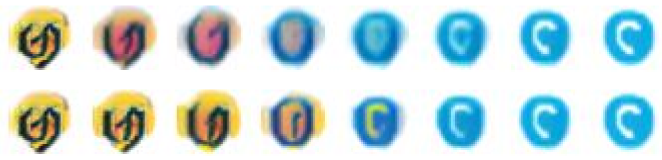


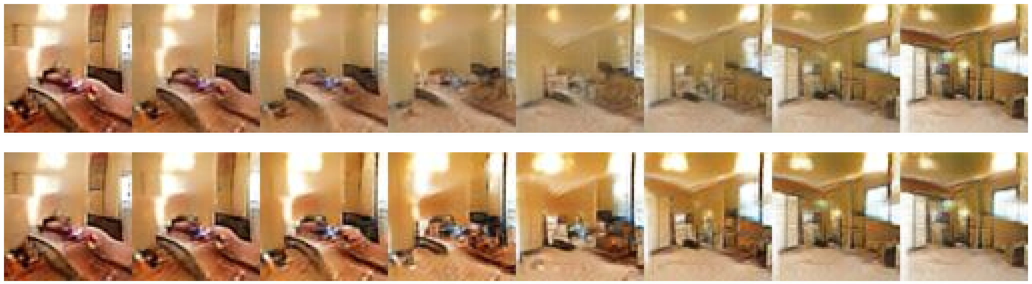
(b) Uniform prior distribution.

(c) Linear midpoint distribution

(d) Matched midpoint distribution (**ours**)

(e) Spherical midpoint distribution (White, 2016)



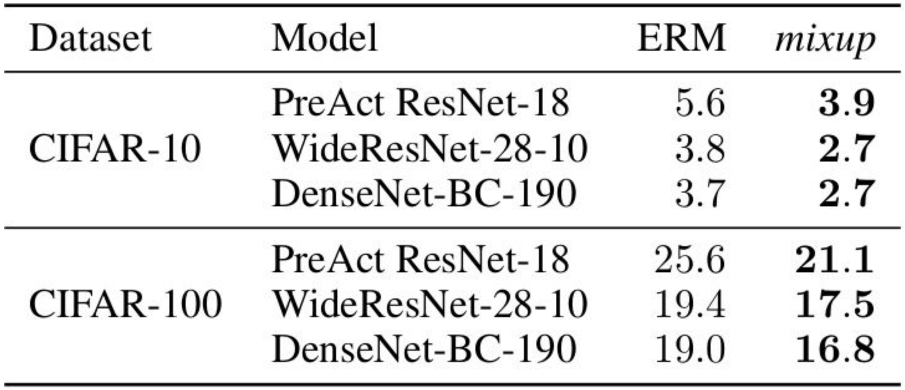


https://arxiv.org/abs/1711.01970

mixup: Beyond Empirical Risk Minimization

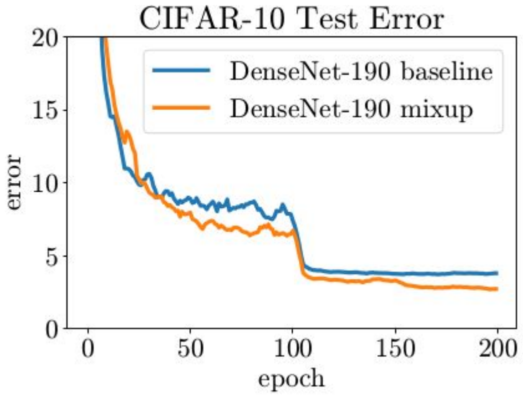
- Proposed regularization method:
  - Draw factor λ from beta distribution ($\alpha = \beta = 0.2$; so mostly near 0/1)
  - Input: linear mixture of two data points $x = \lambda x_1 + (1-\lambda)x_2$
  - Target: linear mixture of respective targets $y = \lambda y_1 + (1-\lambda)y_2$ (with y_1 and y_2 represented as one-hot vectors)

- Improved performance on CIFAR-10/100, ImageNet, Google Commands



Dataset	Model	ERM	*mixup*
CIFAR-10	PreAct ResNet-18	5.6	**3.9**
	WideResNet-28-10	3.8	**2.7**
	DenseNet-BC-190	3.7	**2.7**
CIFAR-100	PreAct ResNet-18	25.6	**21.1**
	WideResNet-28-10	19.4	**17.5**
	DenseNet-BC-190	19.0	**16.8**

(a) Test errors for the CIFAR experiments.

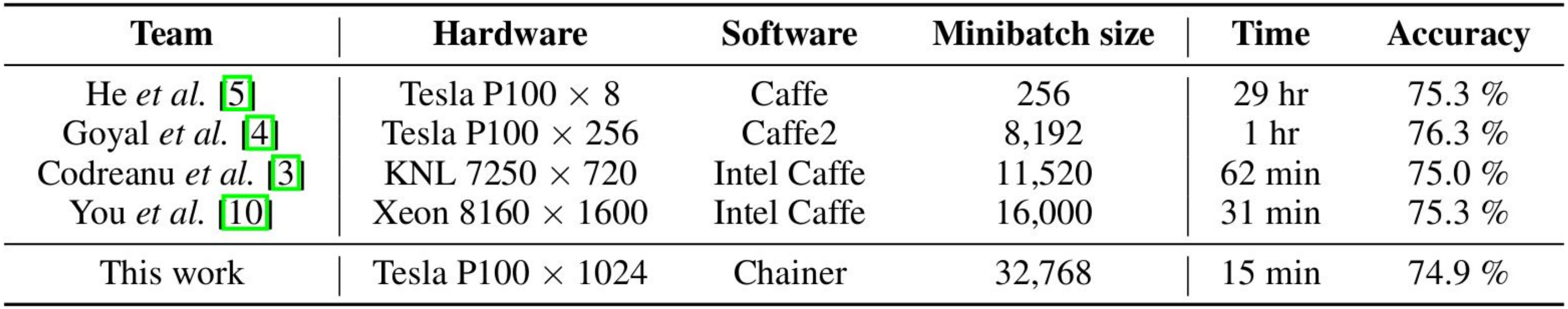


(b) Test error evolution for the best ERM and *mixup* models.

https://arxiv.org/abs/1710.09412
http://www.inference.vc/mixup-data-dependent-data-augmentation/

ImageNet in 15 minutes

- ResNet-50 on ImageNet
- Batchsize: 32k
- Hardware: 1024 NVIDIA P100 (cost: EUR 6000 each)
- Training time: 15 min for 90 epochs



Team	Hardware	Software	Minibatch size	Time	Accuracy
He *et al.* [5]	Tesla P100 × 8	Caffe	256	29 hr	75.3 %
Goyal *et al.* [4]	Tesla P100 × 256	Caffe2	8,192	1 hr	76.3 %
Codreanu *et al.* [3]	KNL 7250 × 720	Intel Caffe	11,520	62 min	75.0 %
You *et al.* [10]	Xeon 8160 × 1600	Intel Caffe	16,000	31 min	75.3 %
This work	Tesla P100 × 1024	Chainer	32,768	15 min	74.9 %

- Some tricks (mostly from Goyal et al.):
  - warm-up with low learning rate
  - warm-up with RMSprop, then switch to SGD
  - no moving averages in Batch Normalization

https://www.preferred-networks.jp/en/news/pr20171110

AI Robot Learned How to Pick up Objects

- University of California, Berkeley, trained a deep learning system on a **simulated** dataset of more than a thousand objects
- DNN exposed to each one's 3D shape and appearance
- Test on physical objects that weren't included in its digital training set

- When the system thought it had a better than 50 percent chance of successfully picking up a new object, it was actually able to do it **98 percent** of the time
  — without having trained on any real-world objects
- shows that a simulated data set can be used to train a model for grasping

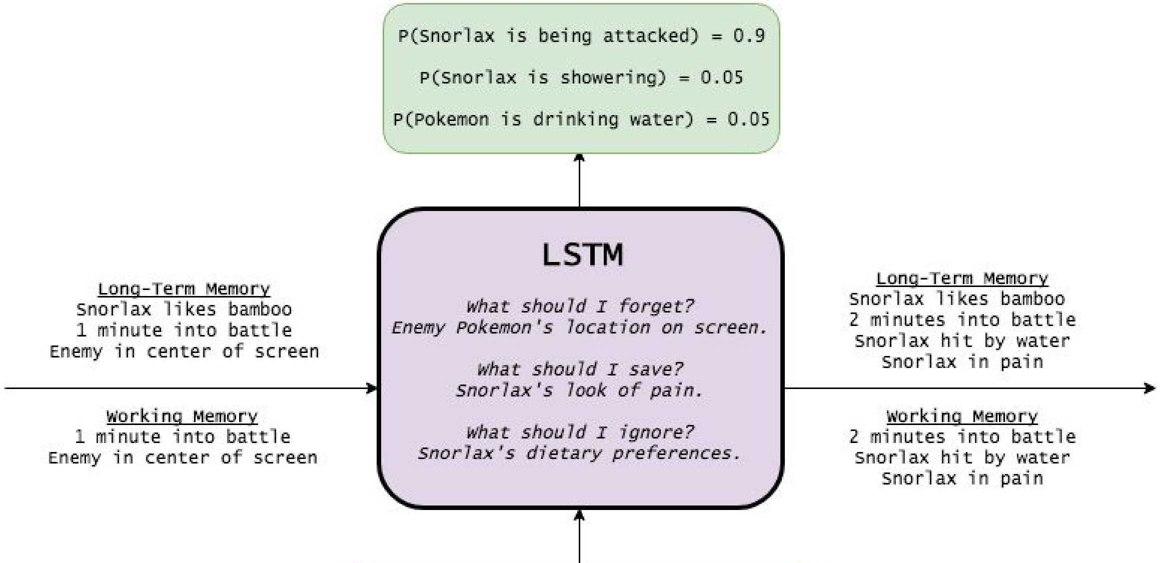
https://futurism.com/an-ai-robot-learned-how-to-pick-up-objects-after-training-only-in-the-virtual-world/

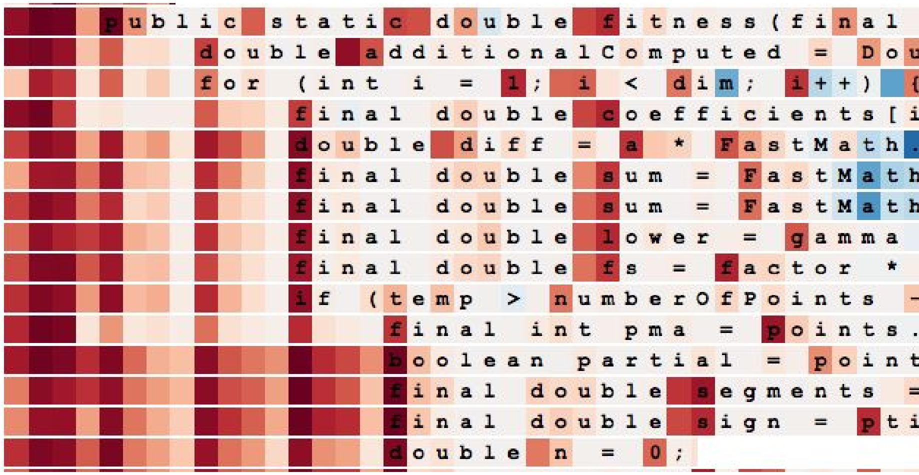
AI Robot Learned How to Pick up Objects

https://futurism.com/an-ai-robot-learned-how-to-pick-up-objects-after-training-only-in-the-virtual-world/

Vienna Deep Learning Meetup

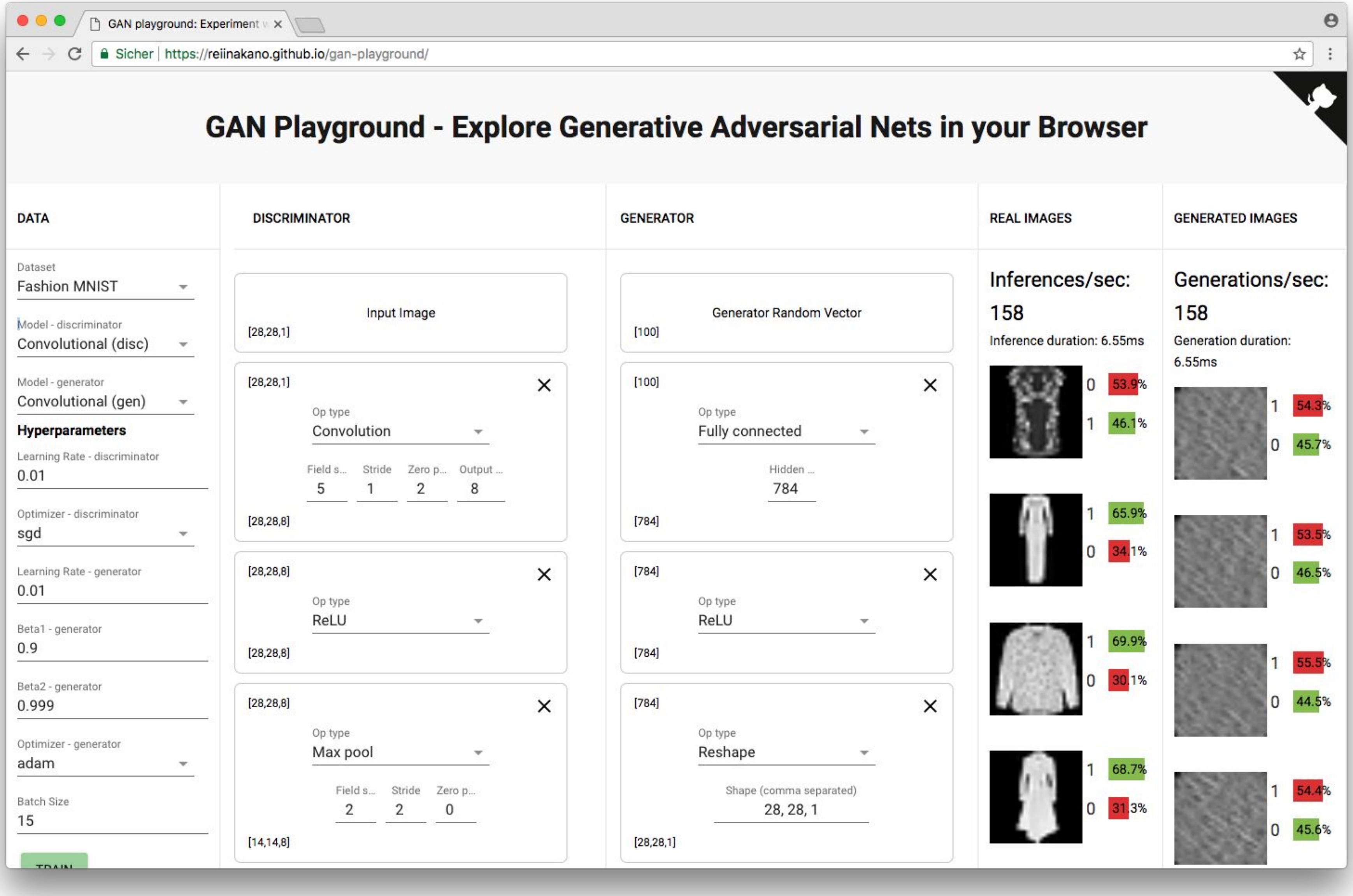
Exploring LSTMs





```
public static double fitness(final
    double additionalComputed = Dou
    for (int i = 1; i < dim; i++) {
        final double coefficients[i
        double diff = a * FastMath.
        final double sum = FastMath
        final double sum = FastMath
        final double lower = gamma
        final double fs = factor *
        if (temp > numberOfPoints -
            final int pma = points.
            boolean partial = point
            final double segments =
            final double sign = pti
            double n = 0;
```

http://blog.echen.me/2017/05/30/exploring-lstms



GAN playground: Experiment w
Sicher | https://reiinakano.github.io/gan-playground/

GAN Playground - Explore Generative Adversarial Nets in your Browser

DATA

Dataset: Fashion MNIST

Model - discriminator: Convolutional (disc)

Model - generator: Convolutional (gen)

Hyperparameters

Learning Rate - discriminator: 0.01

Optimizer - discriminator: sgd

Learning Rate - generator: 0.01

Beta1 - generator: 0.9

Beta2 - generator: 0.999

Optimizer - generator: adam

Batch Size: 15

DISCRIMINATOR

Input Image [28,28,1]

[28,28,1] Op type: Convolution — Field s... 5, Stride 1, Zero p... 2, Output ... 8 — [28,28,8]

[28,28,8] Op type: ReLU — [28,28,8]

[28,28,8] Op type: Max pool — Field s... 2, Stride 2, Zero p... 0 — [14,14,8]

GENERATOR

Generator Random Vector [100]

[100] Op type: Fully connected — Hidden ... 784 — [784]

[784] Op type: ReLU — [784]

[784] Op type: Reshape — Shape (comma separated) 28, 28, 1 — [28,28,1]

REAL IMAGES

Inferences/sec: 158

Inference duration: 6.55ms

0 53.9%	1 46.1%		
1 65.9%	0 34.1%		
1 69.9%	0 30.1%		
1 68.7%	0 31.3%		

GENERATED IMAGES

Generations/sec: 158

Generation duration: 6.55ms

1 54.3%	0 45.7%
1 53.5%	0 46.5%
1 55.5%	0 44.5%
1 54.4%	0 45.6%

https://reiinakano.github.io/gan-playground/
(Chrome only)

Announcements

Post-doc Position

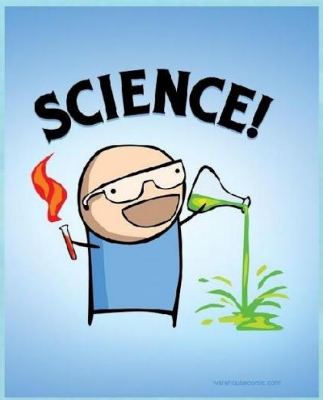
universität wien

Who are we?
Pharmacoinformatics research Group of Prof. Ecker

What are we doing?
Machine learning/ Deep learning for toxicity of chemicals/drugs

We need you!

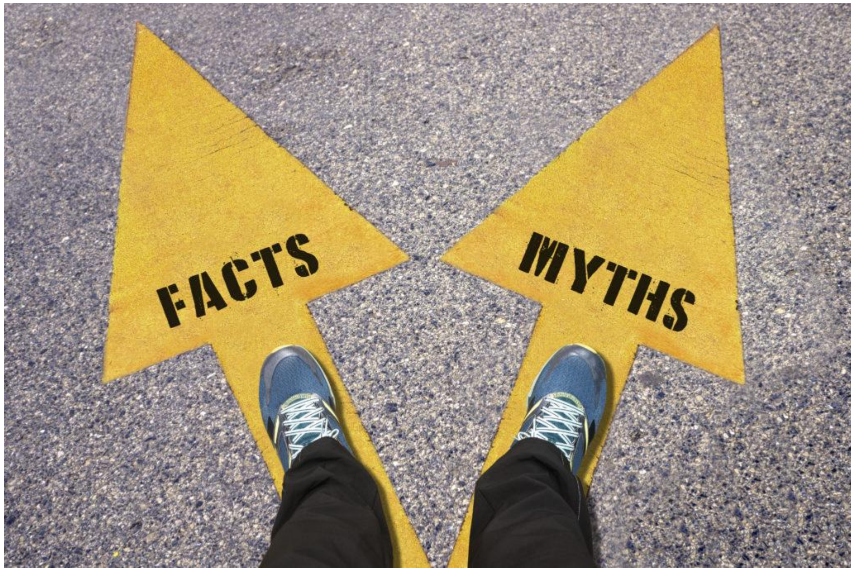
- ✓ PhD (preferably in life sciences)
- ✓ Experience in Machine learning/ Deep learning
- ✓ Scripting/Programming experience: at least 1 language (Python, R,...)



Page: https://pharminfo.univie.ac.at/
Mailto: gerhard.f.ecker@univie.ac.at

Ethics and Bias in DL / AI

- Planned for DL Meetup February or March
- Please send us contributions:
  - Papers
  - Articles
  - Speakers
  - Panelists
  - Questions?



Vienna Deep Learning Meetup

One more Hot Topic...

New Deep Learning PhD-Thesis !

Deep Learning for Event Detection, Sequence Labelling and Similarity Estimation in Music Signals

Doctoral Thesis

to obtain the academic degree of

Doktor der technischen Wissenschaften

in the Doctoral Program

Technische Wissenschaften

- **60-page introduction** to **deep learning** for audio processing
- extended versions of **publications achieving state-of-the-art result** in
  - music/speech detection
  - onset detection
  - segmentation
  - singing voice detection
  - acceleration of music similarity estimation
- previously unpublished negative results from follow-up experiments that didn't work out for each of these tasks
- source code for the singing voice detection experiments
  - https://github.com/f0k/ismir2015
  - https://github.com/f0k/ismir2015/tree/phd_extra
- source code for the wiggly lines drawing demo of ISMIR 2016
  - https://github.com/f0k/singing_horse

Dr. Jan Schlüter

Congratulations!

JKU
JOHANNES KEPLER
UNIVERSITY LINZ

Next

Vienna Deep Learning Meetup

9th January 2018

weXelerate

Thomas Lidy

Jan Schlüter

Alex Schindler