Addressing scientific debt in deep learning research: 4 aspects of our neural nets that made us wonder, and what we learned

Zürcher Hochschule
für Angewandte Wissenschaften

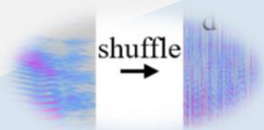
Colloquium of the UZH/ETH Institute of Neuroinformatics, March 18, 2022

Thilo Stadelmann

Agenda

1. Computer vision architecture design

2. Bias in face recognition



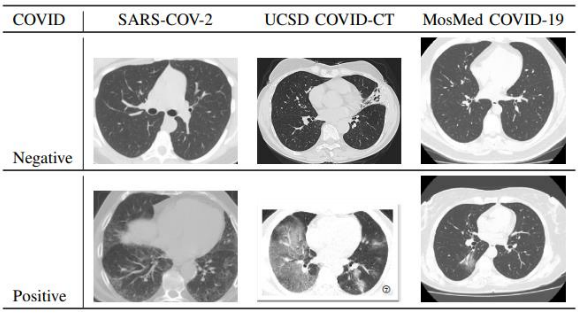
3. Dynamic speech features and sequence learners

4. Communication in multi-agent RL

5. Discussion

Zürcher Hochschule
für Angewandte Wissenschaften

We created a number of practical deep learning applications over the years…

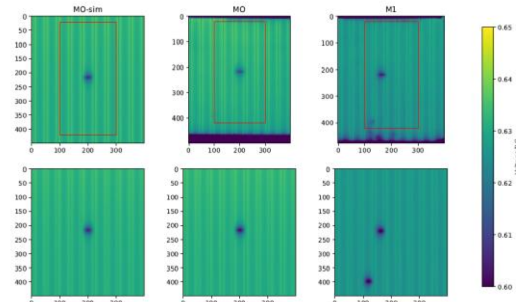


Medical imaging:
domain adaptation for diagnosis

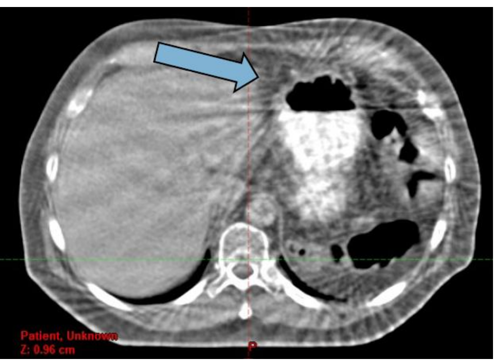
Document analysis:
article segmentation

Document analysis:
optical music recognition

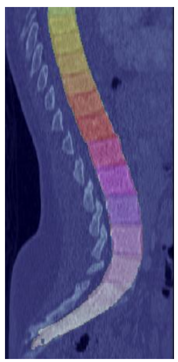
Industrial vision:
quality control



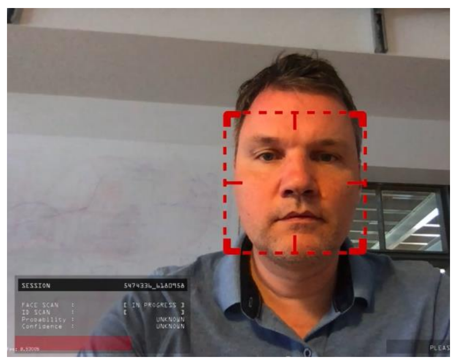
Industrial vision: prediction of solar cell simulation parameters from a real-world picture



Medical imaging:
motion artifact reduction

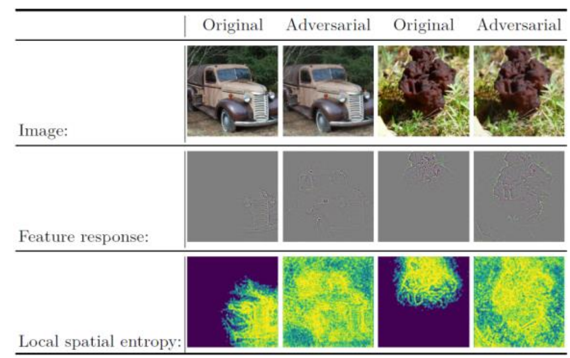


Medical imaging:
vertebrae detection



Biometrics:
robust face recognition

Industrial vision: food waste segmentation



Industrial vision: explainability and adversarial attack detection

Is ImageNet a good basis for deriving CNN architectures for other use cases?

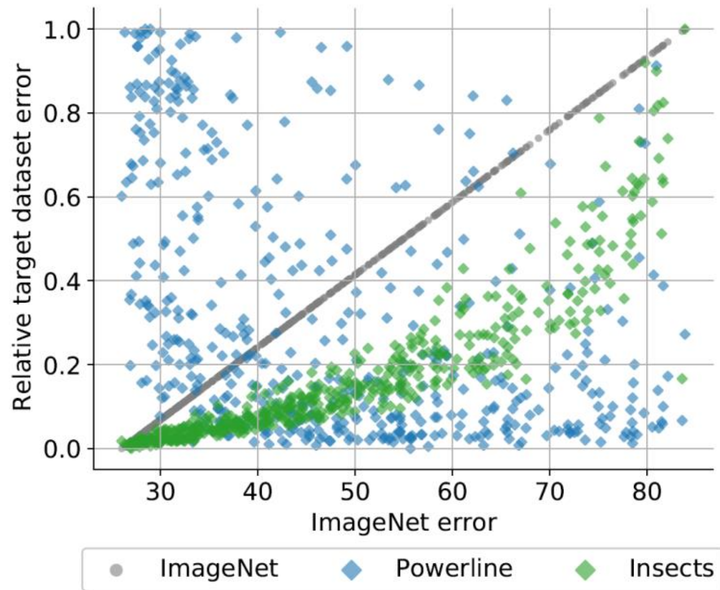


Fig. 1. Is a CNN *architecture* that performs well on ImageNet automatically a good choice for a different vision dataset? This plot suggests otherwise: It displays the relative test errors of 500 randomly sampled CNN architectures on three datasets (ImageNet, Powerline, and Insects) plotted against the test error of the same architectures on ImageNet. The architectures have been trained from scratch on all three datasets. Architectures with low errors on ImageNet also perform well on Insects, on Powerline the opposite is the case.

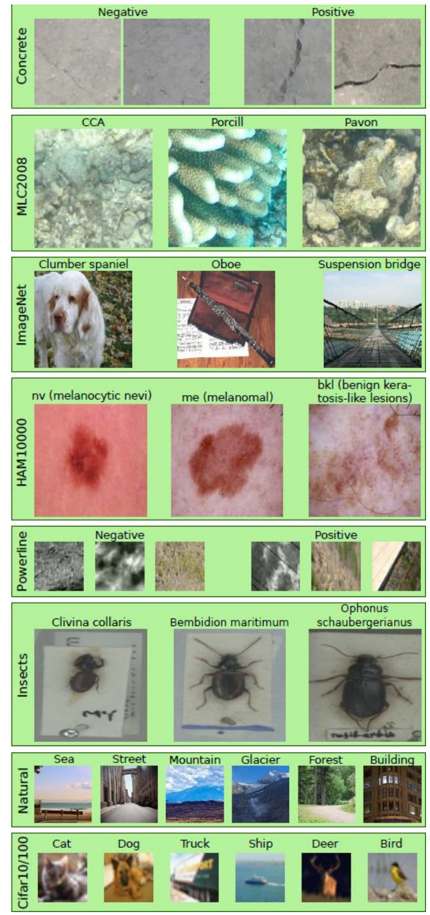
Tuggener, Schmidhuber & Stadelmann: *„ImageNet as a Representative Basis for Deriving Generally Effective CNN Architectures"*, under review, 2022.

Is ImageNet… (contd.)
Study design and results

- 500 randomly sampled architectures from the AnyNetX family (incl. AlexNets, VGGs, ResNets, RegNets)
- Trained from scratch on ImageNet and 8 relevant real-world datasets

DATASET	NO. IMAGES	NO. CLASSES	IMG. SIZE
CONCRETE	40K	2	227 × 227
MLC2008	43K	9	312 × 312
IMAGENET	1.3M	1000	256 × 256
HAM10000	10K	7	296 × 296
POWERLINE	8K	2	128 × 128
INSECTS	63K	291	296 × 296
NATURAL	25K	6	150 × 150
CIFAR10	60K	10	32 × 32
CIFAR100	60K	100	32 × 32

- Tested on (a) a test set from ImageNet and (b) on the same type used for training
- Extensive ablation studies to show validity



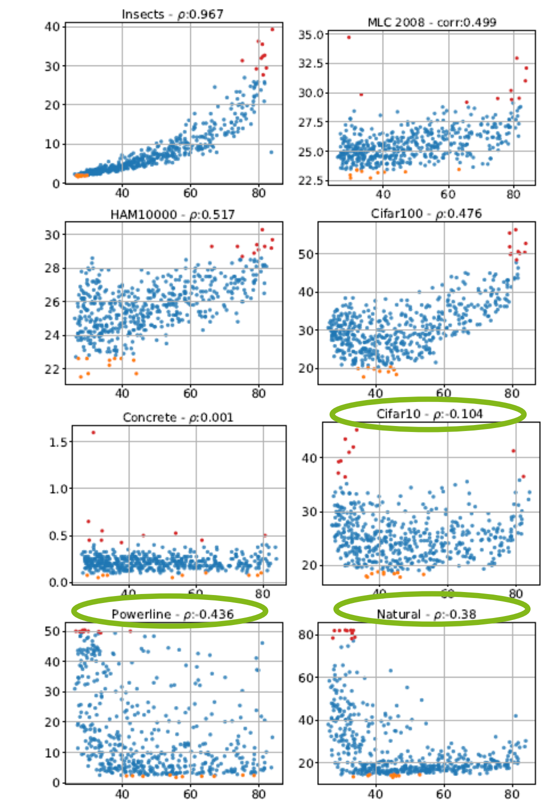


Fig. 4. Test errors of all 500 sampled architectures on target datasets (y-axis) plotted against the test errors of the same architectures (trained and tested) on ImageNet (x-axis). The top 10 performances on the target datasets are plotted in orange and the worst 10 performances in red.

Is ImageNet… (contd.)
Findings

- **Architecture search based on ImageNet** performance **is worse than random search** for at least Natural, Powerline and Cifar10

- **Varying the number of classes in ImageNet** is a cheap and **effective remedy** (i.e., randomly selecting x classes and deleting the rest of the dataset → ImageNet-x)
- …whereas **image-similarity or image size** play **not** an **important** role (e.g., Natural images are most similar to ImageNet's)

- **Hyperparameters cumulative block depth and** cumulative block **width** can **drastically change based on dataset** and are influenced by class count

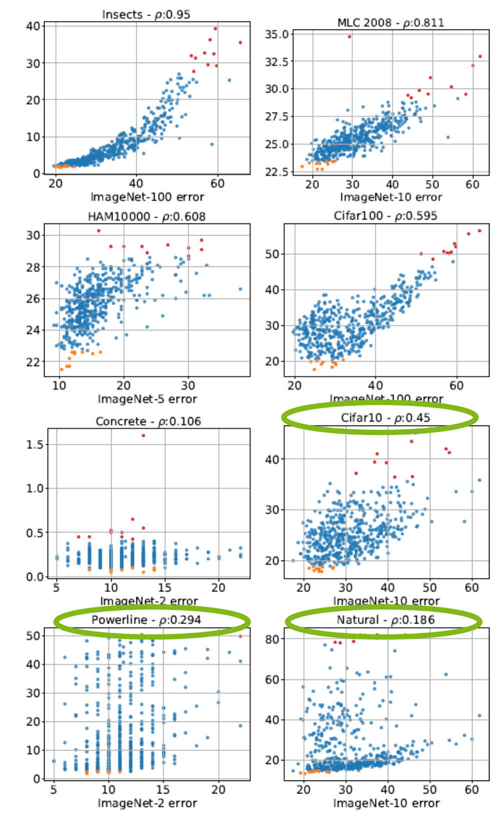
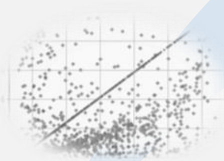


Fig. 6. Test errors of all 500 sampled architectures on target datasets (y-axis) plotted against the test errors of the same architectures on the ImageNet-X (x-axis). The top 10 performances on the target dataset are orange, the worst 10 performances red.

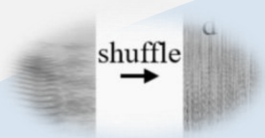
Zürcher Hochschule
für Angewandte Wissenschaften

Agenda

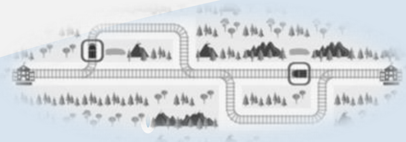


1. Computer Vision architecture design

2. Bias in face recognition



3. Dynamic speech features and sequence learners



4. Communication in multi-agent RL

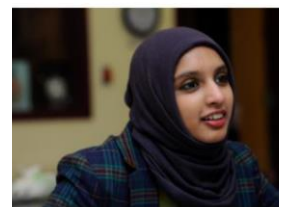
5. Discussion

The problem of bias in face recognition

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

Bias := **different recognition** rates **for** different **sub-groups** of the population, **with** potential **negative effects** for members of disadvantaged groups

Sri Lankan police wrongly identify Brown University student as wanted suspect in terror attack



Different error types, e.g., in a policing application (comparison to suspects):

- Mostly *false positives* for non-whites → **wrongful arrest**
- Mostly *false negatives* for whites → **wrongful letting go**

Buolamwini & Gebru. *«Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification»*. PMLR 2018.

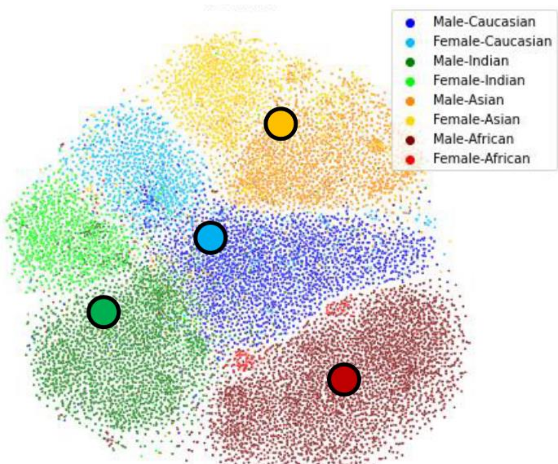
The veil of ignorance for humans and machines

Wehrli, Hertweck, Amirian, Glüge & Stadelmann. *«Bias, awareness, and ignorance in deep-learning-based face recognition»*. AI and Ethics, 2021

Zürcher Hochschule
für Angewandte Wissenschaften

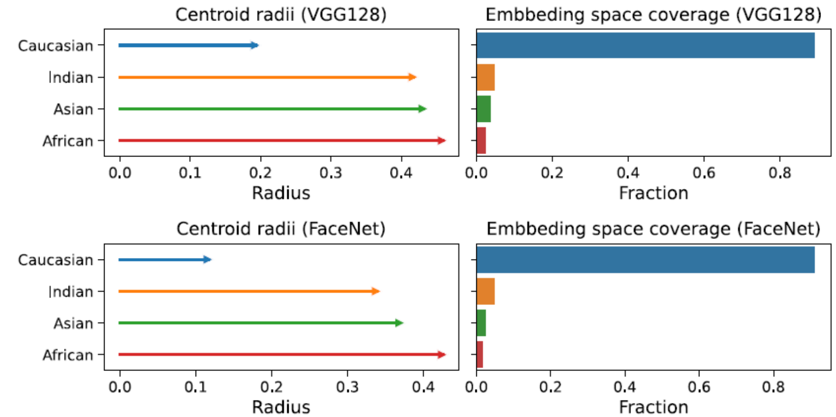
Looking inside the embedding space

Non-Caucasians are more remote and cover little space



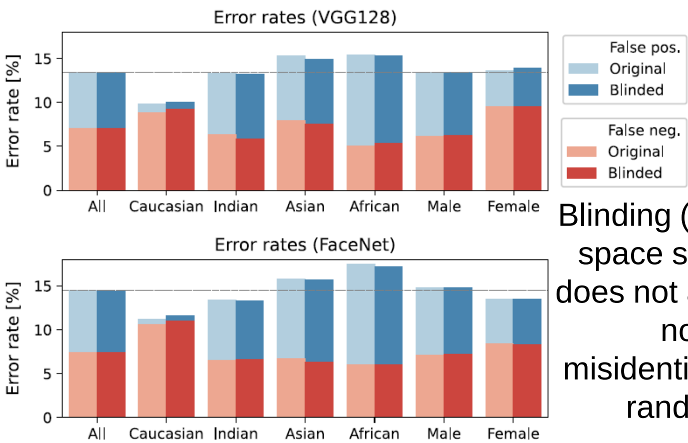
Clusters well → model is very aware of gender/ethnicity

Fig. 5 Left side: radii of the origin of the embedding space to centroids of the clusters related to ethnicity. Right side: coverage of the embedding space associated with ethnicity. This is calculated by classifying (centroid classifier) randomly generated embedding vectors. The fractions add up to 1



→ 2 Caucasian faces have greater distance

→ more often classified as „different“ when using the same decision threshold



Blinding (projecting out the sub-space spanned by centroids) does not affect recognition rates nor bias because misidentifications are scattered randomly in the space

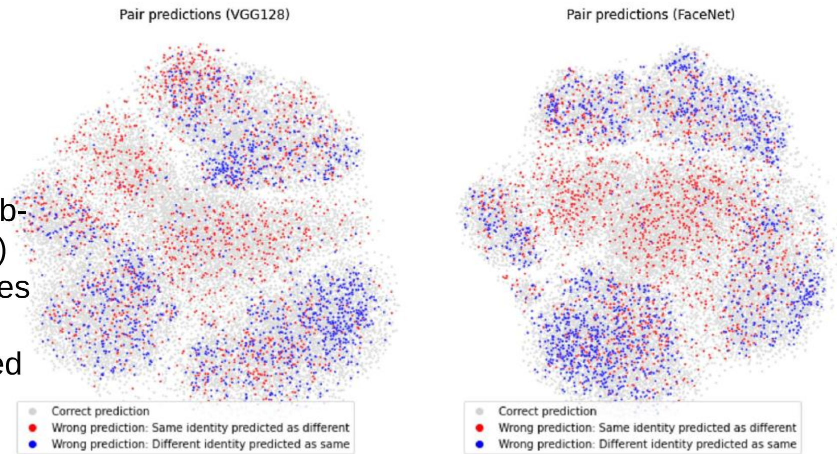
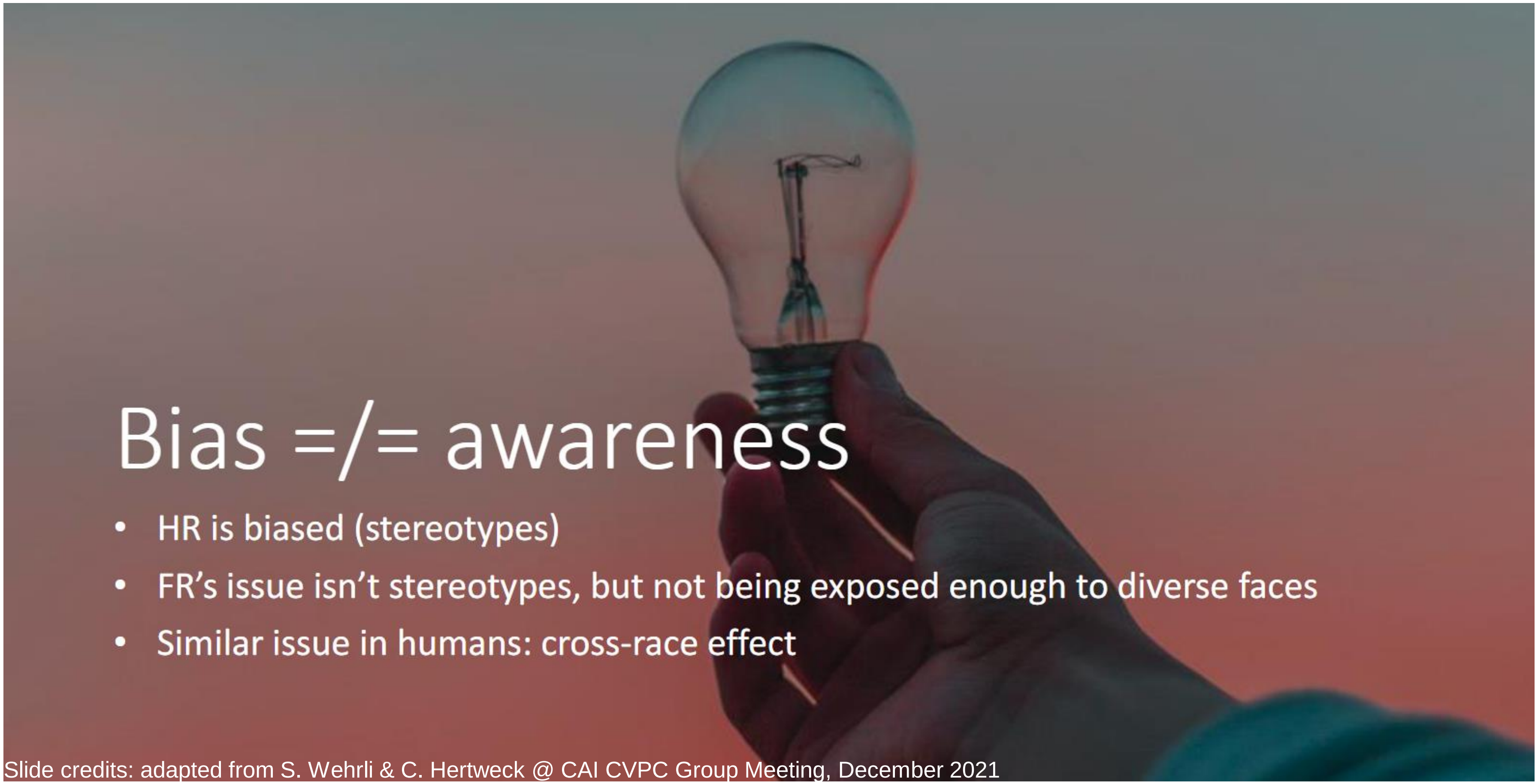


Fig. 6 Relative face recognition error rates for the VGG128 and FaceNet models. The error rates are given for *all* image pairs, for the different ethnic groups (*Caucasian, Indian, Asian,* and *African*) as well as for gender (*male, female*). The horizontal line helps to indicate whether a specific group performs better or worse than the overall average. The colors distinguish the two types of errors: False positives (blue) are pairs of different identities which are mistakenly predicted as identical, whereas false negatives (red) are identical faces mistakenly predicted as different. The brightness indicates the type of embedding. Light: original embeddings. Dark: blinded embedding (color figure online)

Fig. 7 This figure is derived from the t-SNE coordinates shown in Fig. 3. Each point in the plot represents the average coordinates of a pair, either positive with the same identity or negative with different identity. The grey points represent correct predictions of positive (same identity) or negative (different identity) pairs by the face recognition algorithm. Red points are the cases where positive pairs are mistakenly classified as negative pair. Blue points are the cases where negative pairs are mistakenly classified as positive pair. Left: VGG128 model. Right: FaceNet model (color figure online)

Bias in face recognition: results

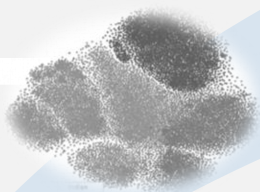


Slide credits: adapted from S. Wehrli & C. Hertweck @ CAI CVPC Group Meeting, December 2021

Zürcher Hochschule
für Angewandte Wissenschaften

Agenda

1. Computer Vision architecture design



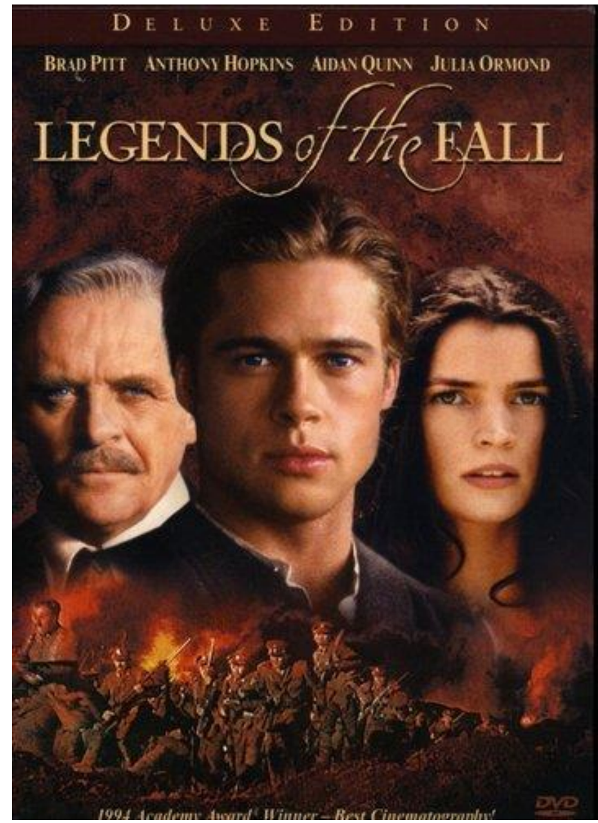
2. Bias in face recognition

3. Dynamic speech features and sequence learners

4. Communication in multi-agent RL

5. Discussion

Automatic speaker recognition performance in different scenarios



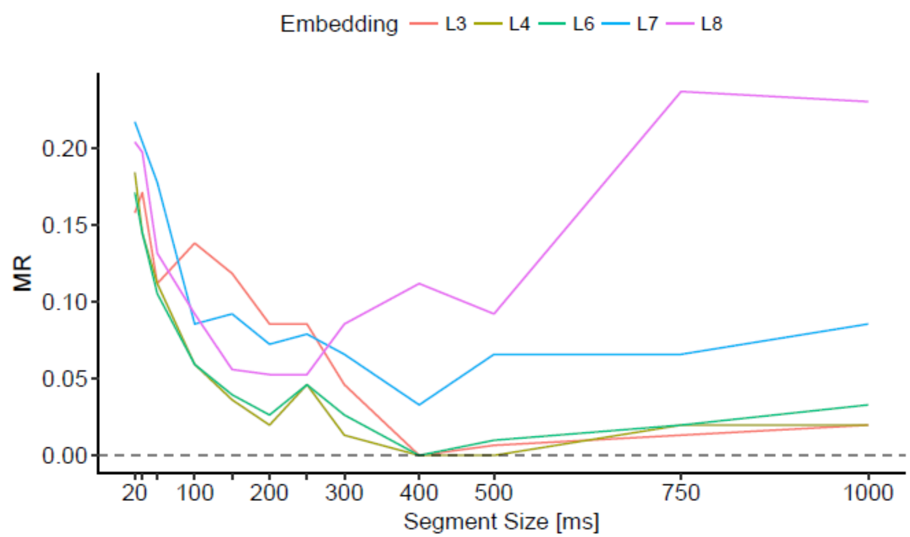
ok

not ok

Literature: in difficult environments, dynamic (temporal) voice features hold important cues

- Stadelmann & Freisleben, ACM MM 2009: **predicted** that speaker recognition **errors to improve one order of magnitude** if temporal aspects (~400ms context) are exploited
- Lukic et al., IEEE MLSP 2016/17: realized predicted effect with CNN
- Stadelmann et al., ANNPR 2018: realized predicted effect with RNN

Method	MR	MR (legacy)
RNN /w PKLD	**2.19%** $\left(\frac{1.25\%+2.5\%+1.25\%+3.75\%}{4}\right)$	**4.38%** (average of 4 runs)
CNN /w PKLD [24]	-	**5%**
CNN /w cross entropy [23]	-	**5%**
ν-SVM [40]	6.25%	-
GMM/MFCC [40]	12.5%	-



Quantifying to which extent DNNs use supra-segmental temporal information

Assumption

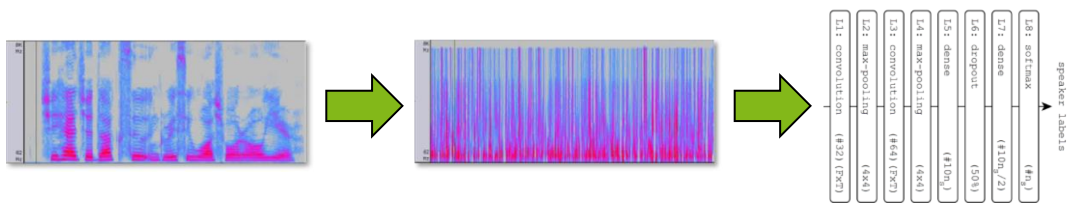
- DNNs are superior voice models ***because*** they model supra-segmental temporal (**SST**) aspects

Evidence

- The **ability is there in principle**: CNNs can use filters along the temporal axis of spectrograms; RNNs have in-built sequence modelling capabilities
- The achieved **results resemble closely the predicted improvements** when modeling temporal aspects: increase in recognition rate, optimal length of temporal context

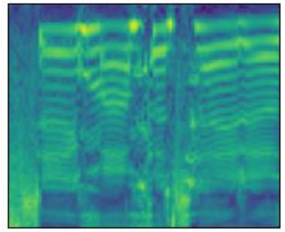
Test

- What happens if we **scramble the time axis** of a spectrogram as a preprocessing to DNN input?

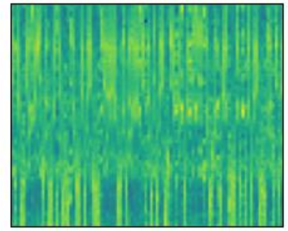


- Rationale: if the sequence of frames is random, the **only usable information are frame-based** acoustic cues (**FBA**) => the **recognition should become worse**, confirming proper exploitation of SSTs

OT



RS



Zürcher Hochschule
für Angewandte Wissenschaften

Setup

METHODOLOGY

3 DNNs: **LUVO** (Lukic, Vogt et al., 2016/17), **LSTM** (Stadelmann et al., 2018) and **ResNet34s** (Xie et al., 2019)

Training details

- **CosFace loss** (Wang et al, 2018) instead of PKLD for computational efficiency and larger margins
- **Per epoch** (64x): draw 1s segment from random starting point from each utterance; batch size 100

Evaluation

- **Evaluate speaker clustering** with Misclassification rate (***MR***) and **speaker verification** with ***EER***
- **Utterance representation**: 1s segments w/ 50% overlap → average over resulting embeddings

EXPERIMENTS

TIMIT dataset

- 630 speakers, studio conditions, 10 sentences/speaker
- Training set: 462 speakers (8 sentences train, 2 val)
- Test set: 168 speakers (10 sentences)

Setup

- As **similar** as possible **to prior work** (2009-2018)
- **Train** each DNN with **original** (**OT**) or **randomized** (**RS**) time axis
- **Evaluate** each trained model **with OT** and **RS** segments
- **Clustering**: hierarchical clustering of 2 utterances (8 or 2 concatenated sentences) per speaker (40 speakers)
- **Verification**: for all test speakers & each sentence: selected 2 matched & 2 unmatched random sentences

Neururer et al. (2022). *«Explaining the (In-)effectiveness of DNNs to Learn Supra-Segmental Temporal Features for Automatic Speaker Recognition»*. Under review.
Stadelmann & Freisleben (2009). *«Unfolding Speaker Clustering Potential: A Biomimetic Approach»*. ACMMM'2009.
Lukic, Vogt, Dürr & Stadelmann (2016). «Speaker Identification and Clustering using Convolutional Neural Networks». MLSP'2016.
Lukic, Vogt, Dürr & Stadelmann (2017). «Learning Embeddings for Speaker Clustering based on Voice Equality». MLSP'2017.
Stadelmann, Glinski-Haefeli, Gerber & Dürr (2018). «Capturing Suprasegmental Features of a Voice with RNNs for Improved Speaker Clustering». ANNPR'2018.
Xie, Nagrani, Chung & Zisserman: *"Utterance-level Aggregation for Speaker Recognition in the Wild"*. ICASSP 2019.
Wang, Wang, Zhou, Ji, Gong, Zhou, ... & Liu: *"Cosface: Large margin cosine loss for deep face recognition."* CVPR 2018.

Results

Speaker clustering on TIMIT
(MR, averaged over 5 runs)

		H50	
		OT	RS
LUVO	OT	**0.00 σ0.00**	9.00 σ2.15
	RS	9.00 σ1.66	1.25 σ0.00
LSTM	OT	1.25 σ1.12	2.75 σ0.50
	RS	2.00 σ1.00	0.25 σ0.50
RESNET34S	OT	**1.00 σ0.94**	11.50 σ4.29
	RS	2.75 σ0.94	**1.00 σ0.94**

Speaker verification on TIMIT
(EER, averaged over 5 runs)

		H50	
		OT	RS
LUVO	OT	6.38 σ0.12	11.90 σ0.46
	RS	8.16 σ0.42	5.78 σ0.16
LSTM	OT	**3.53 σ0.07**	3.90 σ0.12
	RS	4.00 σ0.07	3.54 σ0.05
RESNET34S	OT	**4.96 σ0.19**	9.21 σ1.15
	RS	5.89 σ0.25	5.80 σ0.11

Testing if DNNs can be forced to not rely on frame-based acoustic information alone

1. Make the problem acoustically harder by decreasing the SNR

Speaker verification on VoxCeleb (speech „in the wild“, 5994 speakers, 1+ mio. utterances)

		H50		
		OT	RF	RS
LUVO	OT	6.38 σ0.12	12.02 σ0.51	11.90 σ0.46
	RF	8.55 σ0.49	5.55 σ0.06	6.12 σ0.12
	RS	8.16 σ0.42	**5.33 σ0.18**	5.78 σ0.16
LSTM	OT	**3.53 σ0.07**	4.19 σ0.09	3.90 σ0.12
	RF	3.99 σ0.16	3.78 σ0.10	3.66 σ0.13
	RS	4.00 σ0.07	3.89 σ0.06	3.54 σ0.05
RESNET34S	OT	**4.96 σ0.19**	10.34 σ1.56	9.21 σ1.15
	RF	6.59 σ0.25	6.25 σ0.23	6.37 σ0.35
	RS	5.89 σ0.25	6.11 σ0.31	5.80 σ0.11

		H50		
		OT	RF	RS
LUVO	OT	**25.75 σ0.13**	37.23 σ0.74	36.96 σ0.78
	RF	32.70 σ0.34	27.04 σ0.34	27.99 σ0.30
	RS	33.26 σ0.29	27.91 σ0.32	28.50 σ0.28
LSTM	OT	**20.67 σ0.23**	30.67 σ0.36	30.00 σ0.32
	RF	26.20 σ0.18	22.02 σ0.10	23.57 σ0.09
	RS	28.28 σ1.30	26.30 σ0.59	26.58 σ0.84
RESNET34S	OT	**12.49 σ0.15**	34.11 σ0.54	32.19 σ0.39
	RF	22.05 σ0.43	19.08 σ0.26	20.02 σ0.16
	RS	20.74 σ0.46	21.02 σ0.34	20.36 σ0.23

(EER, averaged over 5 runs)

➔ Being able to exploit **SST** information **helps in the presence of more noise**

Testing if DNNs can be forced to not rely on frame-based acoustic information alone

2. Remove discriminative power of FBAs by equalizing timbre of speakers

Speaker verification on TIMIT-NV (noise-vocoded w/ original amplitude contours in 4 bands)

		H50		
		OT	RF	RS
LUVO	OT	6.38 σ0.12	12.02 σ0.51	11.90 σ0.46
	RF	8.55 σ0.49	5.55 σ0.06	6.12 σ0.12
	RS	8.16 σ0.42	**5.33 σ0.18**	5.78 σ0.16
LSTM	OT	**3.53 σ0.07**	4.19 σ0.09	3.90 σ0.12
	RF	3.99 σ0.16	3.78 σ0.10	3.66 σ0.13
	RS	4.00 σ0.07	3.89 σ0.06	3.54 σ0.05
RESNET34S	OT	**4.96 σ0.19**	10.34 σ1.56	9.21 σ1.15
	RF	6.59 σ0.25	6.25 σ0.23	6.37 σ0.35
	RS	5.89 σ0.25	6.11 σ0.31	5.80 σ0.11

		H50		
		OT	RF	RS
LUVO	OT	32.56 σ0.62	35.32 σ0.46	35.41 σ0.55
	RF	35.16 σ0.52	**30.39 σ0.30**	30.91 σ0.47
	RS	35.25 σ0.69	30.63 σ0.38	31.23 σ0.27
LSTM	OT	**19.34 σ0.16**	27.20 σ0.42	26.12 σ0.44
	RF	22.95 σ0.24	21.48 σ0.40	21.15 σ0.25
	RS	22.82 σ0.40	21.89 σ0.25	21.04 σ0.12
RESNET34S	OT	**21.12 σ0.43**	37.83 σ1.17	36.57 σ1.45
	RF	27.03 σ0.63	23.38 σ0.41	24.02 σ0.25
	RS	27.25 σ1.37	23.57 σ0.46	23.32 σ0.58

(EER, averaged over 5 runs)

➔ Being able to exploit **SST** information **helps with less speaker-discriminating FBAs**
➔ Disclaimer: not evident for speaker clustering using MR

Testing if DNNs can be forced to not rely on frame-based acoustic information alone

2. Remove discriminative power of FBAs by equalizing timbre of speakers

Speaker verification on TIMIT-Syn (re-synthesized w/ original, normalized pitch tracks and phone-level timing information from annotations [Slowsoft synthesizer, similar for MBROLA])

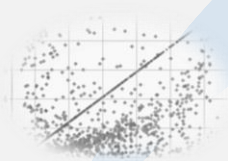
		H50		
		OT	RF	RS
LUVO	OT	6.38 σ0.12	12.02 σ0.51	11.90 σ0.46
	RF	8.55 σ0.49	5.55 σ0.06	6.12 σ0.12
	RS	8.16 σ0.42	**5.33 σ0.18**	5.78 σ0.16
LSTM	OT	**3.53 σ0.07**	4.19 σ0.09	3.90 σ0.12
	RF	3.99 σ0.16	3.78 σ0.10	3.66 σ0.13
	RS	4.00 σ0.07	3.89 σ0.06	3.54 σ0.05
RESNET34S	OT	**4.96 σ0.19**	10.34 σ1.56	9.21 σ1.15
	RF	6.59 σ0.25	6.25 σ0.23	6.37 σ0.35
	RS	5.89 σ0.25	6.11 σ0.31	5.80 σ0.11

		H50		
		OT	RF	RS
LUVO	OT	46.24 σ0.18	48.94 σ0.15	48.97 σ0.23
	RF	47.26 σ0.15	45.98 σ0.34	46.16 σ0.27
	RS	47.14 σ0.22	45.88 σ0.12	**45.66 σ0.12**
LSTM	OT	**40.39 σ0.07**	44.29 σ0.65	42.43 σ1.40
	RF	43.63 σ0.35	41.93 σ0.26	41.64 σ0.25
	RS	43.62 σ0.21	42.55 σ0.34	41.53 σ0.23
RESNET34S	OT	**40.33 σ1.32**	47.28 σ2.06	46.60 σ2.02
	RF	43.44 σ0.86	42.97 σ0.51	42.65 σ0.59
	RS	42.48 σ0.45	43.07 σ0.72	41.59 σ0.36

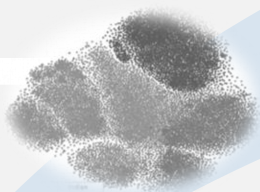
(EER, averaged over 5 runs)

➔ Being able to exploit **SST** information **helps without any speaker-discriminating FBAs**
➔ Disclaimer: less evident for speaker clustering using MR

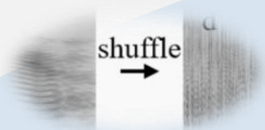
Agenda



1. Computer Vision architecture design



2. Bias in face recognition



3. Dynamic speech features and sequence learners

4. Communication in multi-agent RL

5. Discussion

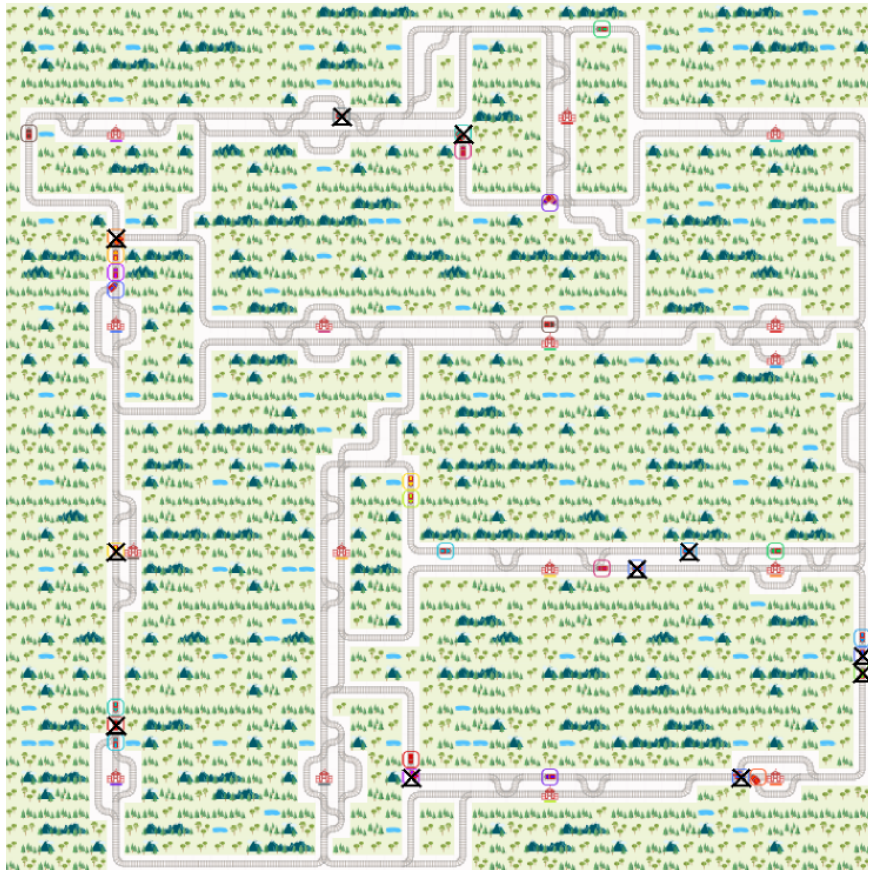
Mutli-agent RL for train rescheduling

Problem description

- How to **adjust for small delays** ("rescheduling") automatically **in a** more and more **packed railway network** like the one of SBB?

- **Closed-form optimization impossible** due to combinatorial explosion of rerouting options

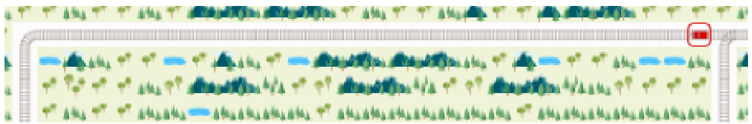
- RL still in its infancy for practical *high-consequence* environments → Flatland challenge to explore options



Lessons learned on RL in rescheduling
(based on a rank-6 entry to the Flatland challenge)

How to make RL sample-efficient:

- Using **task-specific heuristics** to present the agent with percepts only when a decision is necessary (i.e., at switches) increases the performance from 44.5% to 82.9%
- Using **curriculum learning** to learn fundamental behavior in easy environments and gradually increase complexity ensures rank 6/32 in the more realistic Flatland Round 2

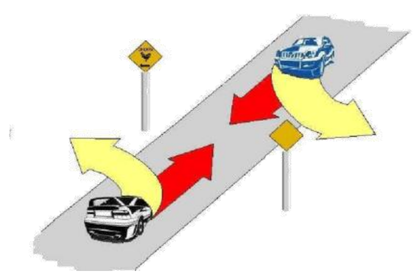


Screenshot from Flatland environment. A train heading to the left. The only reasonable action is to ride forward.

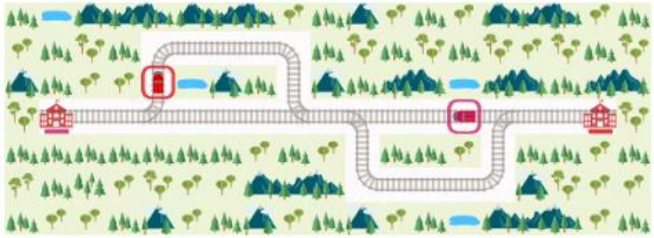
	Level 0	Level 1	Level 2	Level 3	Level 4
Next level on success rate	70%	70%	75%	70%	60%
Nr. of agent	4	8	12	16	20
Env. size	25x25	30x30	40x40	50x50	50x50
Num. cities	5	8	10	12	16
Max. rails between cities	1	2	2	2	2
Max. rails in city	2	2	3	3	3

General remark:

- **Policy gradient methods** seem **generally inappropriate** for high-consequence environments (i.e., one bad action leads to unresolvable catastrophes)
- Reason is **stochasticity**: if distributions over actions are learned and many agents are present in a single environment, the probability of having one bad action in every time step approaches certainty



An emerging machine language?



Humans would communicate to **negotiate** who would take **the detour**

What happens if we **add communication actions** (5 free tokens + EOT) and a shared **communication buffer in the observation** to the RL scenario**?**

Communication process:

1. **Communication loop** is entered upon first comm. action taken by any agent
2. Agents can sequentially **read** the comm. **buffer** and **add** a comm. **action**
3. Comm. loop **ends when** both agents issue the **EOT** action
4. **Then**, both agents can select regular (non-comm.) actions again and **proceed in the environment**

➔ **Does** the **general ability** to negotiate (i.e., exchange an arbitrary long sequence of tokens until mutually agreed to end) **help** in **practically** avoiding collision**?**

A first glimpse

Training

- Reward -1 if agents collide after negotiation; +1 otherwise
- Agents don't know who they are and need to take actions in parallel → cannot stick to go only one way or react to first mover
- 1M episodes training (A3C)

Results

- Success rate increases from 47% to 95%!
- High diversity in machine dialogues!
- (See examples on the right →)

Implications

- Allowing arbitrarily long sequences of 5 tokens can lead to Turing-completeness
- But what happens actually?

Timestep	Actions agent 1|2	Outcome
0 1	4 | 2 5 | 5	Success
0 1 2	3 | 0 1 | 5 5 | 5	Success
0 1	3 | 5 5 | 5	Success
0 1 2 3	3 | 1 3 | 2 5 | 0 5 | 5	Crash
0 1 2 3 4	3 | 2 5 | 3 5 | 4 2 | 5 5 | 5	Success
0 1 2	4 | 3 3 | 1 5 | 5	Success

Zürcher Hochschule
für Angewandte Wissenschaften

Discussion

- What puzzling aspects of your research have you so far ignored in hunt of a different goal?
- Do you think there is a lesson to learn from searching for an explanation?
- Do you think it pays off to take these detours?

- Ideas for forcing DNNs to pick up temporal patterns of a voice?
- Ideas for continuing the RL & communication work?

About us:

- Director of Centre for AI, head CVPC Group: Prof. Dr. Thilo Stadelmann
  Email: stdm@zhaw.ch
  Phone: +41 58 934 72 08
- Head NLP Group: Prof. Dr. Mark Cieliebak
  Email: ciel@zhaw.ch
  Phone: +41 58 934 72 39

Further contacts:

- info.cai@zhaw.ch, datalab@zhaw.ch, info.office@data-innovation.org, office-switzerland@claire-ai.org

Datalab
zhaw.ch/datalab

data innovation alliance

SwissNLP
Swiss Association for
Natural Language Processing

CLAIRE

Zürcher Hochschule
für Angewandte Wissenschaften

APPENDIX

Sample projects

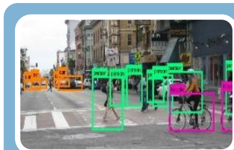
Zürcher Hochschule
für Angewandte Wissenschaften

The ZHAW Centre for Artificial Intelligence

Foundation: Machine Learning & Deep Learning
Cross-cutting concerns: Ethics, Generality

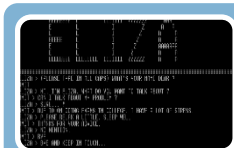
Autonomous Learning Systems

- *Reinforcement Learning*
- *Multi-Agent Systems*
- *Embodied AI*



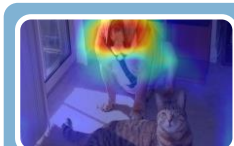
Computer Vision, Perception and Cognition

- *Pattern Recognition*
- *Machine Perception*
- *Neuromorphic Engineering*



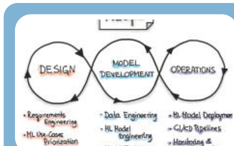
Natural Language Processing

- *Dialogue Systems*
- *Text Analytics*
- *Spoken Language Technologies*



Trustworthy AI

- *Explainable AI*
- *Robust Deep Learning*
- *AI & Society*



AI Engineering

- MLOps
- Data-Centric AI
- Continuous Learning

Areas of application & cooperation: medicine & health, IoT, robotics, AI ethics & regulation, predictive maintenance, automatic quality control, document analysis, chat bots, biometrics, earth observation, digital farming, meteorology, autonomous driving, further data science use cases in industries like manufacturing / finance / insurance / commerce / transportation / energy etc.

Education at the CAI

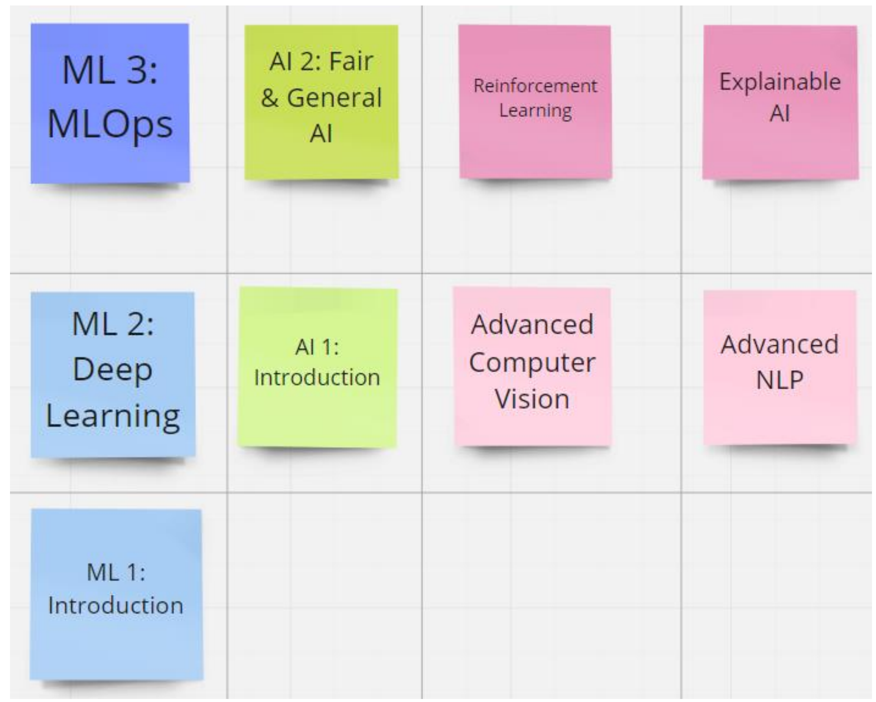
TEACHING ENGAGEMENT

- B.Sc. Computer Science & Data Science
- M.Sc. Engineering (CS, DS)
- Ph.D. in cooperation with e.g.

- Continuing education in AI & ML

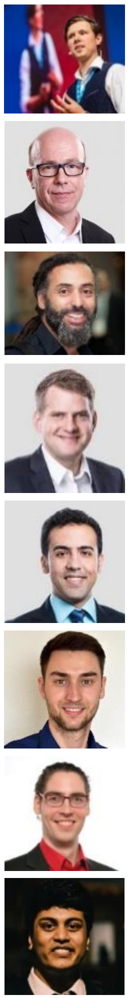
- Special mentoring program for CAI-affiliated students

UNDERGRAD PORTFOLIO

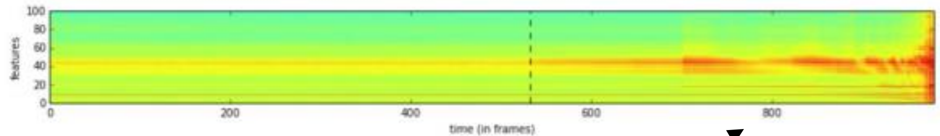


Zürcher Hochschule
für Angewandte Wissenschaften

Computer Vision, Perception & Cognition Group



Machine learning-based
Pattern Recognition

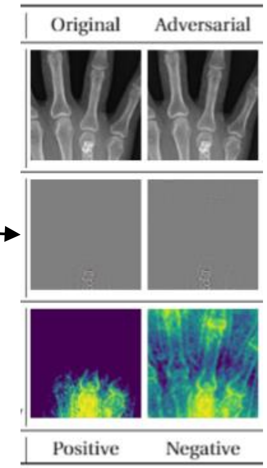


Robust applications

[3]

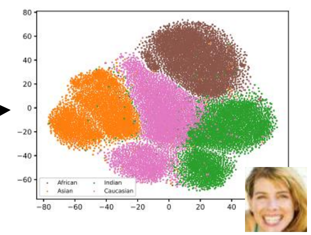
[1]

[2]

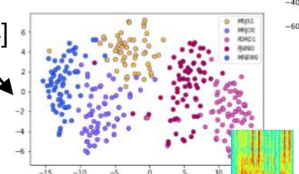


Biometrics

[5]



[4]



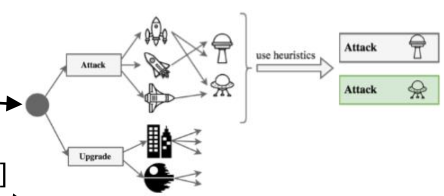
Document Analysis

[7]

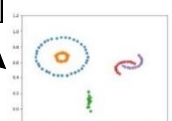
[6]

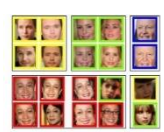
Learning to act

[1]

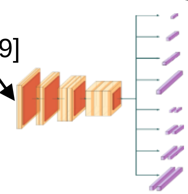


[8]

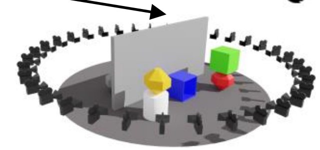




[9]



[10]



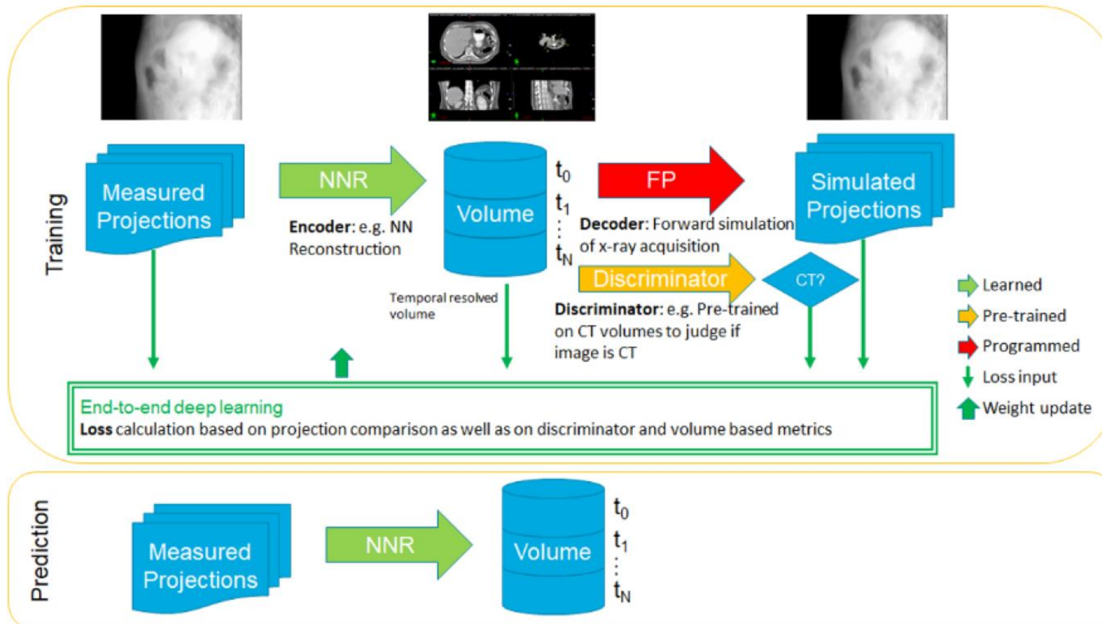
CVPC Group: references for overview

1. Thilo Stadelmann, Mohammadreza Amirian, Ismail Arabaci, Marek Arnold, Gilbert François Duivesteijn, Ismail Elezi, Melanie Geiger, Stefan Lörwald, Benjamin Bruno Meier, Katharina Rombach, and Lukas Tuggener. "**Deep Learning in the Wild**". In: Proceedings of the 8th IAPR TC 3 Workshop on Artificial Neural Networks for Pattern Recognition (**ANNPR'18**), Springer, LNAI 11081, pp. 17-38, Siena, Italy, September 19-21, 2018.
2. Mohammadreza Amirian, Friedhelm Schwenker, and Thilo Stadelmann. "**Trace and Detect Adversarial Attacks on CNNs using Feature Response Maps**". In: Proceedings of the 8th IAPR TC 3 Workshop on Artificial Neural Networks for Pattern Recognition (**ANNPR'18**), Springer, LNAI 11081, pp. 346-358, Siena, Italy, September 19-21, 2018.
3. Thilo Stadelmann, Vasily Tolkachev, Beate Sick, Jan Stampfli, and Oliver Dürr. "**Beyond ImageNet - Deep Learning in Industrial Practice**". In: Martin Braschler, Thilo Stadelmann, and Kurt Stockinger (Editors). **"Applied Data Science - Lessons Learned for the Data-Driven Business"**. **Springer**, 2019.
4. Thilo Stadelmann, Sebastian Glinski-Haefeli, Patrick Gerber, and Oliver Dürr. "**Capturing Suprasegmental Features of a Voice with RNNs for Improved Speaker Clustering**". In: Proceedings of the 8th IAPR TC 3 Workshop on Artificial Neural Networks for Pattern Recognition (**ANNPR'18**), Springer, LNAI 11081, pp. 333-345, Siena, Italy, September 19-21, 2018.
5. Stefan Glüge, Mohammadreza Amirian, Dandolo Flumini, and Thilo Stadelmann. "**How (Not) to Measure Bias in Face Recognition Networks**". In: Proceedings of the 9th IAPR TC 3 Workshop on Artificial Neural Networks for Pattern Recognition (**ANNPR'20**), Springer, LNAI, Winterthur, Switzerland, September 02-04, 2020.
6. Lukas Tuggener, Yvan Putra Satyawan, Alexander Pacha, Jürgen Schmidhuber, and Thilo Stadelmann. "**The DeepScoresV2 Dataset and Benchmark for Music Object Detection**". In: Proceedings of the 25th International Conference on Pattern Recognition (**ICPR'20**), IAPR, Milan, Italy, January 10-15 (online), 2021.
7. Benjamin Meier, Thilo Stadelmann, Jan Stampfli, Marek Arnold, and Mark Cieliebak. "**Fully convolutional neural networks for newspaper article segmentation**". In: Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition (**ICDAR'17**). 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto Japan, November 13-15, 2017. Kyoto, Japan: CPS.
8. Benjamin Bruno Meier, Ismail Elezi, Mohammadreza Amirian, Oliver Dürr, and Thilo Stadelmann. "**Learning Neural Models for End-to-End Clustering**". In: Proceedings of the 8th IAPR TC 3 Workshop on Artificial Neural Networks for Pattern Recognition (**ANNPR'18**), Springer, LNAI 11081, pp. 126-138, Siena, Italy, September 19-21, 2018.
9. Lukas Tuggener, Mohammadreza Amirian, Fernando Benites, Pius von Däniken, Prakhar Gupta, Frank-Peter Schilling, and Thilo Stadelmann. "**Design Patterns for Resource-Constrained Automated Deep-Learning Methods**". **AI** section "Intelligent Systems: Theory and Applications" 1(4):510-538, MDPI, Basel, Switzerland, Novemer 06, 2020.
10. Dano Roost, Ralph Meier, Giovanni Toffetti Carughi, and Thilo Stadelmann. "**Combining Reinforcement Learning with Supervised Deep Learning for Neural Active Scene Understanding**". In: Proceedings of the Active Vision and Perception in Human(-Robot) Collaboration Workshop at IEEE RO-MAN 2020 (**AVHRC'20**), online, August 31, 2020.

DIR3CT: Deep Image Reconstruction through X-Ray Projection-based 3D Learning of Computed Tomography Volumes

Collaboration with Inst. of Appl. Math. & Physics

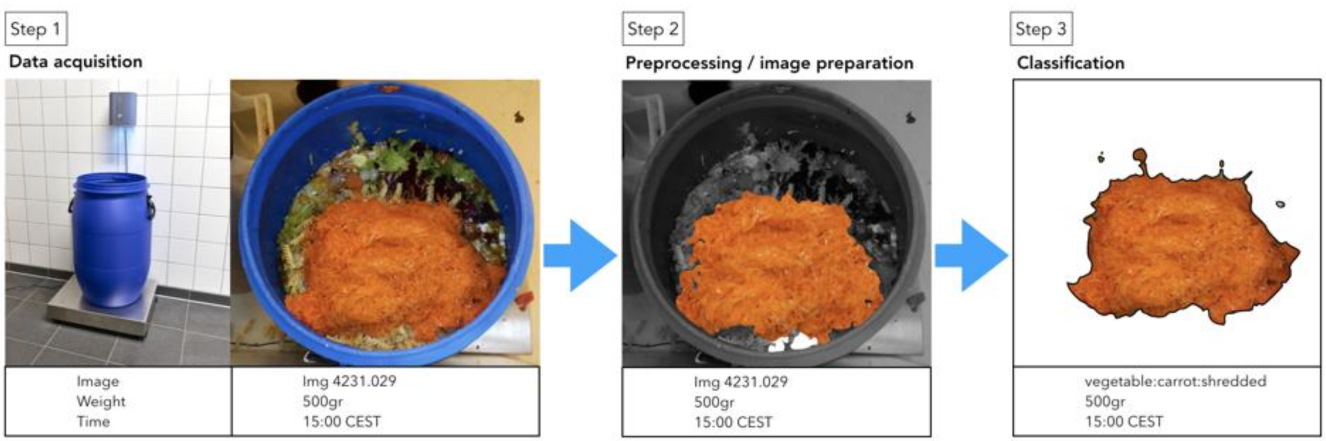
- Topic: Compensation of motion artefacts in 3D CBCT reconstructed volumes using deep learning
- InnoSuisse, total volume **1.13 MCHF**
- Duration: 02/2020 – 05/2022
- Industry partner **Varian Medical Systems** (world market leader radiation therapy)
- Two involved ZHAW institutes **CAI & IAMP** (approx. 8 ZHAW researchers involved)
  - Focus CAI: 3D reconstruction using deep learning (supervised & unsupervised)
  - Focus IAMP: Physical modeling and simulation of motion, anatomical constraints
- Highly ambitious and technologically challenging



Zürcher Hochschule
für Angewandte Wissenschaften

Food Waste Analysis
Collaboration with the Inst. Of Embedded Systems

- Automatic detection of food waste in restaurants
- Embedded Machine Learning for waste classification
- Savings potential: At least CHF 2,500 per month per kitchen
- Research: Embedded System Design with GPU Edge Processing, Automatic Food Waste Classification
- Joint Innosuisse Project InES / CAI, July 19 – Aug 21



KITRO

www.kitro.ch

Zürcher Hochschule
für Angewandte Wissenschaften

ADA: Automated Data Analyst
Collaboration with EPFL MLO Lab

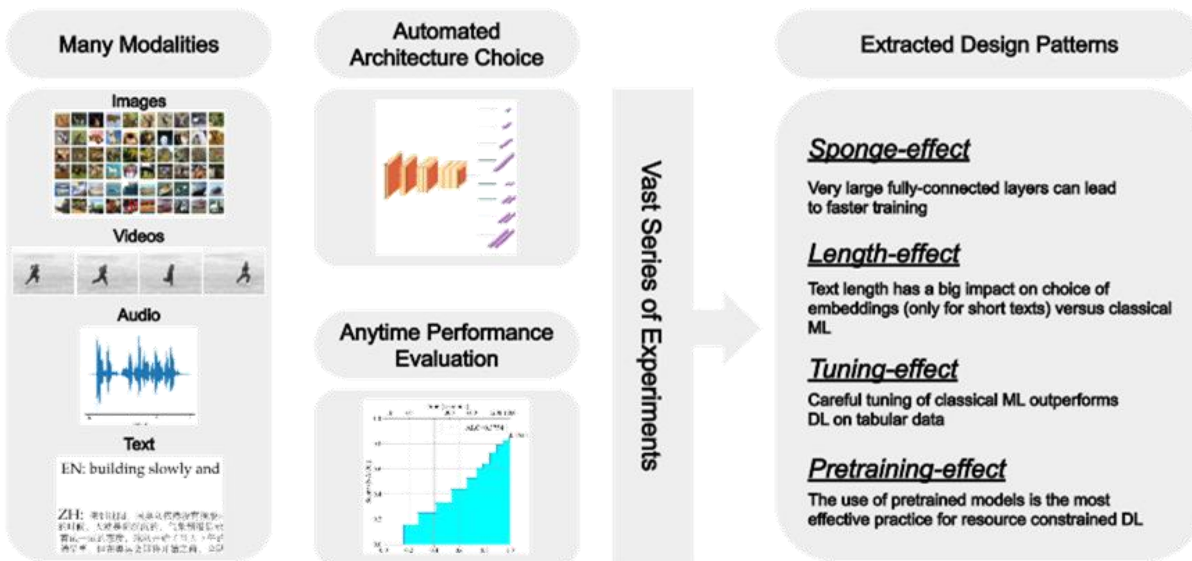
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Innosuisse – Swiss Innovation Agency

The project
- Target: in-house solution of industrial partner to improve turnover in standard analytics projects
- Challenge: optimize hyperparameters smarter than with well initialized random perturbations
- Result: top ranks in Google AutoDL'2020 competition

Design Patterns for Resource Constrained Automated Deep Learning Methods



Stadelmann, Amirian, Arabaci, Arnold, Duivesteijn, Elezi, Geiger, Lörwald, Meier, Rombach & Tuggener (2018). *«Deep Learning in the Wild»*. ANNPR'2018.
Tuggener, Amirian, Rombach, Lörwald, Varlet, Westermann & Stadelmann (2019). *«Automated Machine Learning in Practice: State of the Art and Recent Results»*. SDS'19.
Tuggener, Amirian, Benites, von Däniken, Gupta, Schilling & Stadelmann (2020). *«Design Patterns for Resource-Constrained Automated Deep-Learning Methods»*. AI 1(4).

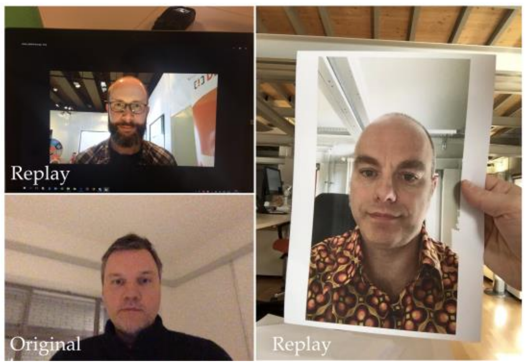
LIBRA: Face matching & anti-spoofing
Collaboration with Inst. of Appl. Math. & Physics

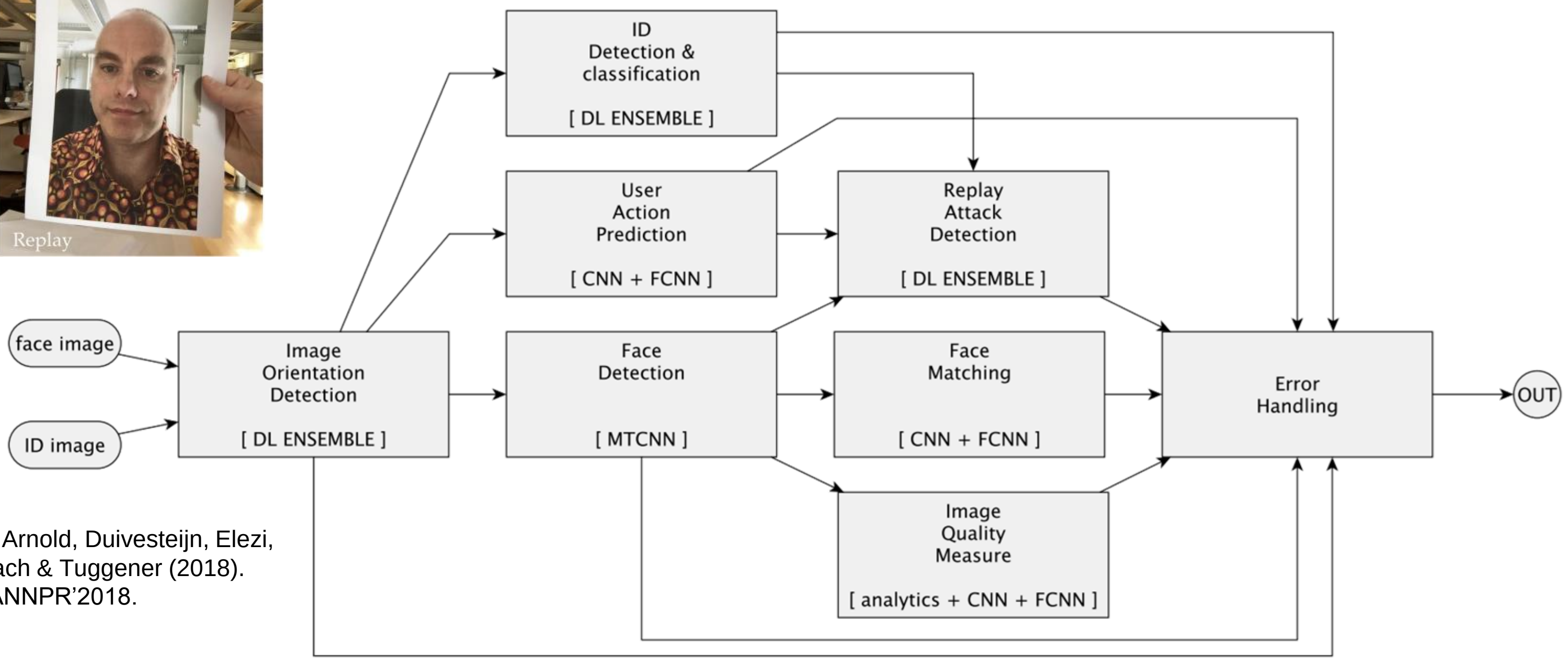
[!] DEEPIMPACT

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Innosuisse – Swiss Innovation Agency





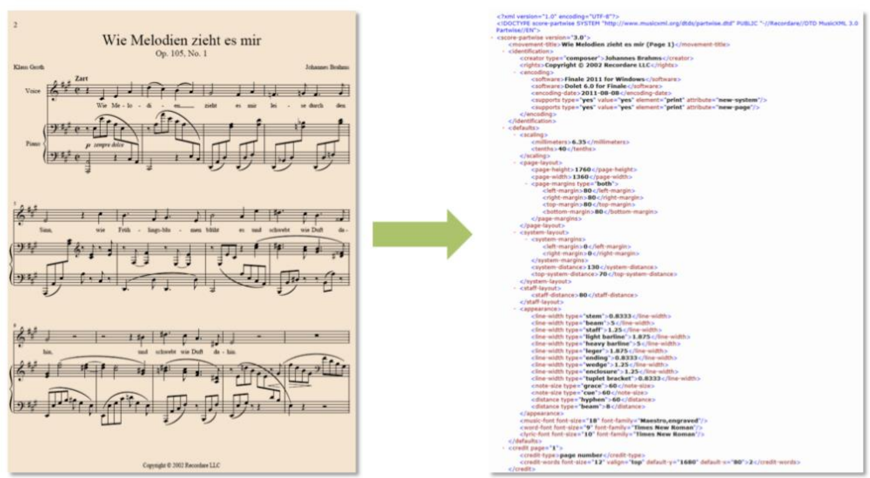
Stadelmann, Amirian, Arabaci, Arnold, Duivesteijn, Elezi, Geiger, Lörwald, Meier, Rombach & Tuggener (2018). *«Deep Learning in the Wild»*. ANNPR'2018.

DeepScore – Music OCR via Deep Neural Nets
Collaboration with IDSIA

Goal: Raise the accuracy of optical music recognition (OMR) by one order of magnitude to facilitate paper-free work of professional musicians

Challenge: Transfer the recent success of deep learning methods on numerous pattern recognition tasks (e.g., OCR) to the domain of music notation (which is 2D, without benchmarks, many syntactical constraints)

Solution: Enhance the open music scanner Audiveris by a new symbol classifier and segmenter based on convolutional neural networks to output musicXML



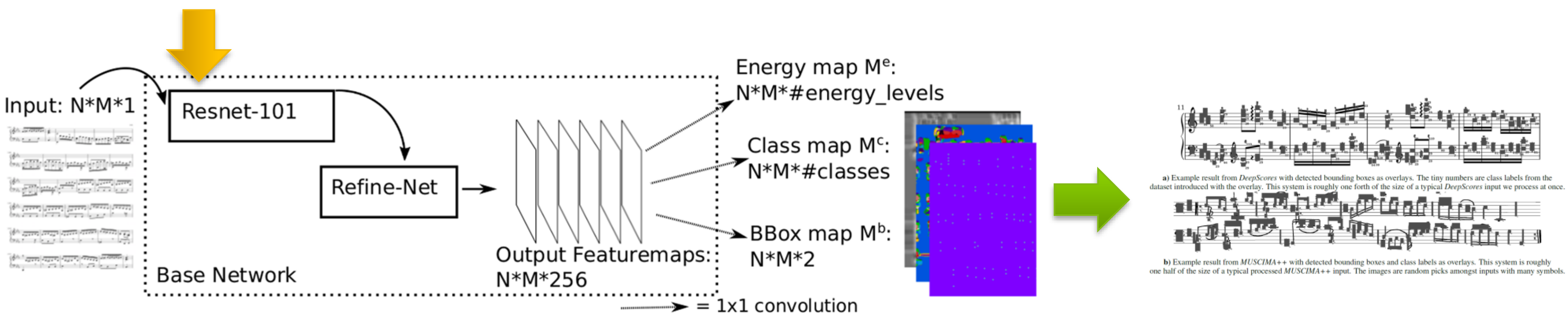
DeepScore – challenges & solutions

SCOREPAD

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

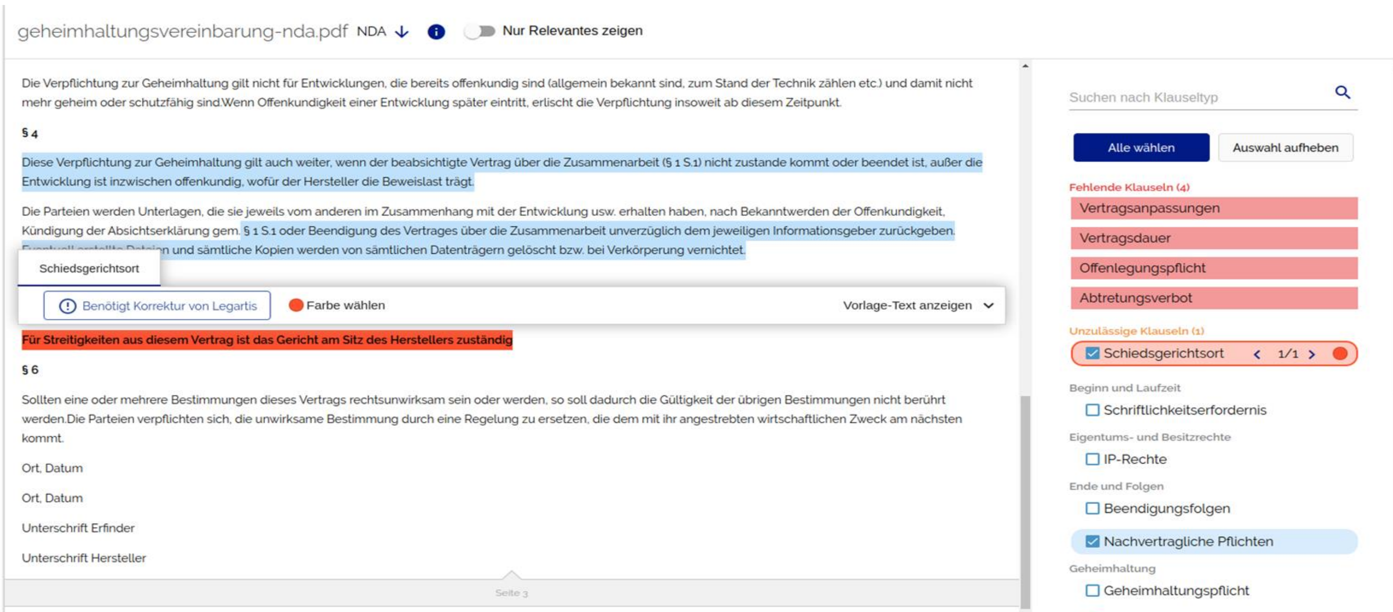
Swiss Confederation

Innosuisse – Swiss Innovation Agency



Tuggener, Elezi, Schmidhuber, Pelillo & Stadelmann (2018). *«DeepScores – A Dataset for Segmentation, Detection and Classification of Tiny Objects»*. ICPR’2018.
Tuggener, Elezi, Schmidhuber & Stadelmann (2018). *«Deep Watershed Detector for Music Object Recognition»*. ISMIR’2018.
Tuggener, Satyawan, Pacha, Schmidhuber & Stadelmann (2020). *«The DeepScoresV2 Dataset and Benchmark for Music Object Detection»*. ICPR’2020.

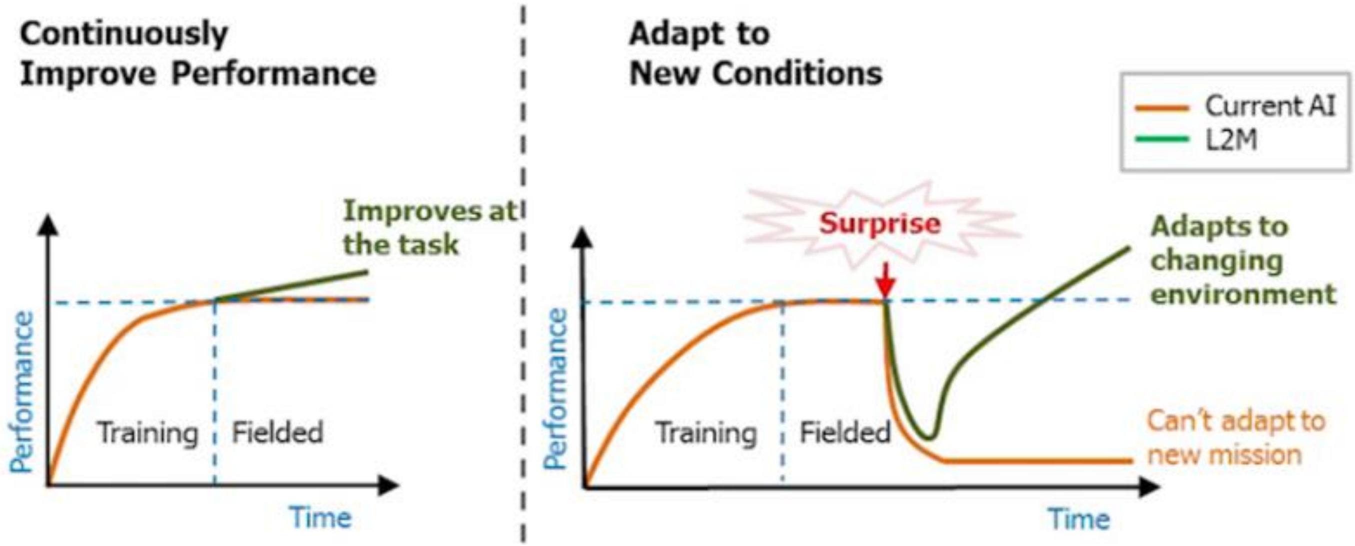
SCAI: Smart Contract Analytics using Artificial Intelligence



Innosuisse project (480'000 CHF)

Multi-label text classification: Classify contractual provisions (120+ labels)
Outlier detection: Find problematic provisions
Entity recognition: Detect companies, costs, penalties, jurisdiction etc.
Multilingual: EN, DE

LIHLITH: Lifelong Learning for Dialogue Systems



- EU CHIST-ERA and SNF project (220'000 CHF)
  - Fundamental research project
  - What happens with dialogue systems after deployment? How does it learn new things continuously and autonomously?
  - How to react when the algorithm is confronted with an unknown situation?
  - Our contribution: benchmark to evaluate lifelong machine learning for natural language interfaces to databases

QualitAI

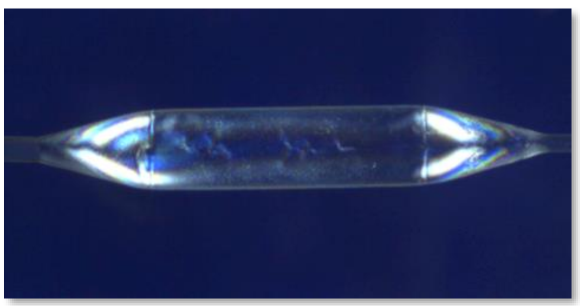
Optical Quality Control for MedTech Products

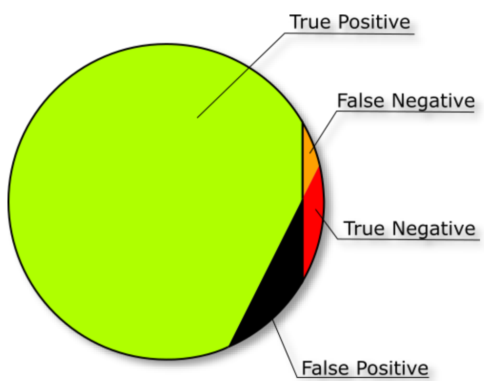
Goal: semi-automatic quality control of industrial goods with computer vision

Challenge: Work with small amounts of imbalanced data

Approach:

- Use state-of-the art deep learning models
- Use transfer learning, few-shot learning, image improvement to enable small data app





Zürcher Hochschule
für Angewandte Wissenschaften

QualitAI – enabling model interpretability

BW-TEC

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

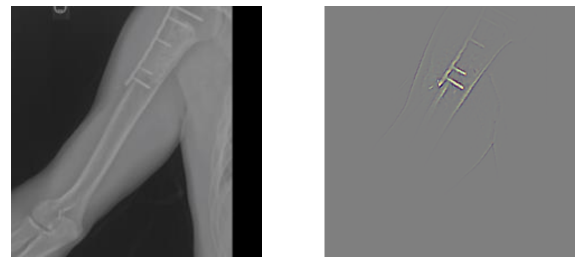
Swiss Confederation

Innosuisse – Swiss Innovation Agency

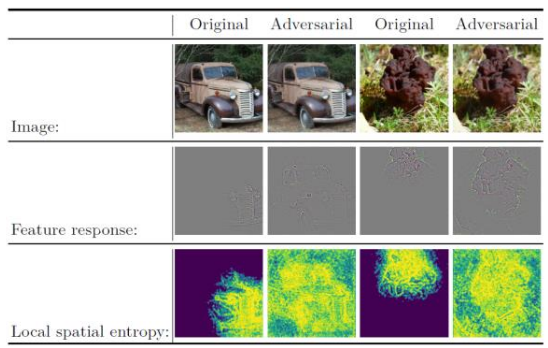
- Helps the developer in «debugging», needed by the user to trust
  → visualizations of learned features, training process, learning curves etc. should be «always on»

negative X-ray

positive X-ray

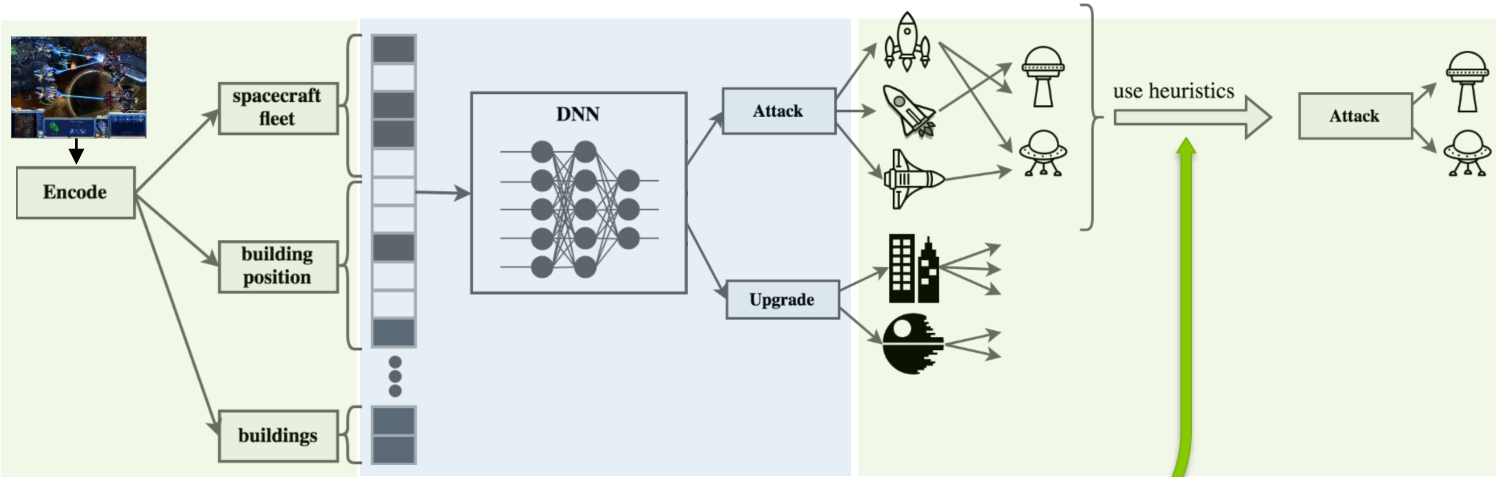


- Defends against adversarial attacks
  → thresholding local spatial entropy easily detects many adversarial attacking schemes through «lost focus»



Stadelmann, Amirian, Arabaci, Arnold, Duivesteijn, Elezi, Geiger, Lörwald, Meier, Rombach & Tuggener (2018). *«Deep Learning in the Wild»*. ANNPR'2018.
Amirian, Schwenker & Stadelmann (2018). *«Trace and Detect Adversarial Attacks on CNNs using Feature Response Maps»*. ANNPR'2018.
Amirian, Tuggener, Chavarriaga, Satyawan, Schilling, Schwenker, & Stadelmann (2021). «Two to Trust: AutoML for Safe Modelling and Interpretable Deep Learning for Robustness». ECAI'2020 workshops.

FarmAI: Automatic game playing
Collaboration with Inst. for Data Analysis & Process Design



Reinforcement learning: deep Q network

Large discrete action space → use heuristic
- makes exploration difficult
- elongates training time

Delayed and sparse reward → do reward shaping
- sequence of actions crucial to get a reward

Distance encoding → use reference points

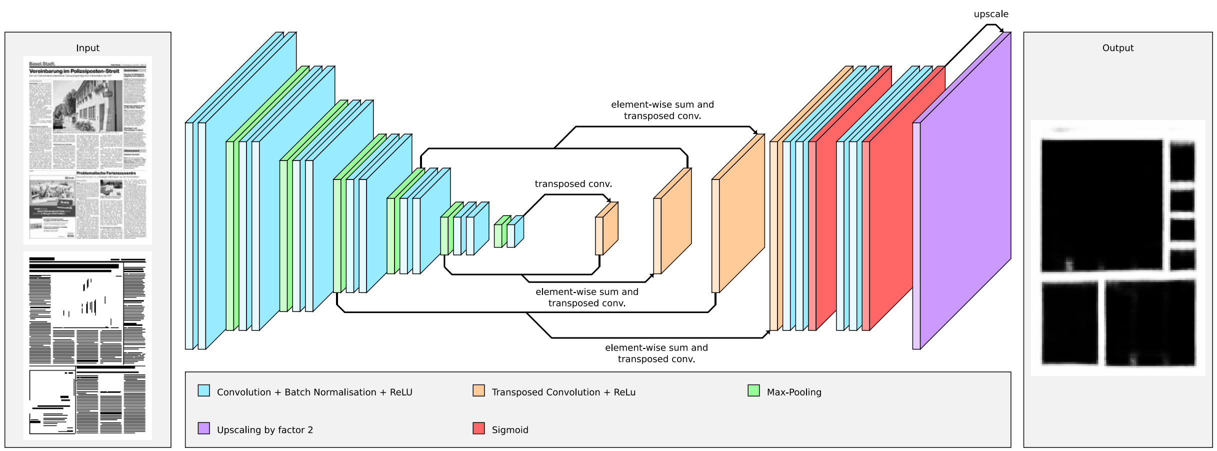
Transfer Learning → difficult: more complex environment needs other action sequence

Stadelmann, Amirian, Arabaci, Arnold, Duivesteijn, Elezi, Geiger, Lörwald, Meier, Rombach & Tuggener (2018). *«Deep Learning in the Wild»*. ANNPR'2018.

Project example: PANOPTES
Newspaper article segmentation for print media monitoring

Goal
- **Automatically segment newspaper pages** into constituting articles for automatic print media monitoring

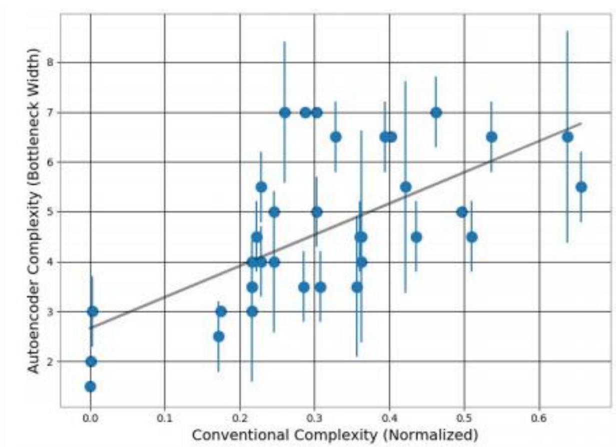
Approach
- **Image-based** approach with **deep neural networks** that learn layouting principles from examples



Meier, Stadelmann, Stampfli, Arnold, & Cieliebak. *“Fully convolutional neural networks for newspaper article segmentation”*. ICDAR 2017.

Stadelmann, Tolkachev, Sick, Stampfli, & Dürr. *“Beyond ImageNet - Deep Learning in Industrial Practice”*. In: Braschler et al. (Eds). “Applied Data Science – Lessons Learned for the Data-Driven Business”, Springer, 2019.

Project example: Complexity 4.0
Collaboration with HSG et al.



Goal
- **Reduce** unnecessary **complexity of product variability** in production environments in a data-driven (~automatable) fashion

Project team
- Business partners: **2 different industries** with large production facilities in CH
- **Economists**: ITEM-HSG (technology management, business models)
- **Engineers**: ZHAW-Engineering (machine learning), ZHAW-Life Sciences (simulation)

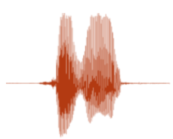
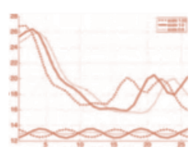
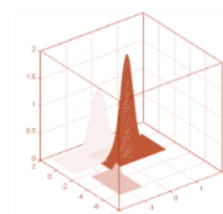
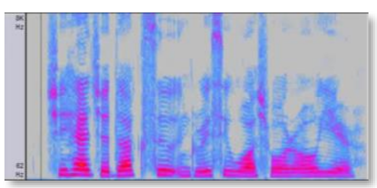
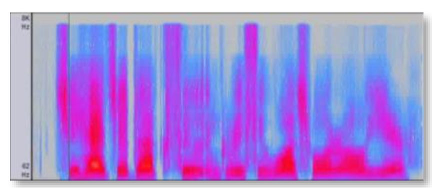
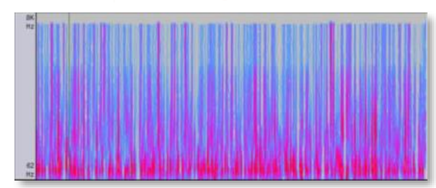
Results
- *"The paradigm of **data-driven decision support** can […] enter the domain of a highly qualified business consultant, **deliver**ing the **quantitative results** necessary to ponder informed **management decisions**."*
- *"**It is merely the knowledge** of what methods and technologies are possible and available **that** currently **hinders the faster adoption** of the data-driven paradigm in businesses."*

Hollenstein, Lichtensteiger, Stadelmann, Amirian, Budde, Meierhofer, Füchslin, & Friedli *"Unsupervised Learning and Simulation for Complexity Management in Business Operations"*. In: Braschler et al. (Eds). "Applied Data Science – Lessons Learned for the Data-Driven Business", Springer, 2019.

Talkalyzer
Contact: Prof. Dr. Thilo Stadelmann

Goal: Speaker Recognition in meetings on mobile devices
Challenge: Build reliable speaker models

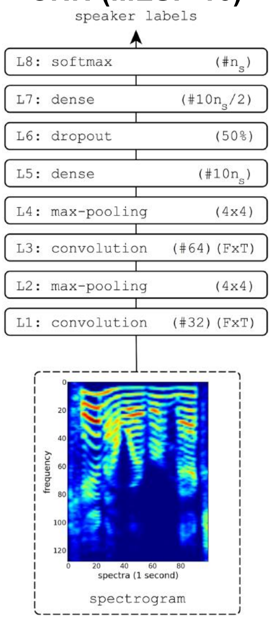
Approach:
- Loosen iid. assumption on feature vectors
- Use Deep Neural Network approach on continuous audio features
  ➔ find typical sounds of a speaker in a spectrogram

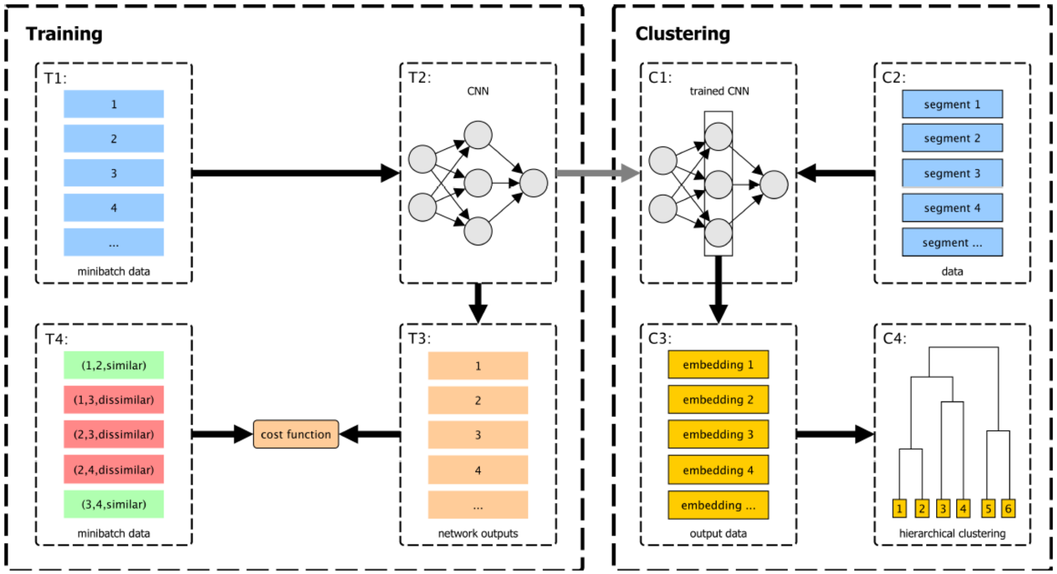
Zürcher Hochschule
für Angewandte Wissenschaften

Talkalyzer – exploiting time information

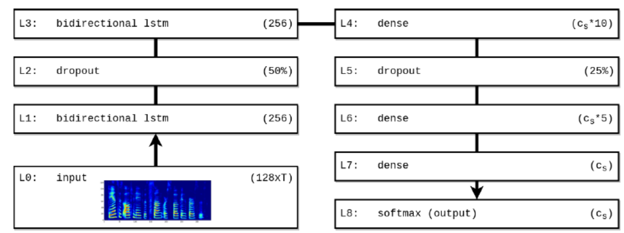
CNN (MLSP'16)



CNN & clustering-loss (MLSP'17)



RNN & clustering-loss (ANNPR'18)

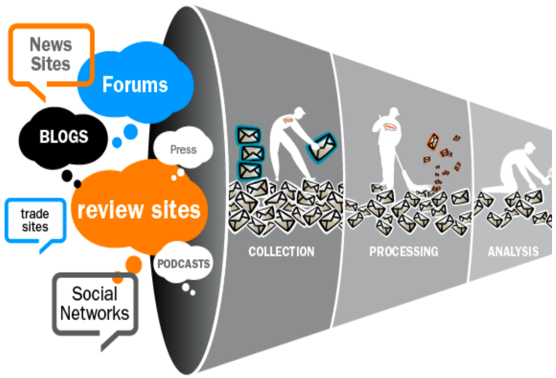


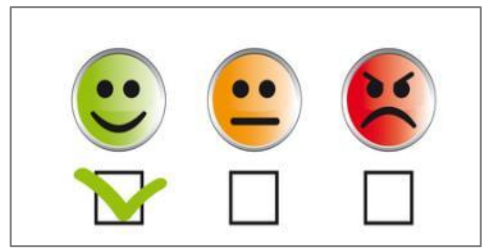
Method	MR	MR (legacy)
RNN /w PKLD	**2.19%** ($\frac{1.25\%+2.5\%+1.25\%+3.75\%}{4}$)	4.38% (average of 4 runs)
CNN /w PKLD [24]	-	5%
CNN /w cross entropy [23]	-	5%
ν-SVM [40]	6.25%	-
GMM/MFCC [40]	12.5%	-

Lukic, Vogt, Dürr & Stadelmann (2016). «Speaker Identification and Clustering using Convolutional Neural Networks». MLSP'2016.
Lukic, Vogt, Dürr & Stadelmann (2017). «Learning Embeddings for Speaker Clustering based on Voice Equality». MLSP'2017.
Stadelmann, Glinski-Haefeli, Gerber & Dürr (2018). «Capturing Suprasegmental Features of a Voice with RNNs for Improved Speaker Clustering». ANNPR'2018.

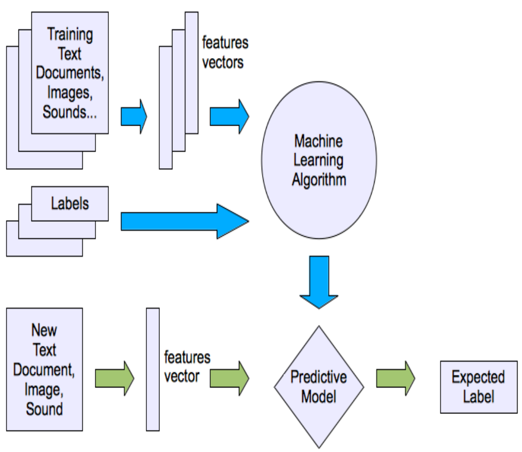
Sentiment Analysis
Contact: Dr. Mark Cieliebak

Challenge:

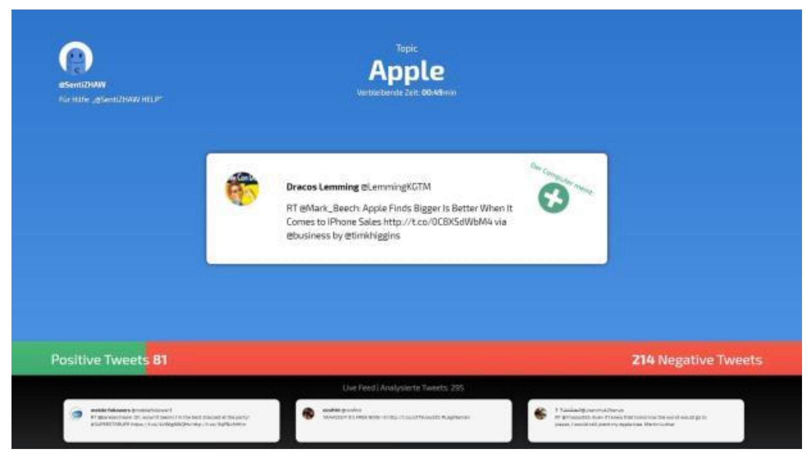




Approach:



Demo:



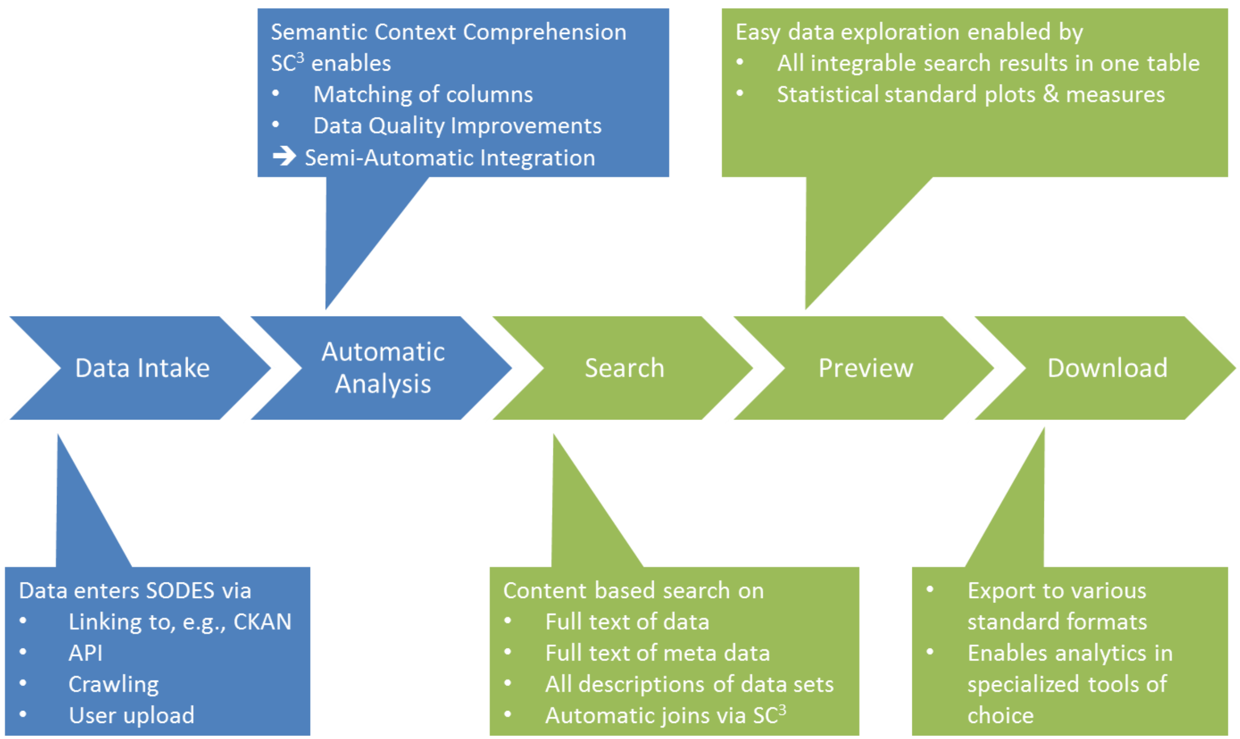
SODES – Swiss Open Data Exploration System

Contact: Prof. Dr. Mark Cieliebak

Challenge: Open Data promises to be a gold mine – but accessing and combining data from different data sources turns out to be non-trivial and very time consuming

Goal: A platform that enables easy and intuitive access, integration and exploration of different data sources

Solution:



Zürcher Hochschule
für Angewandte Wissenschaften

DaCoMo – Data-driven Condition Monitoring
Contact: Prof. Dr. Thilo Stadelmann

Mechmine
data science technologies

Situation: Maintaining big (rotating) machinery is expensive, defect is more expensive

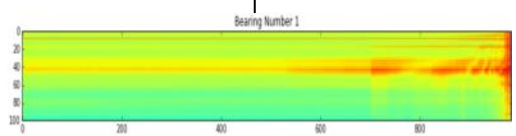
Goal: Schedule maintenance shortly before defect is expected, not merely regularly

Challenge: Develop an approach that adapts to each new machine automatically

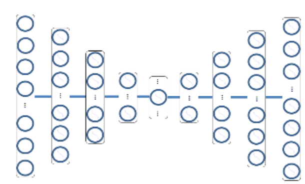
Solution: Use machine learning approaches for anomaly detection to learn the normal state of each machine and deviations of it purely from observed sensor signals; the approach combines classic and industry-proven features with e.g. deep learning auto-encoders

vibration sensors

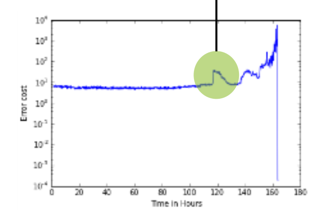
feature extraction



e.g., RNN autoencoder



early detection of fault



Stadelmann, Tolkachev, Sick, Stampfli, & Dürr. *"Beyond ImageNet - Deep Learning in Industrial Practice"*. In: Braschler et al. (Eds). "Applied Data Science – Lessons Learned for the Data-Driven Business", Springer, 2019.

Zürcher Hochschule für Angewandte Wissenschaften

Influencer Detection in Social Media

Target Specific, Interactive

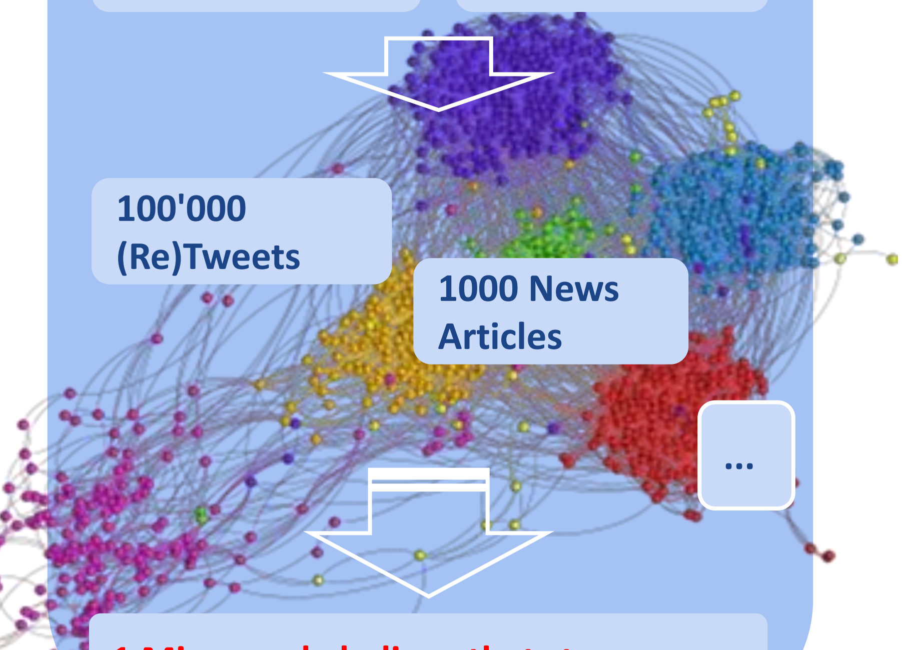
Contact: Dr. Mark Cieliebak

1 Person

1 Story

1 Blog Post

1 Tweet



100'000 (Re)Tweets

1000 News Articles

...

1 Mio people believe that story

How to find **this** person in 1 second

?

Social Media Monitoring

Data Source APIs

Periodic Data Fetch

Analyze Unstructured Text

- NLP
- Full-Text Index
- Graph Structure

Fast Retrieval