MENSCH GEGEN MASCHINE

AKTUELLER STAND DER KÜNSTLICHEN INTELLIGENZ IN FORSCHUNG UND ANWENDUNGEN.

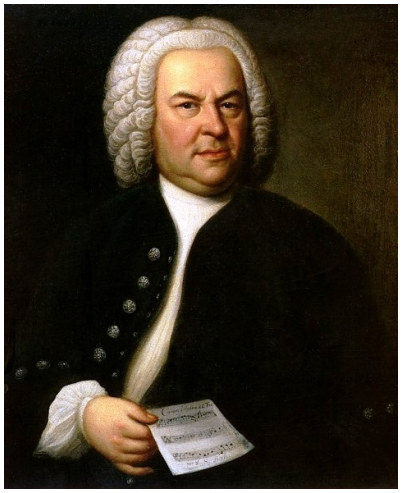
Prof. Dr. Frank Schönefeld
T-Systems Multimedia Solutions GmbH

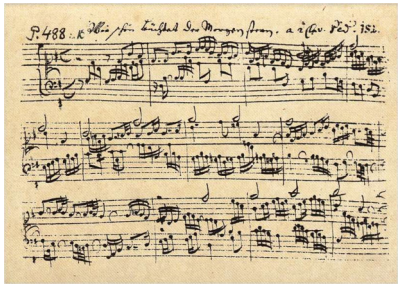
T · · Systems · Let's power higher performance

Johann Sebastian Bach

Bildquelle: https://www1.wdr.de/radio/wdr3/musik/johann-sebastian-bach-106~_v-ARDGrosserTeaser.jpg

WO DEEPART – DA AUCH DEEPBACH?
Wie gut klingt KI heute?





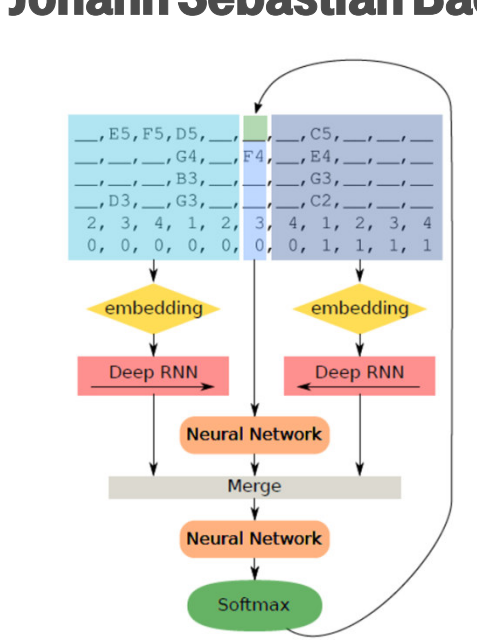


Figure 4. Graphical representations of DeepBach's neural network architecture for the soprano prediction p_1.

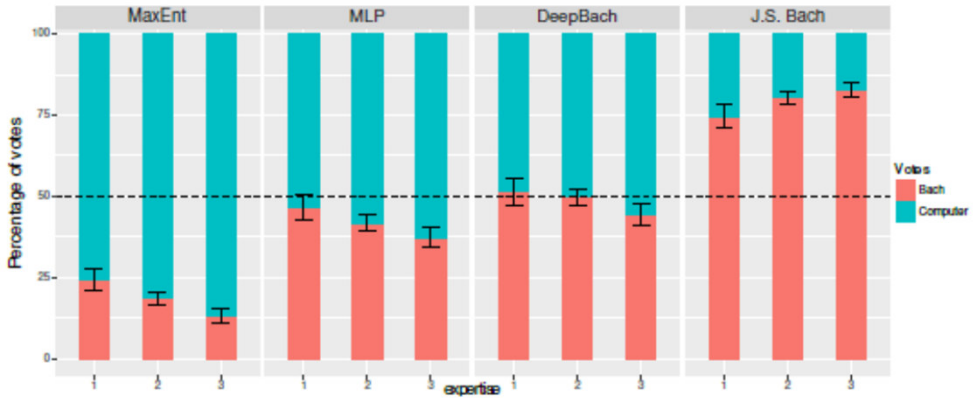


Figure 5. Results of the "Bach or Computer" experiment. The figure shows the distribution of the votes between "Computer" (blue bars) and "Bach" (red bars) for each model and each level of expertise of the voters (from 1 to 3), see Sect. 3.2 for details.

Quelle:
https://upload.wikimedia.org/wikipedia/commons/6/6a/Johann_Sebastian_Bach.jpg
http://artsites.ucsc.edu/faculty/cope/experiments.htm
Source: https://arxiv.org/pdf/1612.01010.pdf

MENSCH GEGEN MASCHINE

- **SPIELEN**
- WAHRNEHMEN
- TEXTVERSTEHEN
- KREATIVITÄT

SPIELEN: SCHACH MATT.
Mit Positionsbewertung und Brute Force zum Erfolg.

1997

Quelle: http://blogs-images.forbes.com/davidewalt/files/2011/05/garry-kasparov-deep-blue-ibm.jpg

SPIELEN: GO – EIN „PERFECT INFORMATION“ GAME.
Die letzte Bastion ist gefallen.

2016

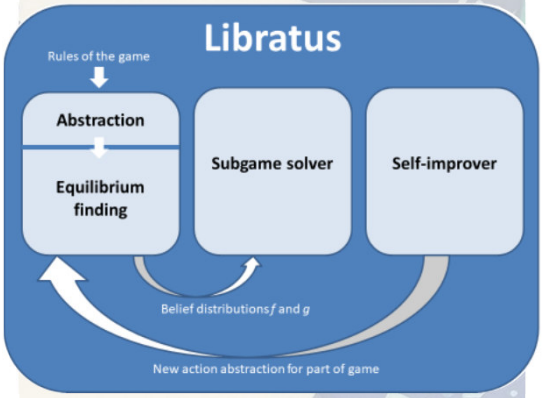
Quelle: http://images02.futurezone.at/46-79007756.jpg/186.421.945

POKER – EIN NON-INFORMATION COMPLETE GAME.
No he can‘t read my Pokerface.

2018

Heads-Up No-Limit Texas Hold’em

- Has become the main ***benchmark and challenge problem*** in AI for imperfect-information games
- No-Limit Betting = Continuous Action Space
  - Technically, 10^{161} situations since bets must be integers
- The most popular variant of poker in the world
  - Played in the World Series of Poker Main Event
  - Featured in *Casino Royale* and *Rounders*
- Two-player variant allows an objective winner to be determined
- No prior AI has been able to beat top humans



Final Result

- Libratus beat the top humans in this game by a lot
  - 14.7 BB / 100
  - Statistical significance 99.98%, i.e., p-value of 0.0002
  - Each human lost to Libratus

Source: Noah Brown: From Poker AI to Negotiation AI: Dealing with Hidden Information. TTI Vanguard Conference:June 2018

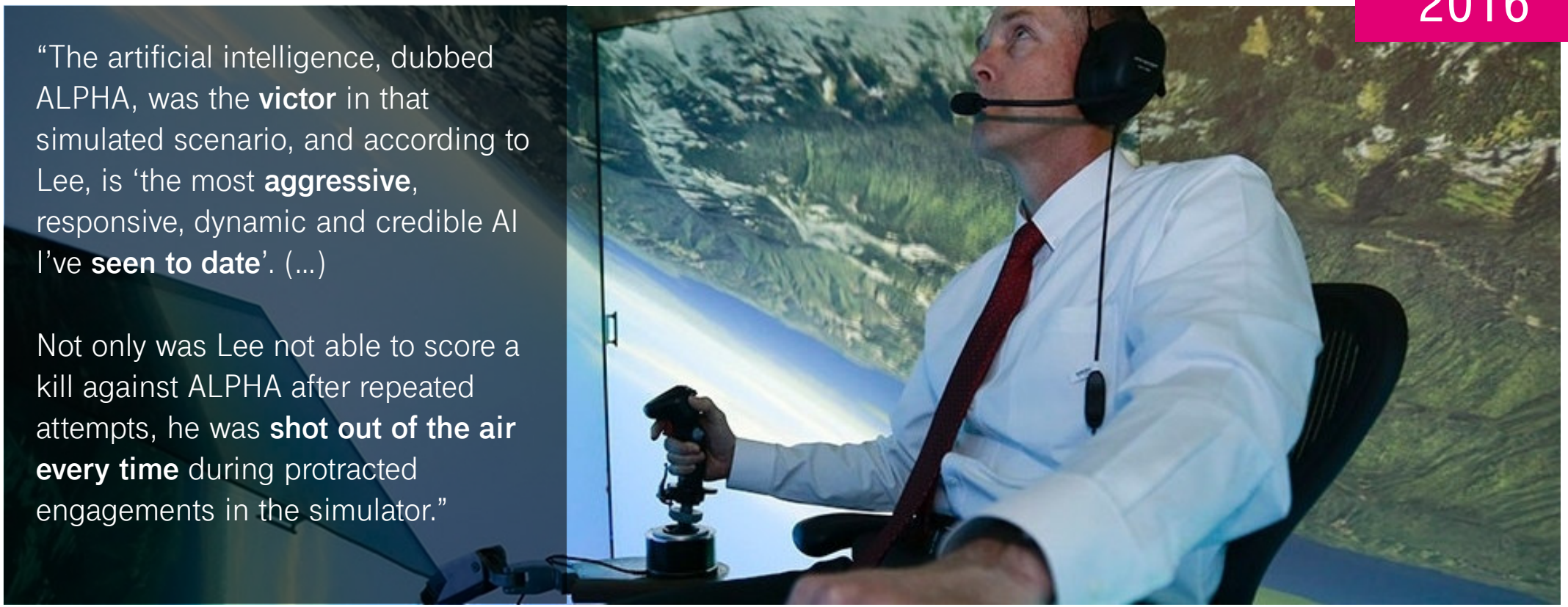
STARCRAFT 2: TERRANS VS PROTOSS VS ZERG.
Auch Multiplayer Games? – Auch Multiplayer Games!

2019

Quelle 1: https://starcraft2.com/de-de/media
Quelle 2: https://www.scientificamerican.com/article/ai-beats-top-human-players-at-strategy-game-starcraft-ii/

WENN AUS SPIEL ERNST WIRD.
Noch ist es Simulation.

2016



Quelle: http://magazine.uc.edu/editors_picks/recent_features/alpha.html

MENSCH GEGEN MASCHINE

SPIELEN

▶ **WAHRNEHMEN**

TEXTVERSTEHEN

KREATIVITÄT

DEEP NEURAL NETWORKS:
Vom Konkreten zum Abstrakten.

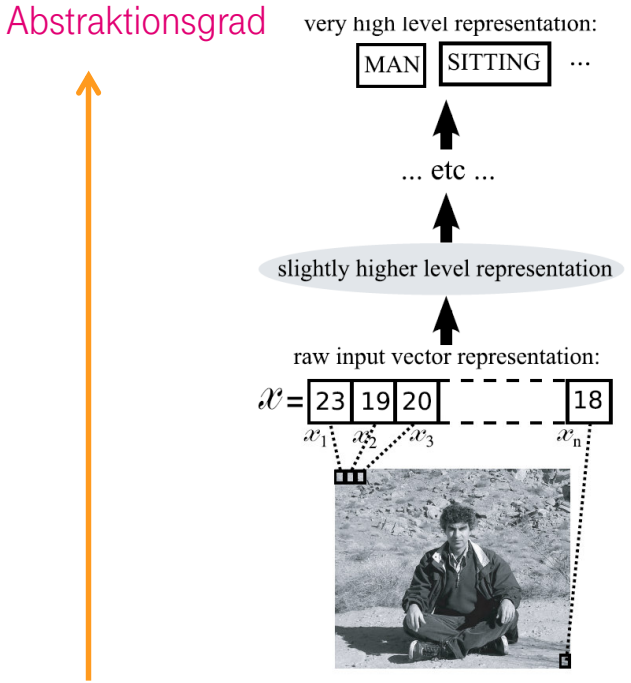
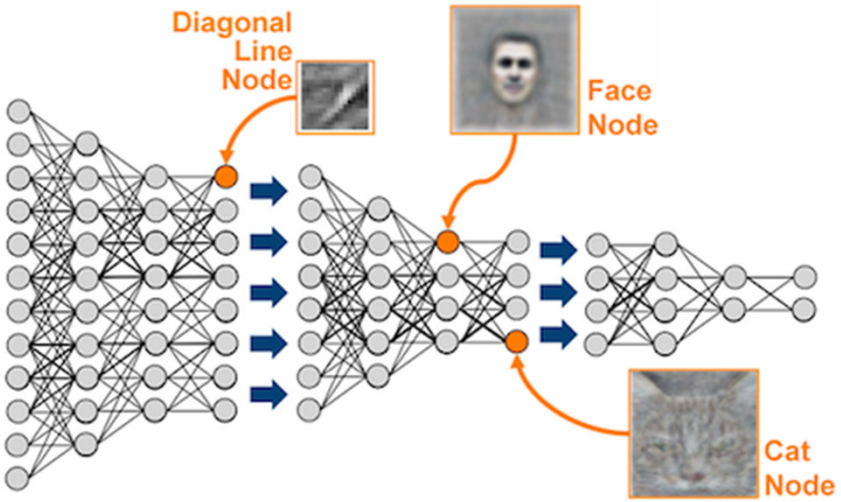


Fig. 1.1 We would like the raw input image to be transformed into gradually higher levels of representation, representing more and more abstract functions of the raw input, e.g., edges, local shapes, object parts, etc. In practice, we do not know in advance what the "right" representation should be for all these levels of abstractions, although linguistic concepts might help guessing what the higher levels should implicitly represent.

Source: Learning Deep Architectures for AI; Yoshua Bengio 2009.



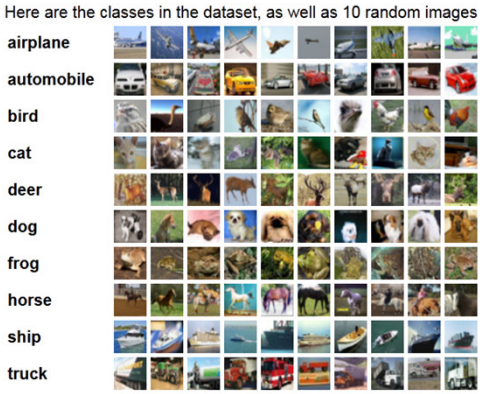
Source: http://theanalyticsstore.ie/deep-learning/.

WETTBEWERBE IN COMPUTER VISION. MNIST. CIFAR. ILSVRC. ...

MNIST Mixed National Institute of Standards and Technology

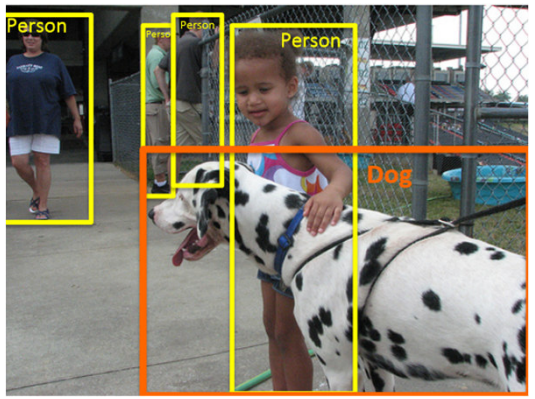
- 60.000 Trainings Bilder
- 10.000 Test Bilder
- Fehlerrate eines CNN (Convolutional Neural Network): 0.23 %) 2012 Schmidhuber Schweiz.

CIFAR Canadian Institute for Advanced Research



- The CIFAR-10 (100) dataset consists of 60000 32x32 colour images in 10 (100) classes,
- 50000 training images and 10000 test images.
- Winner 2014: 4,5% Error Rate Deep CNN (Graham, Uni Warwick)

ImageNet Large Scale Visual Recognition Challenge



- 200 object classes with 456,567 images for detection
- 1000 object classes with 1,431,167 images for localisation
- Winner 2014 Det: GoogLeNet with top-5 error-rate of 7% (Deep CNN) – wins 142 out of 200 cl.

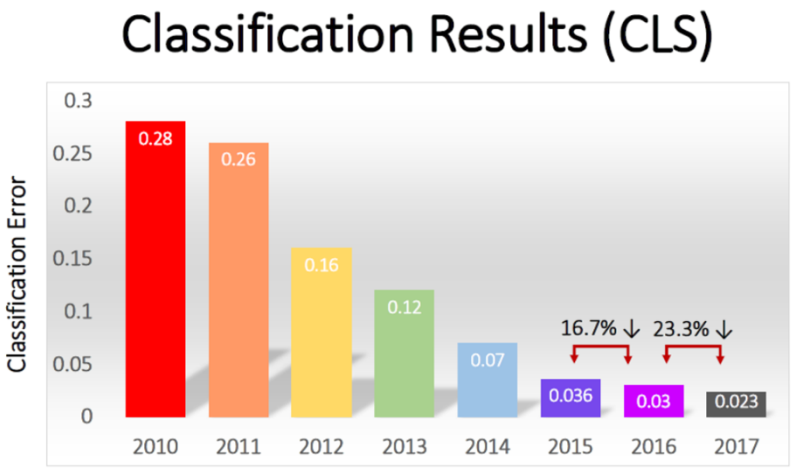
WETTBEWERBE IN COMPUTER VISION.
Welche Trennschärfe wird benötigt?

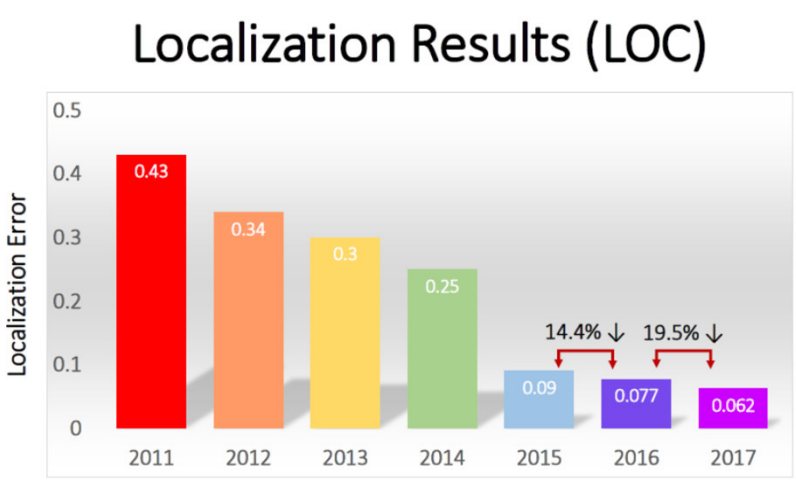
(a) Siberian husky (b) Eskimo dog

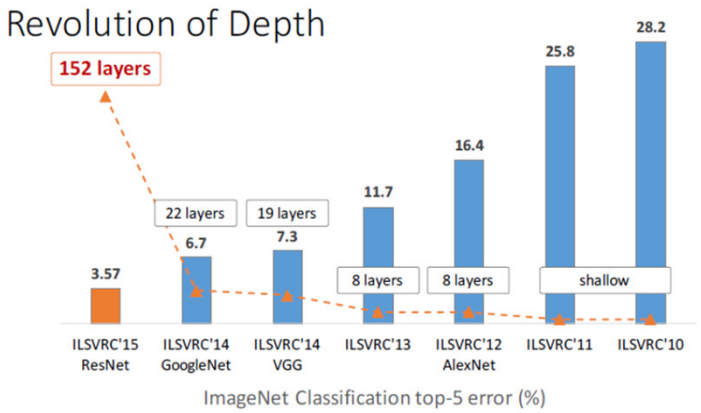
Figure 1: Two distinct classes from the 1000 classes of the ILSVRC 2014 classification challenge.

Source: Going deeper with convolutions. arxiv: 1409.4842v1 Sep 2014

DER ILSVRC – ENTWICKLUNG BIS 2017.
Immer Besser. Immer Tiefer. Woanders.







ILSVRC2017 CLS Results - 'External' Data

Team Name	Error(%)
NUS-Qihoo_DPNs	0.0271
BDAT	0.0300

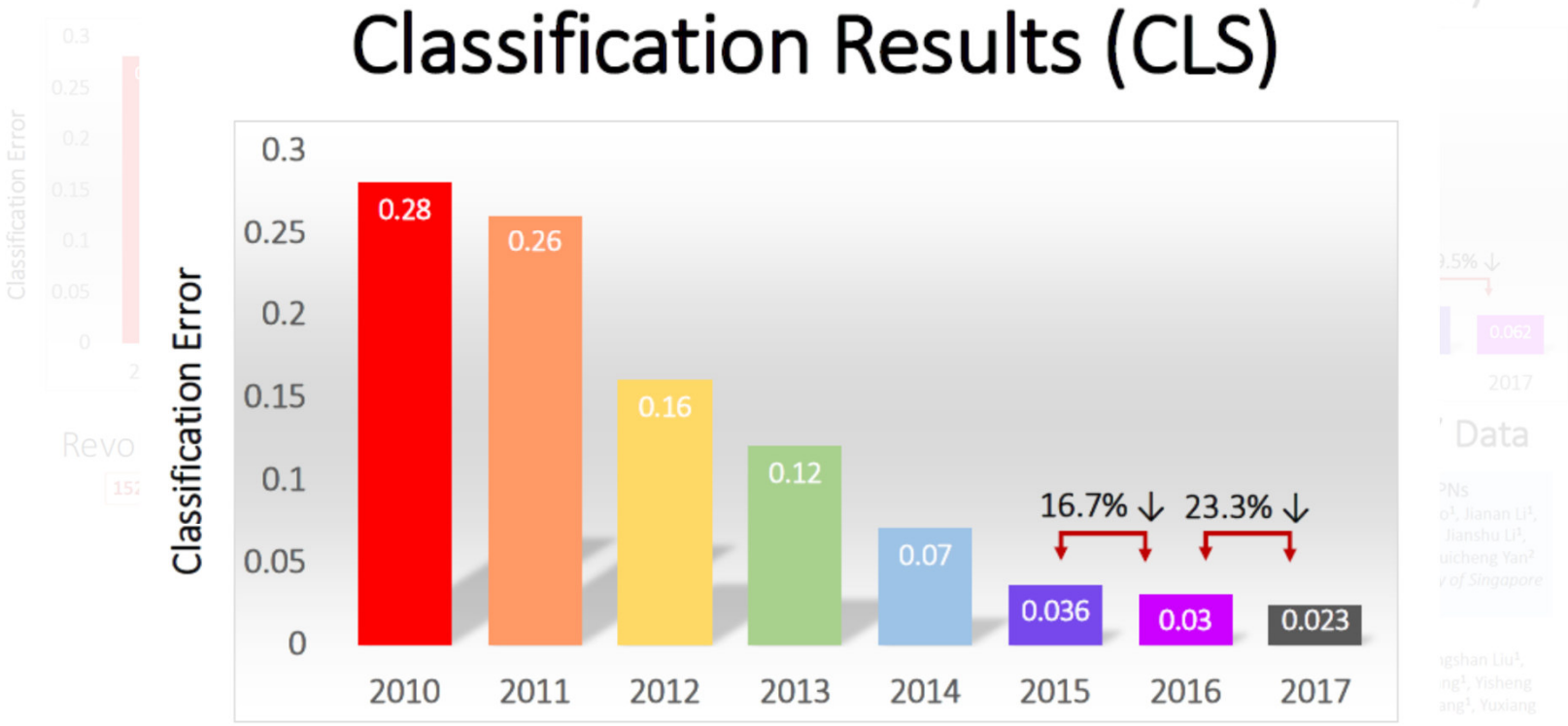
NUS-Qihoo_DPNs
Yunpeng Chen[1], Huaxin Xiao[1], Jianan Li[1], Xuecheng Nie[1], Xiaojie Jin[1], Jianshu Li[1], Jiashi Feng[1], Jian Dong[2], Shuicheng Yan[2]
1. NUS - National University of Singapore
2. Qihoo 360

BDAT
Hui Shuai[1], Zhenbo Yu[1], Qingshan Liu[1], Xiaotong Yuan[1], Kaihua Zhang[1], Yisheng Zhu[1], Guangcan Liu[1], Jing Yang[1], Yuxiang Zhou[2], Jiankang Deng[2]
1. Nanjing University of Information Science & Technology
2. Imperial College London

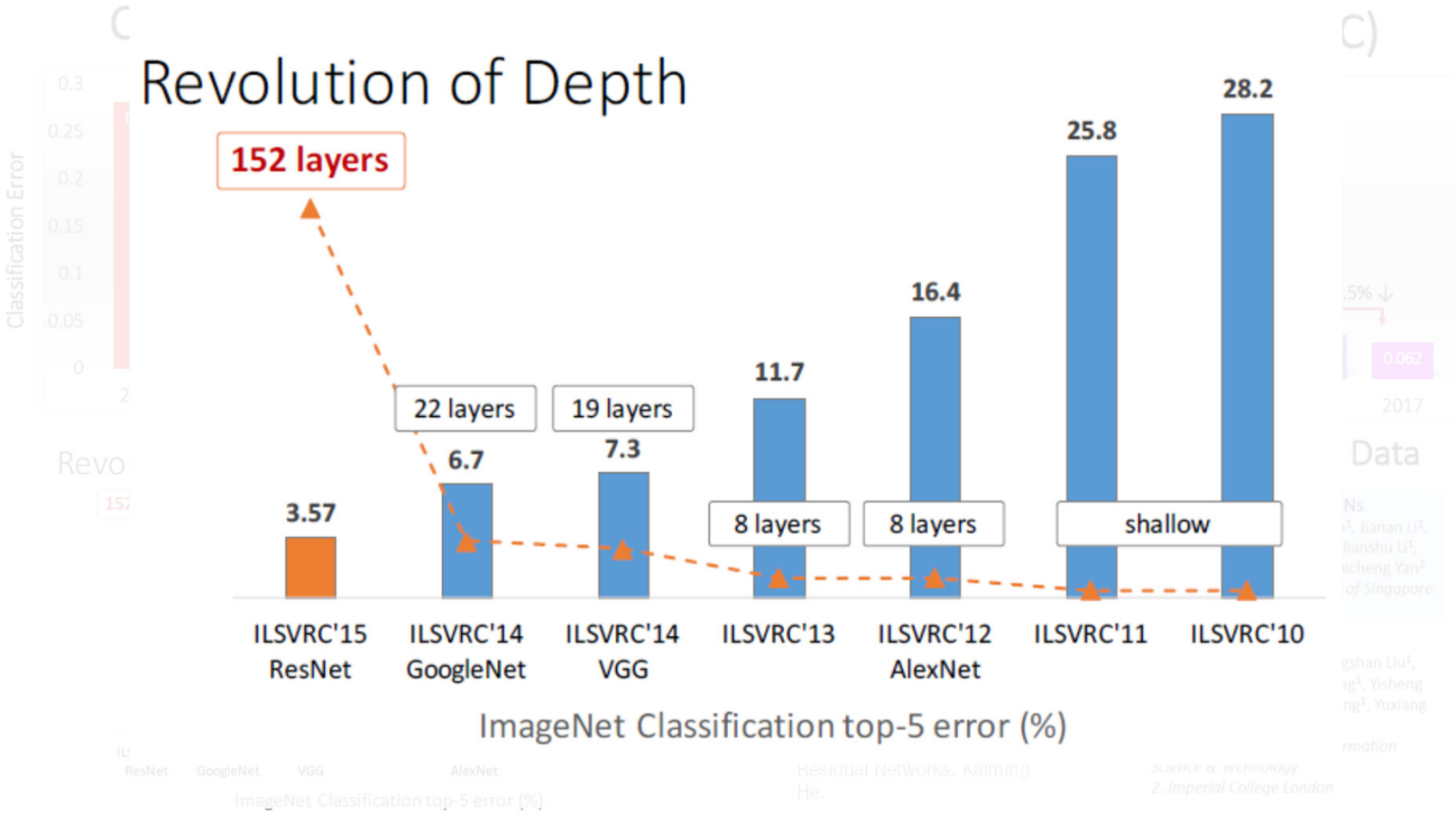
Source: ImageNet 2017 Challenge. E Park et.al. und ICML 2016 Tutorial on Deep Residual Networks. Kaiming He.

T · · Systems · Let's power higher performance

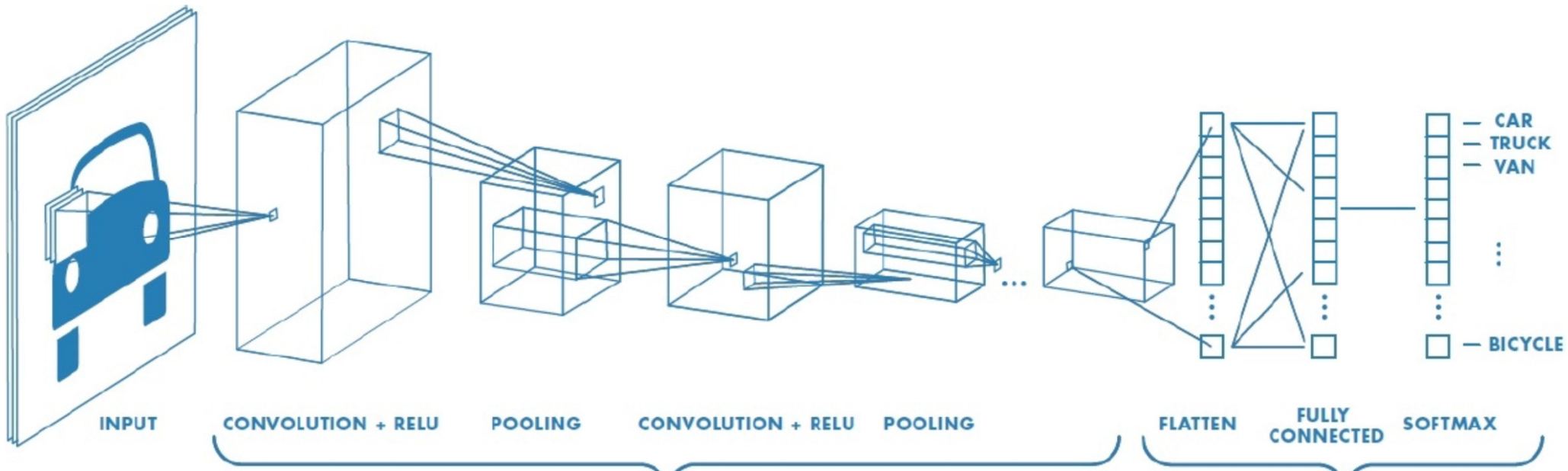
DER ILSVRC – ENTWICKLUNG BIS 2017.
Immer Besser. Immer Tiefer. Woanders.



DER ILSVRC – ENTWICKLUNG BIS 2017.
Immer Besser. Immer Tiefer. Woanders.



CONVOLUTIONAL NEURAL NETWORKS:
Die dominierende KI Architektur der Bildverarbeitung/-erkennung.



Quelle: introtodeeplearning.com

Learn weights for convolutional filters and fully connected layers

MENSCH GEGEN MASCHINE

SPIELEN

WAHRNEHMEN

▶ **TEXTVERSTEHEN**

KREATIVITÄT

GLUE SuperGLUE Paper Code Tasks Leaderboard FAQ Diagnostics Submit Login

Quelle: https://gluebenchmark.com/leaderboard/

	Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	WNLI	AX
	1	T5 Team - Google	T5		90.3	71.6	97.5	92.8/90.4	93.1/92.8	75.1/90.6	92.2	91.9	96.9	92.8	94.5	53.1
	2	ERNIE Team - Baidu	ERNIE		90.0	72.2	97.5	93.0/90.7	92.9/92.5	75.2/90.8	91.2	90.8	96.0	90.9	94.5	49.4
	3	Microsoft D365 AI & MSR AI & GATECH	MT-DNN-SMART		89.9	69.5	97.5	93.7/91.6	92.9/92.5	73.9/90.2	91.0	90.8	99.2	89.7	94.5	50.2
+	4	王玮	ALICE v2 large ensemble (Alibaba DAMO NLP)		89.7	73.2	97.1	93.9/91.9	93.0/92.5	74.8/91.0	90.8	90.6	95.9	87.4	94.5	48.7
+	5	Microsoft D365 AI & UMD	FreeLB-RoBERTa (ensemble)		88.4	68.0	96.8	93.1/90.8	92.3/92.1	74.8/90.3	91.1	90.7	95.6	88.7	89.0	50.1
	6	Junjie Yang	HIRE-RoBERTa		88.3	68.6	97.1	93.0/90.7	92.4/92.0	74.3/90.2	90.7	90.4	95.5	87.9	89.0	49.3
	7	Facebook AI	RoBERTa		88.1	67.8	96.7	92.3/89.8	92.2/91.9	74.3/90.2	90.8	90.2	95.4	88.2	89.0	48.7
+	8	Microsoft D365 AI & MSR AI	MT-DNN-ensemble		87.6	68.4	96.5	92.7/90.3	91.1/90.7	73.7/89.9	87.9	87.4	96.0	86.3	89.0	42.8
	9	GLUE Human Baselines	GLUE Human Baselines		87.1	66.4	97.8	86.3/80.8	92.7/92.6	59.5/80.4	92.0	92.8	91.2	93.6	95.9	-
	10	Stanford Hazy Research	Snorkel MeTaL		83.2	63.8	96.2	91.5/88.5	90.1/89.7	73.1/89.9	87.6	87.2	93.9	80.9	65.1	39.9
	11	XLM Systems	XLM (English only)		83.1	62.9	95.6	90.7/87.1	88.8/88.2	73.2/89.8	89.1	88.5	94.0	76.0	71.9	44.7
	12	Zhuosheng Zhang	SemBERT		82.9	62.3	94.6	91.2/88.3	87.8/86.7	72.8/89.8	87.6	86.3	94.6	84.5	65.1	42.4
	13	Danqi Chen	SpanBERT (single-task training)		82.8	64.3	94.8	90.9/87.9	89.9/89.1	71.9/89.5	88.1	87.7	94.3	79.0	65.1	45.1
	14	Kevin Clark	BERT + BAM		82.3	61.5	95.2	91.3/88.3	88.6/87.9	72.5/89.7	86.6	85.8	93.1	80.4	65.1	40.7
	15	Nitish Shirish Keskar	Span-Extractive BERT on STILTs		82.3	63.2	94.5	90.6/87.6	89.4/89.2	72.2/89.4	86.5	85.8	92.5	79.8	65.1	28.3
	16	Jason Phang	BERT on STILTs		82.0	62.1	94.3	90.2/86.6	88.7/88.3	71.9/89.4	86.4	85.6	92.7	80.1	65.1	28.3
+	17	Jacob Devlin	BERT: 24-layers, 16-heads, 1024-hidden		80.5	60.5	94.9	89.3/85.4	87.6/86.5	72.1/89.3	86.7	85.9	92.7	70.1	65.1	39.6
	18	Neil Houlsby	BERT + Single-task Adapters						7/84.3	87.3/86.1	71.5/89.4	85.4	85.0	92.4	71.6	65.1	9.2

Click on a submission to see more information

WHAT IS GLUE? (I)
General Language Understanding Evaluation.

MNLI Multi-Genre Natural Language Inference	**QQP** Quora Question Pairs	**QNLI** Question Natural Language Inference	**SST-2 Stanford Sentiment Treebank**
Given a pair of sentences, the goal is to predict whether the second sentence is an entailment, contradiction, or neutral with respect to the first one.	is a binary classification task where the goal is to determine if two questions asked on Quora are semantically equivalent	The positive examples are (question, sentence) pairs which do contain the correct answer, and the negative examples are (question, sentence) from the same paragraph which do not contain the answer.	binary single-sentence classification task consisting of sentences extracted from movie reviews with human annotations of their sentiment
Example: Die Demonstration verlief friedlich. Es gab 50 Verletzte.	Unterschied zwischen A und B; vs B und A	Wer schlug die Reformationsthesen an die Kirchentür zu Wittenberg? (Luther vs sein Diener)	...mein Lieblingsfilm vs schau ich mir kein 2. mal an.

Source: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

WHAT IS GLUE? (II)
General Language Understanding Evaluation.

CoLA The Corpus of Linguistic Acceptability	**STS-B** The Semantic Textual Similarity Benchmark	**MRPC** Microsoft Research Paraphrase Corpus	**RTE** Recognizing Textual Entailment	WNLI Winograd NLI Natural Language inference dataset
Is a binary single-sentence classification task, where the goal is to predict whether an English sentence is linguistically “acceptable” or not	is a collection of sentence pairs drawn from news headlines and other sources. They were annotated with a score from 1 to 5 denoting how similar the two sentences are in terms of semantic meaning.	Consists of sentence pairs automatically extracted from online news sources, with human annotations for whether the sentences in the pair are semantically equivalent	is a binary entailment task similar to MNLI, but with much less training data	is a reading comprehension task in which a system must read a sentence with a pronoun and select the referent of that pronoun from a list of choices.
Example: They can sing. Usually, any lion is majestic.	How old are you? What is your age? How are you?	Das Rote Meer trennt Afrika von Asien. Zwischen Afrika und Asien verläuft das RM.	Beide mochten den Ort. (Mag Peter den Ort?)	Jim [yelled at/comforted] Kevin because he was so upset. Who was upset? **Answers:** Jim/Kevin.

Source: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

ARCHITECTURE AND PRINCIPLES OF BERT
Bidirectional Encoder Representations from Transformers.

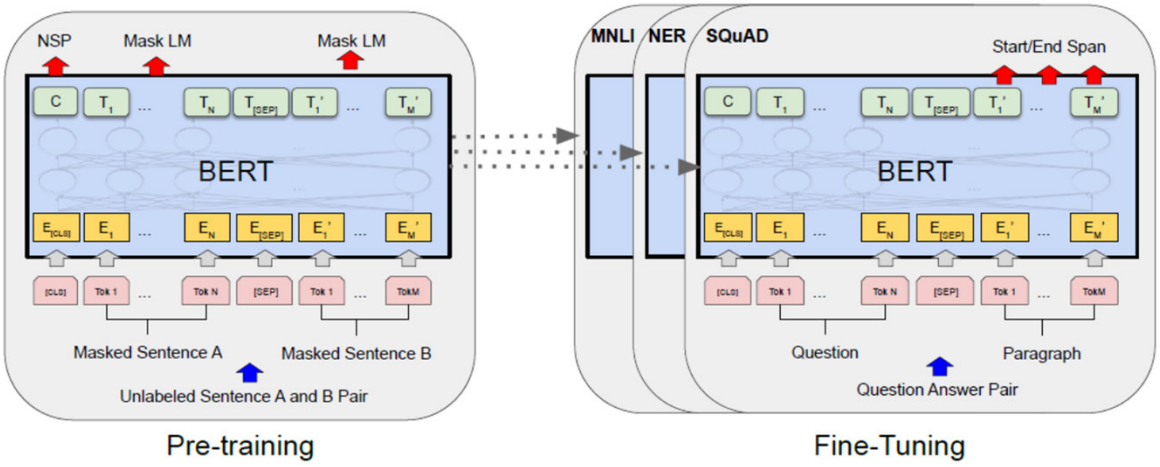


Figure 1: Overall pre-training and fine-tuning procedures for BERT. Apart from output layers, the same architectures are used in both pre-training and fine-tuning. The same pre-trained model parameters are used to initialize models for different down-stream tasks. During fine-tuning, all parameters are fine-tuned. [CLS] is a special symbol added in front of every input example, and [SEP] is a special separator token (e.g. separating questions/answers).

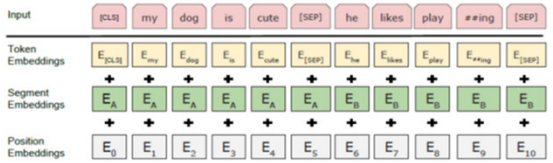


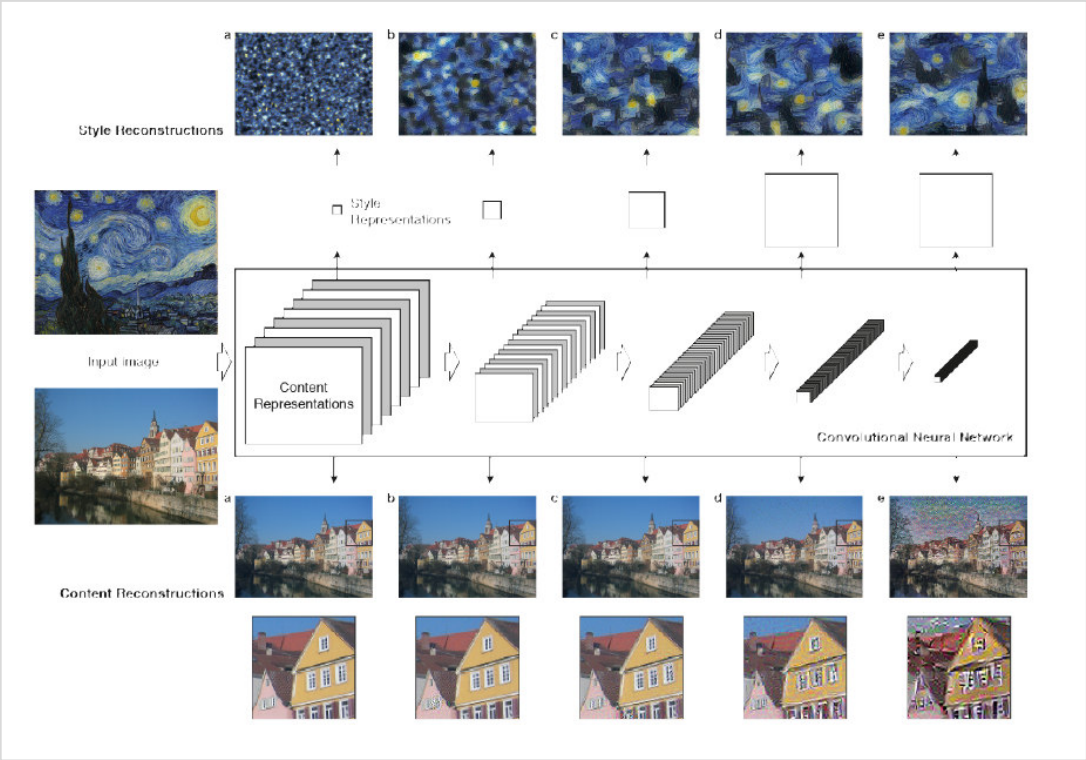
Figure 2: BERT input representation. The input embeddings are the sum of the token embeddings, the segmentation embeddings and the position embeddings.

Source: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Devlin et.al.

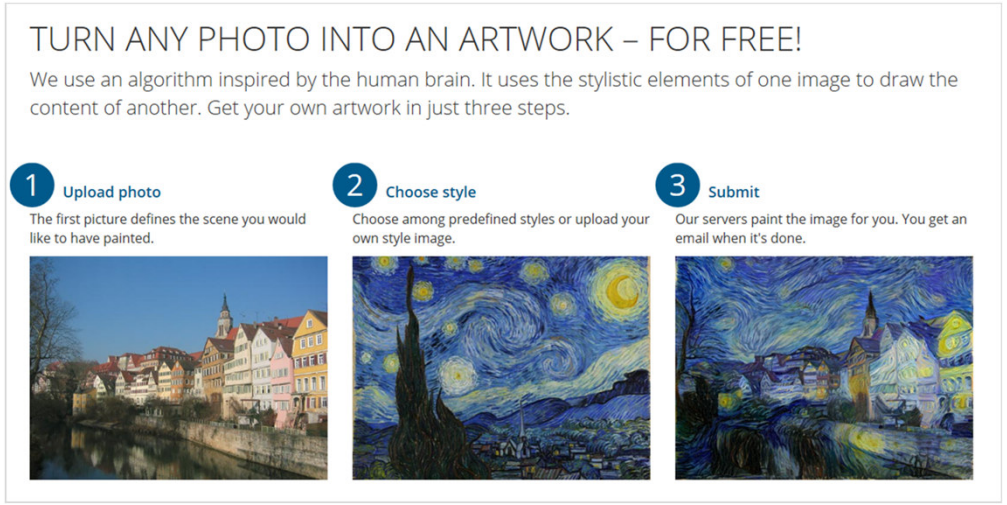
MENSCH GEGEN MASCHINE

SPIELEN
WAHRNEHMEN
TEXTVERSTEHEN
▶ **KREATIVITÄT**

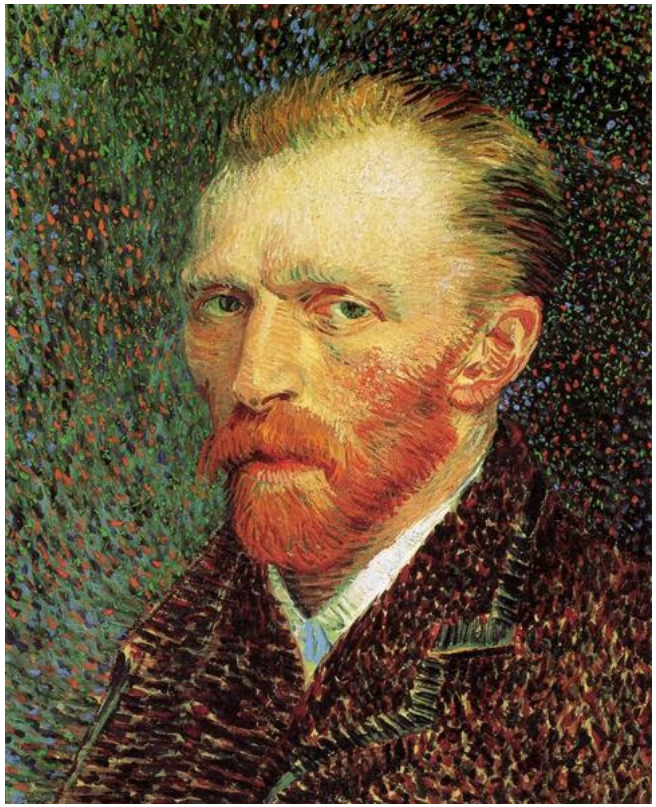
KÜNSTLICHE KUNST
Mit wenigen Klicks zum eigenen Van Gogh!



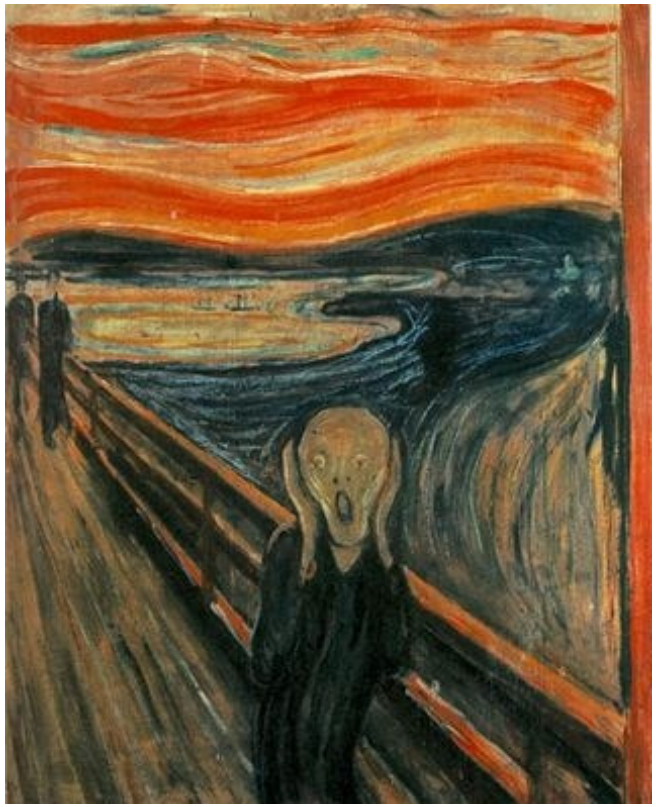
A Neural Algorithm of Artistic Style. Leon A. Gatys, Alexander S. Ecker, Matthias Bethge Arxiv: 2015.



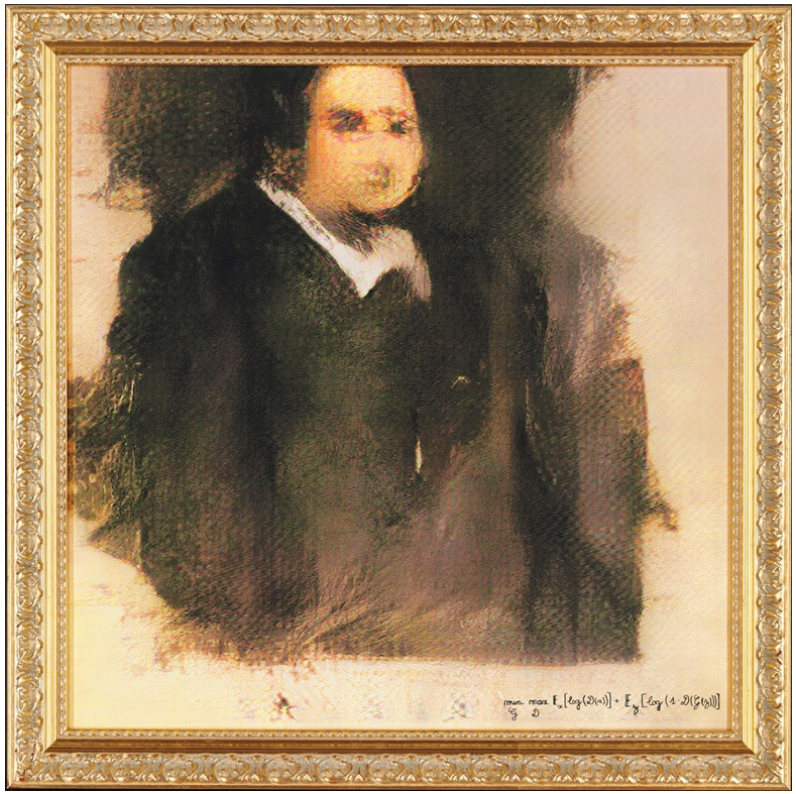
KÜNSTLICHE KUNST
Mit wenigen Klicks zum eigenen Van Gogh!



KÜNSTLICHE KUNST
Mit wenigen Klicks zum eigenen Munch!



KÜNSTLICHE KUNST
Das erste KI Portrait – Versteigert bei Christie‘s: 432.500 $



- Portrait of Edmond de Belamy
- From a distance, the 70 cm by 70 cm portrait printed on canvas and hung in a gilded wood frame, looks like it belongs in a museum of classical art. But upon closer inspection, the artist’s signature — the mathematical formula that created it (min G max D x [log (D(x))] + z [log (1 – D (G(z)))]) — reveals that the artist was not human.
- But in November 2019, another in the Belamy series, “La Baronne de Belamy,” sold at Sotheby’s without the same success. “La Baronne” sold for only US$25,000, just slightly more than its estimated value.

Source: theconversation.com

KÜNSTLICHE KUNST
Do we like it?

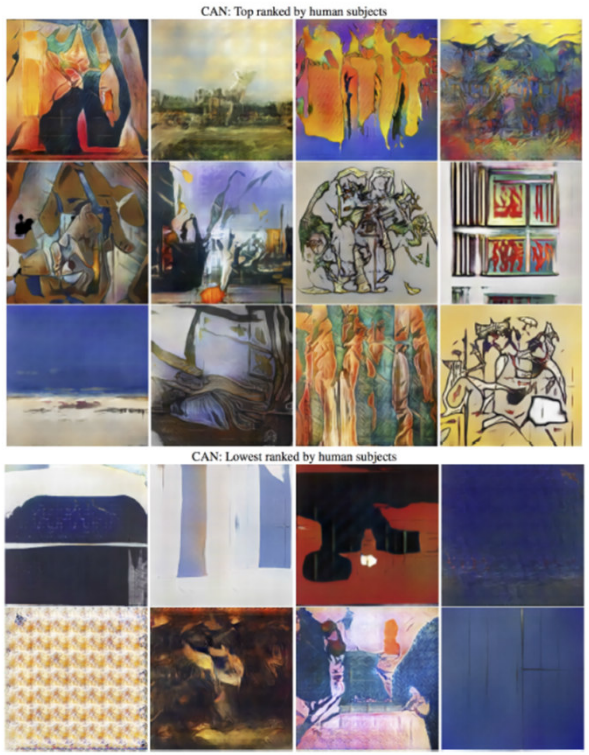


Figure 5: Example of images generated by CAN. Top: Images ranked high in "likeness" according to human subjects. Bottom: Images ranked the lowest by human subjects.

Table 2: Means and standard deviations of responses of Experiment I

Painting set	Q1 (std)	Q2 (std)
CAN	53% (18%)†	3.2 (1.5) ‡
DCGAN [18] (64x64)	35% (15%) †	2.8 (0.54) ‡
Abstract Expressionist	85% (16%)	3.3 (0.43)
Art Basel 2016	41% (29%)	2.8 (0.68)
Artist sets combined	62% (32%)	3.1 (0.63)

All images are resized to 512x512 resolution

† Q1 t-test (CAN vs. DCGAN) p-value = $1.9932e-15$

‡ Q2 t-test (CAN vs. DCGAN) p-value = $9.3634e-06$

Q1: Do you think the work is created by an artist or generated by a computer? The User has to choose one of two answers: artist or computer.

Q2: User asked to rate how they like the image in a scale 1 (extremely dislike to 5 (extremely like).

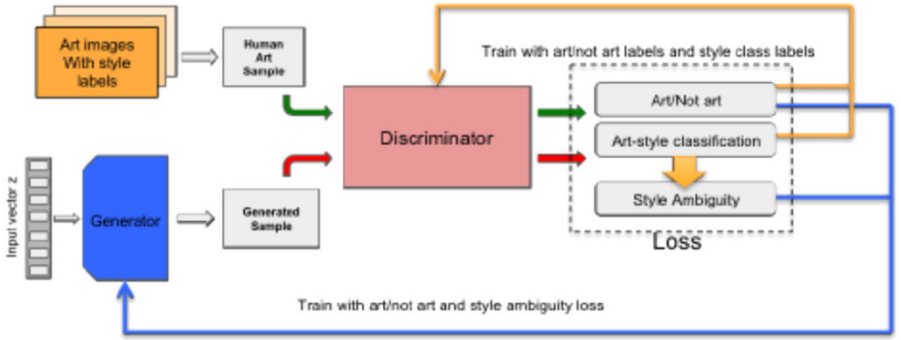


Figure 2: Block diagram of the CAN system.

Quelle: A. Elgammal: CAN Creative Adversarial Networks Generating Art by Learning about Styles and deviating from Style norms. https://arxiv.org/pdf/1706.07068.pdf

AUCH GEDICHTE?

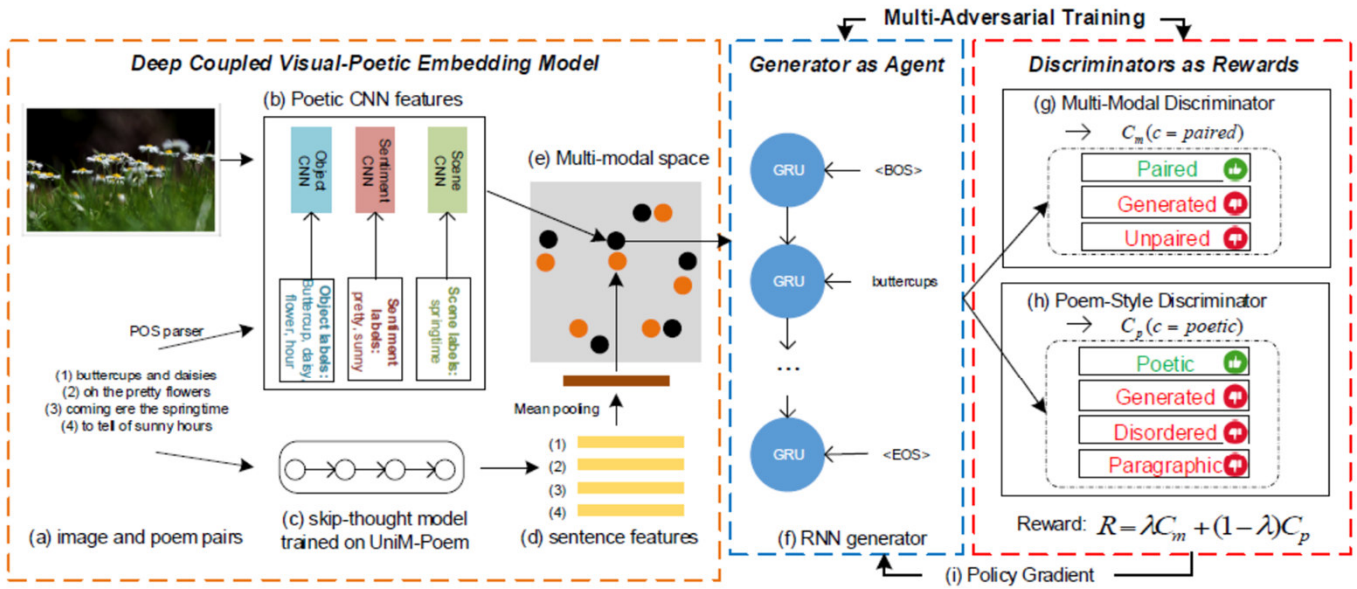


Figure 2: The framework of poetry generation with multi-adversarial training. A deep coupled visual-poetic model (e) is trained by human annotated image-poem pairs (a). The image features (b) are poetic multi-CNN features obtained by fine-tuning CNNs with the extracted poetic symbols (e.g., objects, scenes and sentiments) by a POS parser [28] from poems. The sentence features (d) of poems are extracted from a skip-thought model (c) trained on the largest public poem corpus (UniM-Poem). A RNN-based sentence generator (f) is trained as agent and two discriminators considering multi-modal (g) and poem-style (h) critics of a generated poem to a given image provide rewards to policy gradient (i). POS parser extracts Part-Of-Speech words from poems.

Beyond Narrative Description: Generating Poetry from Images by Multi-Adversarial Training

Bei Liu*
Kyoto University
liubei@dl.kuis.kyoto-u.ac.jp

Jianlong Fu†
Microsoft Research Asia
jianf@microsoft.com

Makoto P. Kato
Kyoto University
mpkato@acm.org

Masatoshi Yoshikawa
Kyoto University
yoshikawa@i.kyoto-u.ac.jp

AUCH GEDICHTE?

Show and tell (1CNN)
i am a coal-truck
by a broken heart
i have no sound
the sound of my heart
i am not

Show and tell (3CNNs)
i will find a little bird
that shivers and falls the morrow
and every day of night and seas immortal night
i know that all the world shall be alone
and the wild wild horses
the women of the great city of the sea
and i like to be a jellyfish
i will never find a way

I2P-GAN w/o discriminator
when we can go
the sun is singing in the forest rain
and the mist is the sound of the sea
and the soul is the golden sun
and the light of god is gone

I2P-GAN w/ Dm
he sun is singing in the forest rain
and like the wind in its warm june
is still my heart is a falcon
but it sings in the night
i feel like the meaning it

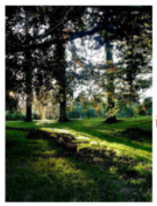
SeqGAN
the sun is shining on the sea
shining on the wind and a sudden
green
and round the little little boy they said
and look at the little noises the with a
with a coffee a silver penny a huge a
drum
steer the spider and you

Regions-Hierarchical
the sun was shining on the sea
the waves are crashing in
the light's spilled out of heaven and
flowing growing
the sun is warm and slow

I2P-GAN w/ Dp
the sun is singing
the sound is raining
i will catch you
don't look at you
you are the sky
you're beginning

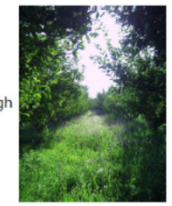
I2P-GAN
the sun is singing in the forest wind
and let us go to the wind of the sun
let the sun be free
let us be the storm of heaven
and let us be the slow sun
we keep our own strength together
we live in love and hate

Figure 4: Example of poems generated by eight methods for an image. Words in read indicate poeticness.

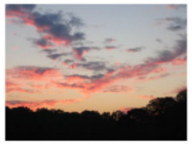


the sun rays struck my face
warm tingles to my fingertips
the light showed me a path
i should walk down
i spoke and the whispers of the breeze
told me to close my eyes
i lost my way in a paradise

i have been a great city
spinning and shout
the sound of the road
washed away
the mountain passes through
the streets are gone
the silence is raining
it sits still in silence
glint its own



the sun is a beautiful thing
in silence is drawn
between the trees
only the beginning of light



i will arise and go by the sea gate i watch it fly
and let me lie in the dark green valleys
and let me sing to the sun
i know you are not beautiful enough to me
and you know that you are so much
you are so much you can see you
love you will never fly
if you are always

when the sun shines through the snow and the night
is somebody really need you
if only can be really always just
just as if you can be a different way to be
walking through the sky is it possible
you know that is you want
to be pure as you can see
the light leaves you can see
the cherry blossoms

and now i am tired of my own
let me be the freshening blue
haunted through the sky bare
and cold water
warm blue air shimmering
brightly never arrives
it seems to say

the sun is shining
the wind moves
naked trees
you dance

i have seen the wind blows
my heart filled with air
my eyes are bleared with stinging
i am
a woman is a reminder
of love that's just
just one thing

Figure 5: Example of poems generated by our approach I2P-GAN.

Beyond Narrative Description: Generating Poetry from Images by Multi-Adversarial Training

Bei Liu*
Kyoto University
liubei@dl.kuis.kyoto-u.ac.jp

Jianlong Fu*
Microsoft Research Asia
jianf@microsoft.com

Makoto P. Kato
Kyoto University
mpkato@acm.org

Masatoshi Yoshikawa
Kyoto University
yoshikawa@i.kyoto-u.ac.jp

AUCH GEDICHTE?

the sun is shining
the wind moves
naked trees
you dance

Beyond Narrative Description: Generating Poetry from Images by Multi-Adversarial Training

Bei Liu*
Kyoto University
liubei@dl.kuis.kyoto-u.ac.jp

Jianlong Fu†
Microsoft Research Asia
jianf@microsoft.com

Makoto P. Kato
Kyoto University
mpkato@acm.org

Masatoshi Yoshikawa
Kyoto University
yoshikawa@i.kyoto-u.ac.jp

KÜNSTLICHE SOFTWARE
Wenn das Netz das Netz trainiert.

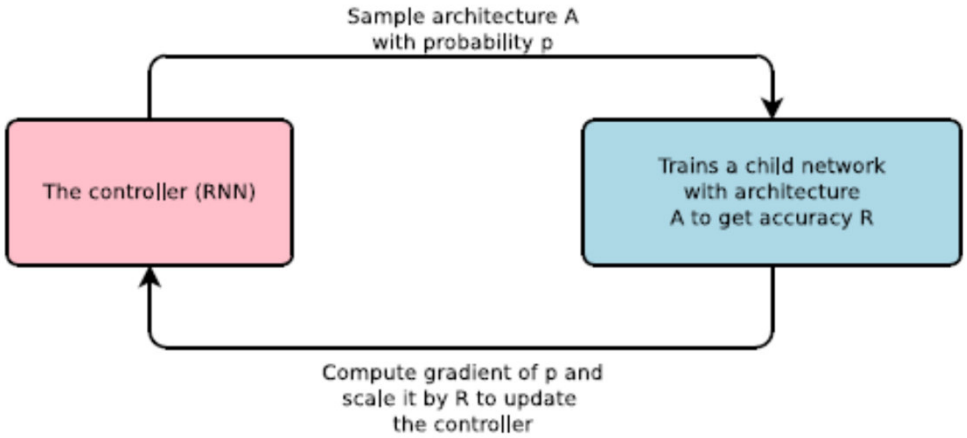


Figure 1: An overview of Neural Architecture Search.

Quelle: Zoph and Le (Google Brain): Neural Architecture Search with Reinforcement Learning. Arxiv: 1611.01578

Model	Depth	Parameters	Error rate (%)
Network in Network (Lin et al., 2013)	-	-	8.81
All-CNN (Springenberg et al., 2014)	-	-	7.25
Deeply Supervised Net (Lee et al., 2015)	-	-	7.97
Highway Network (Srivastava et al., 2015)	-	-	7.72
Scalable Bayesian Optimization (Snoek et al., 2015)	-	-	6.37
FractalNet (Larsson et al., 2016)	21	38.6M	5.22
with Dropout/Drop-path	21	38.6M	4.60
ResNet (He et al., 2016a)	110	1.7M	6.61
ResNet (reported by Huang et al. (2016c))	110	1.7M	6.41
ResNet with Stochastic Depth (Huang et al., 2016c)	110	1.7M	5.23
	1202	10.2M	4.91
Wide ResNet (Zagoruyko & Komodakis, 2016)	16	11.0M	4.81
	28	36.5M	4.17
ResNet (pre-activation) (He et al., 2016b)	164	1.7M	5.46
	1001	10.2M	4.62
DenseNet ($L = 40, k = 12$) Huang et al. (2016a)	40	1.0M	5.24
DenseNet($L = 100, k = 12$) Huang et al. (2016a)	100	7.0M	4.10
DenseNet ($L = 100, k = 24$) Huang et al. (2016a)	100	27.2M	3.74
DenseNet-BC ($L = 100, k = 40$) Huang et al. (2016b)	190	25.6M	3.46
Neural Architecture Search v1 no stride or pooling	15	4.2M	5.50
Neural Architecture Search v2 predicting strides	20	2.5M	6.01
Neural Architecture Search v3 max pooling	39	7.1M	4.47
Neural Architecture Search v3 max pooling + more filters	39	37.4M	3.65

Table 1: Performance of Neural Architecture Search and other state-of-the-art models on CIFAR-10.

TÄUSCHUNG VON KI

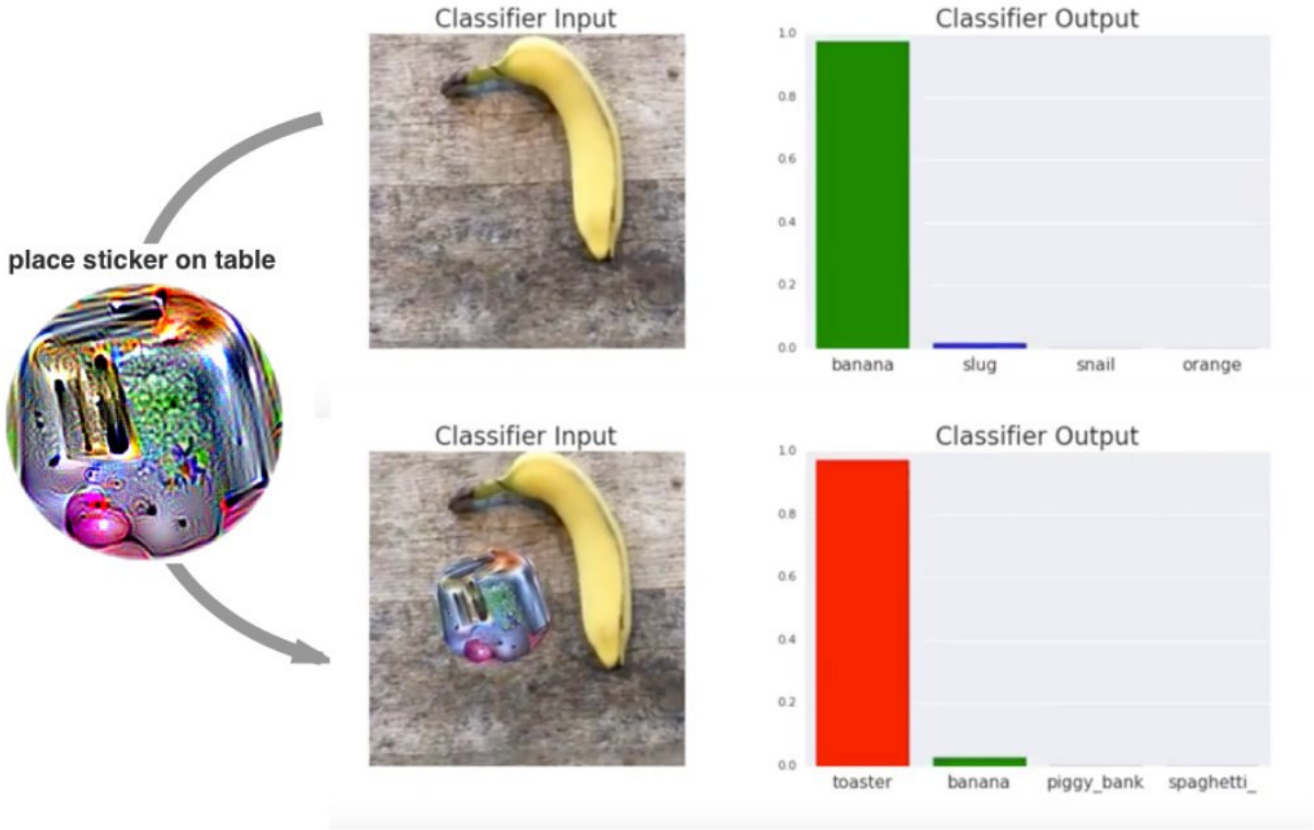
TÄUSCHUNG VON KI

T · · Systems · Let's power higher performance

KANN MAN KI TÄUSCHEN?
Wie man zur Hyäne oder auch zum Killerwal wird - DEMO.

Creating Adversarial Examples – Using Tensorflow. Source: https://github.com/Hvass-Labs/TensorFlow-utorials Pedersen

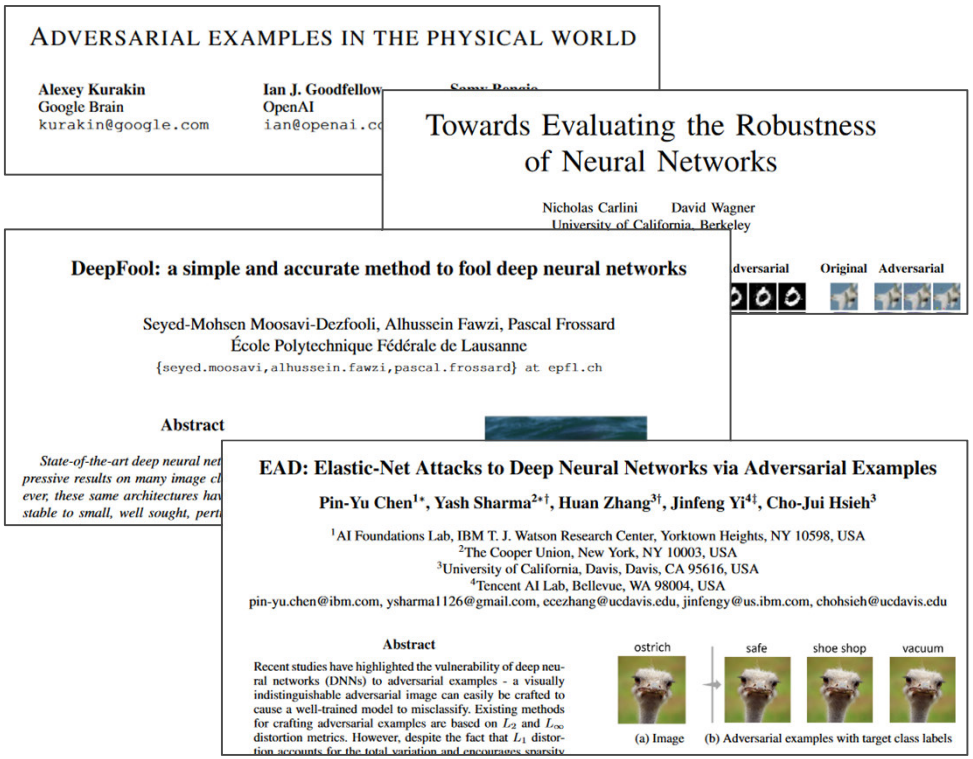
KANN MAN KI TÄUSCHEN?
Man kann.



Adversarial Patch [Tom B. Brown, Dandelion Mané, Aurko Roy, Martín Abadi and Justin Gilmer/Neural Information Processing Systems 2017]

We present a method to create universal, robust, targeted adversarial image patches in the real world. The patches are universal because they can be used to attack any scene, robust because they work under a wide variety of transformations, and targeted because they can cause a classifier to output any target class. These adversarial patches can be printed, added to any scene, photographed, and presented to image classifiers; even when the patches are small, they cause the classifiers to ignore the other items in the scene and report a chosen target class.

MAN KANN MAN KI TÄUSCHEN!
Adversarial Examples – Von FGsM zu DeepFool.



ADVERSARIAL EXAMPLES IN THE PHYSICAL WORLD

Alexey Kurakin
Google Brain
kurakin@google.com

Ian J. Goodfellow
OpenAI
ian@openai.co

Towards Evaluating the Robustness of Neural Networks

Nicholas Carlini David Wagner
University of California, Berkeley

DeepFool: a simple and accurate method to fool deep neural networks

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Pascal Frossard
École Polytechnique Fédérale de Lausanne
{seyed.moosavi,alhussein.fawzi,pascal.frossard} at epfl.ch

Abstract

EAD: Elastic-Net Attacks to Deep Neural Networks via Adversarial Examples

Pin-Yu Chen[1]*, Yash Sharma[2]*†, Huan Zhang[3]†, Jinfeng Yi[4]‡, Cho-Jui Hsieh[3]

[1]AI Foundations Lab, IBM T. J. Watson Research Center, Yorktown Heights, NY 10598, USA
[2]The Cooper Union, New York, NY 10003, USA
[3]University of California, Davis, Davis, CA 95616, USA
[4]Tencent AI Lab, Bellevue, WA 98004, USA
pin-yu.chen@ibm.com, ysharma1126@gmail.com, ecezhang@ucdavis.edu, jinfengy@us.ibm.com, chohsieh@ucdavis.edu

Abstract

Recent studies have highlighted the vulnerability of deep neural networks (DNNs) to adversarial examples - a visually indistinguishable adversarial image can easily be crafted to cause a well-trained model to misclassify. Existing methods for crafting adversarial examples are based on L_2 and L_∞ distortion metrics. However, despite the fact that L_1 distortion accounts for the total variation and encourages sparsity

ostrich | safe | shoe shop | vacuum

(a) Image (b) Adversarial examples with target class labels

Adversarial Examples:

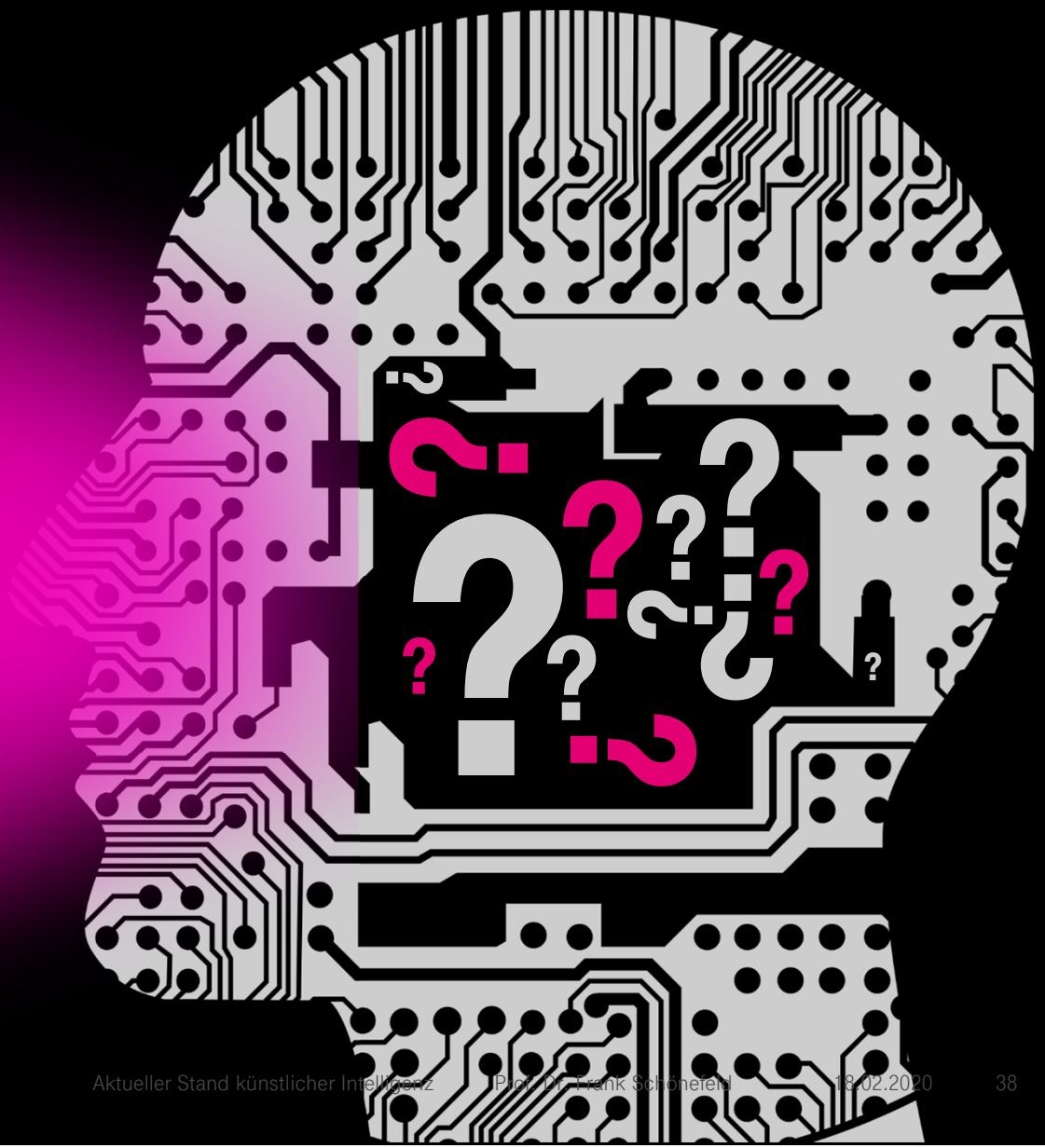
Sammlung von Methoden, um KI (Neuronale Netze) systematisch in die Irre zu führen.

- Blackbox vs Whitebox Methoden
- Iterative vs Non-iterative Methoden
- Targeted Attacks vs Non-Targeted Attacks

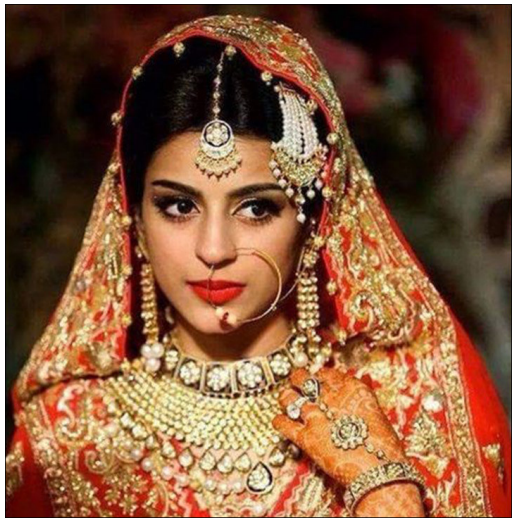
- Adversarial Examples Methods
  - Basic iterative Method
  - Carlini-Wagner Method
  - DeepFool Method
  - Fast Feature Gradient Method
  - Fast Gradient (Sign) Method
  - Projected Gradient Descent Method
  - Momentum Difference Method
  - Jacobian-based saliency Method
  - ...see

Source: https://cleverhans.readthedocs.io/en/latest/source/attacks.html

WEITERE OFFENE FRAGEN

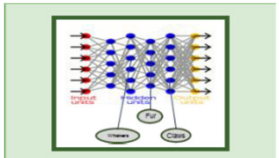


DEBIASING. EXPLAINABILITY. AI ETHICS.
Ich weiß nicht, was soll es bedeuten.

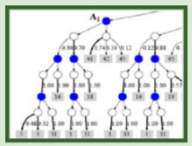


Oriental fairy tale or

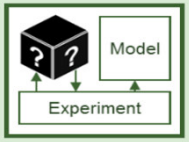
Indian Bride?



Deep Explanation
Modified deep learning techniques to learn explainable features



Interpretable Models
Techniques to learn more structured, interpretable, causal models



Model Induction
Techniques to infer an explainable model from any model as a black box

Source: Gunning, DARPA. XAI: Explainable Artificial Intelligence. TTI Vanguard June 2018.

KONZERN | 11.05.2018 | VERENA FULDE | 2 KOMMENTARE

Die neun Leitlinien der Telekom zum Einsatz von künstlicher Intelligenz

Teilen Drucken Text vorlesen

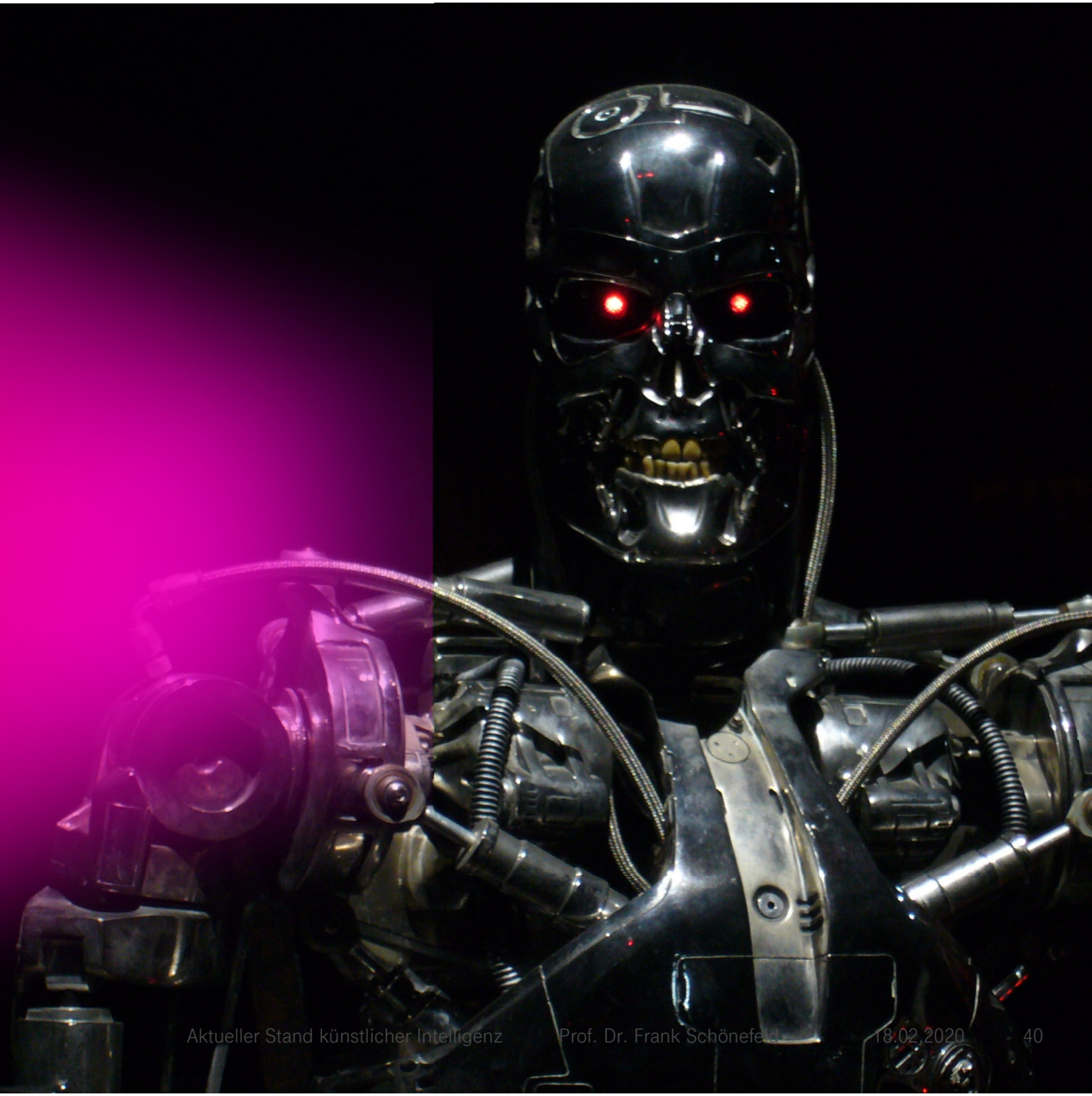
Diese Leitlinien sind auf unsere Geschäftsmodelle ausgerichtete, selbst-bindende Richtlinien. Sie definieren, wie wir als Deutsche Telekom mit KI umgehen wollen und wie wir unsere auf KI basierenden Produkte und Services künftig entwickeln.

Präambel

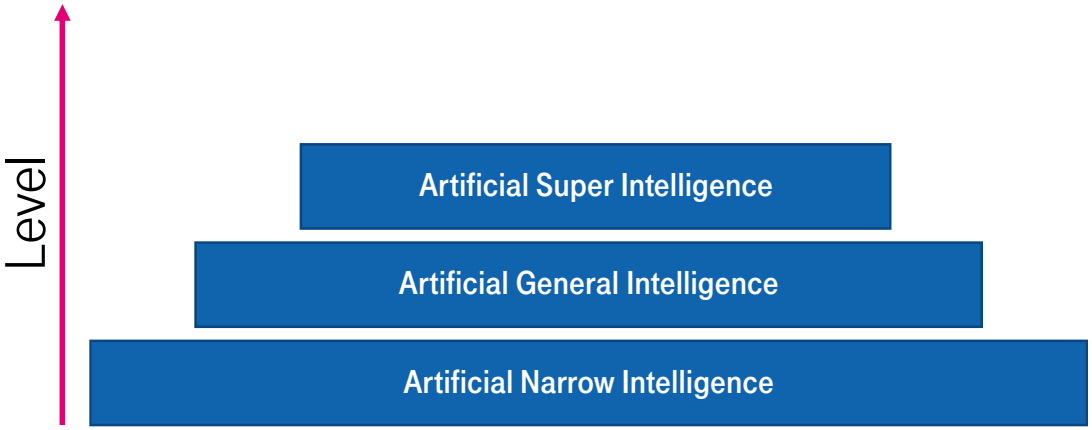
Künstliche Intelligenz muss unsere Handlungen unterstützen. KI soll den Menschen unterstützen und seine Fähigkeiten erweitern, anstatt ihn einzuschränken.

1. **Wir übernehmen Verantwortung.**
   Klare Definition, wer verantwortlich für welches KI-System ist.
2. **Wir gehen sorgsam mit Künstlicher Intelligenz um.**
   Unsere KI-Systeme und ihre Verwendung folgen dem für die Menschen geltenden Recht und Gesetz.
3. **Wir stellen unsere Kunden in den Mittelpunkt.**
   KI soll das Leben unserer Kunden vereinfachen und bereichern.
4. **Wir stehen für Transparenz.**
   Transparenz, wenn ein Kunde mit einem KI-System kommuniziert und Transparenz über die Nutzung der Kundendaten.
5. **Wir bieten Sicherheit.**
   Unsere Kundendaten sind vor unautorisiertem, externen Zugriff geschützt (Datensicherheit und Datenschutz).
6. **Wir legen das Fundament.**
   Gründliche Analyse und Evaluierungen als Basis für die Weiterentwicklung und stete Verbesserung unserer KI-Systeme.
7. **Wir behalten den Überblick.**
   Ständige Bereitschaft, in unsere KI-Systeme einzugreifen, um Schäden zu vermeiden bzw. zu reduzieren.
8. **Wir leben das Kooperationsmodell.**
   Obwohl wir den Menschen an erster Stelle sehen, können Vorteile aus einer Interaktion von Mensch und Maschine gezogen werden.
9. **Wir teilen und erklären.**
   Verbreitung von Wissen über KI und Vermittlung von relevanten Kompetenzen hierzu.

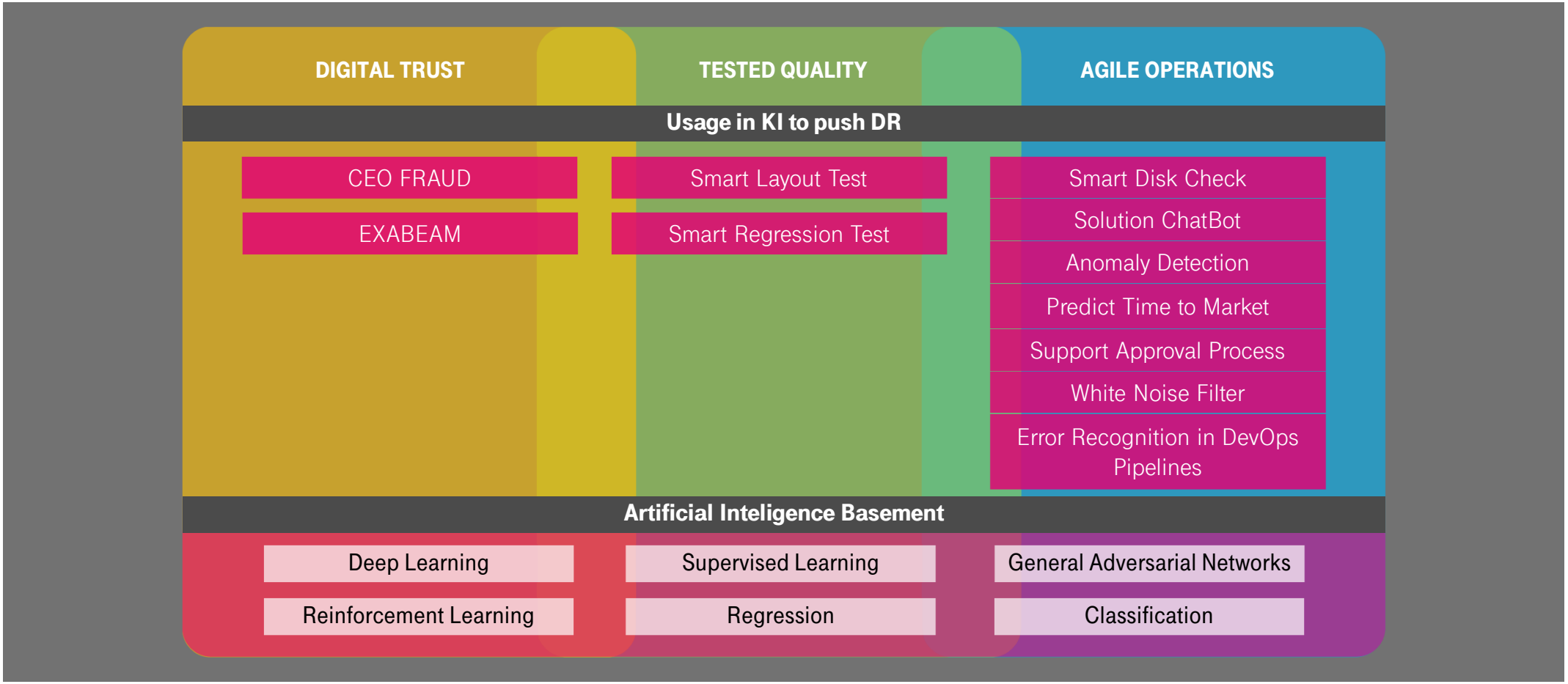
WOHIN FÜHRT DAS?
BEDROHUNGEN?



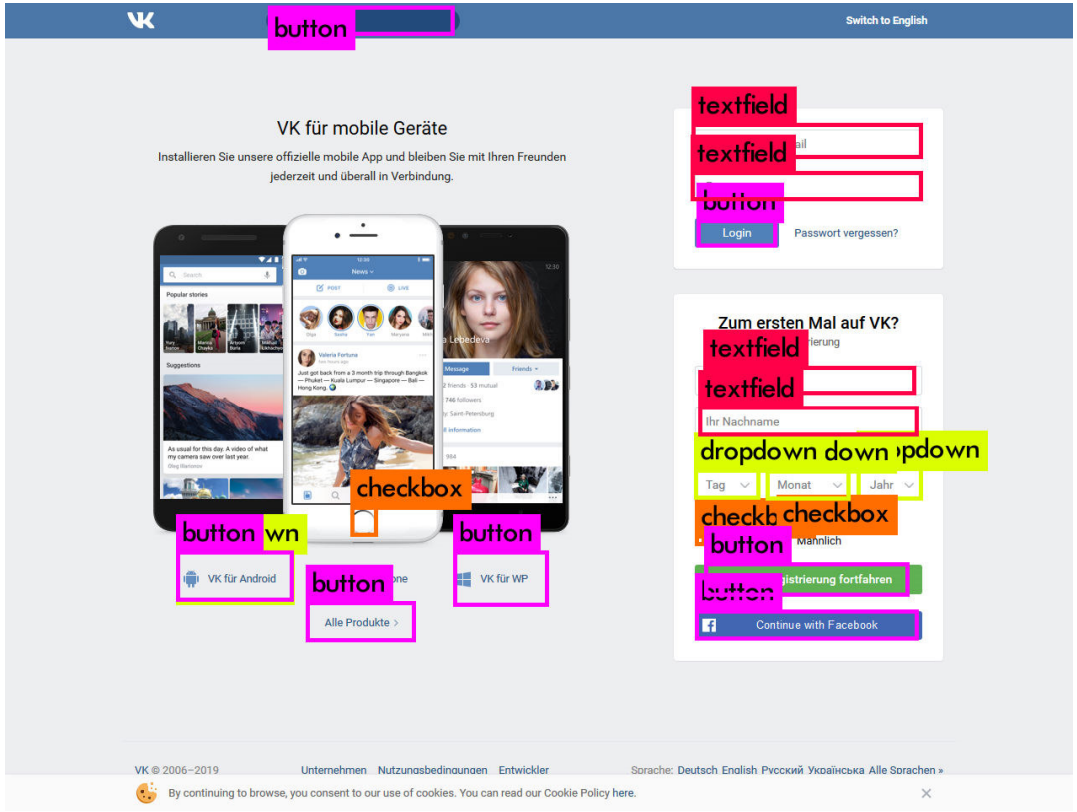
STUFEN DER KÜNSTLICHEN INTELLIGENZ.



KI IN (MY) BUSINESS



SMART REGRESSION TESTS



FAZIT

FAZIT
Going Deeper.

1. **Künstliche Intelligenz** hat seit 2012 enorme Fortschritte gemacht und insbesondere im Bereich **Erkennung/Perception (Wahrnehmung)** dramatische Verbesserungen erreicht. Diese beruhen insbesondere auf der Anwendung der **DeepLearning** (und hier insbesondere CNNs) Technologie für Bild-, Video- und Spracherkennung (**Attention based Mechanisms**). Ein weiterer Fortschritt für Lernen ohne Vorwissen wurde mit **Reinforcement Lernansätzen** erreicht. (GO)

2. Diese Ansätze führen zu vielen **verbesserten Anwendungen**, die auf ausreichend stabiler Erkennung basieren. Nach wie vor können diese Ansätze **überlistet** werden (adversarial examples). Sie liefern heute eng umrissene Anwendungen, der Lernfortschritt basiert auf einer großen Menge an Daten.

3. Obwohl auch in Gebieten, die man gemeinhin mit Kreativität verbindet (Musik, Malstile, Software-Schreiben) zum Teil überraschende Ergebnisse vorliegen ist das Erreichen fortgeschrittener, **allgemeiner künstlicher Intelligenz** noch **sehr weit entfernt**. Für viele Fragen (z.B. eigener Wille, Selbst-Bewusstsein, emotionale Intelligenz) gibt es keine (kaum) Forschungsansätze.

Source: https://www.horsesforsources.com/storage/app/media/OBrien/Digital%20customer%20stories.jpg

https://www.kunsthalle-kuehlungsborn.de/category/ausstellungen/

Vielen Dank

Kontakt:

Prof. Dr. Frank Schönefeld
T-Systems Multimedia Solutions
Geschäftsleitung

+49 351 2820 2500
Frank.Schoenefeld@t-systems.com

T · · Systems · Let's power higher performance