Jürgen Schmidhuber
You_again Shmidhoobuh
Ultimate Trend
Will history converge around 2050 = Ω?

Pattern starts at Ω - 13.8 B years:
Big Bang
Ω - $\frac{1}{4}$ of this time

Ω - 3.5 B years:
Life
| $\Omega$ - $\frac{1}{4}$ of this time | $\Omega$ - 0.9 B years: Animal-like life |
$\Omega - \frac{1}{4}$ of this time  \hspace{1cm} $\Omega - 220$ M years: Mammals
Ω - ¼ of this time

Ω - 55 M years:
Primates
Ω - $\frac{1}{4}$ of this time

Ω - 13 M years:
Hominids
$\Omega - \frac{1}{4}$ of this time

$\Omega - 3.5$ M years:

Stone tools
$\Omega - \frac{1}{4}$ of this time

$\Omega - 850,000$ years:
Controlled fire
Ω - 210,000 years: Anatomically modern man

Ω - ¼ of this time
Ω - \( \frac{1}{4} \) of this time

Ω - 50,000 years:
Behaviorally
modern man
Ω - ¼ of this time

Ω - 13,000 years:
Neolithic revolution
Ω - ¼ of this time

Ω - 3,300 years: Iron age
$\Omega$ - ¼ of this time

$\Omega$ - 800 years:
Guns & rockets
Ω - ¼ of this time

Ω - 200 years:
Industrial revolution
$\Omega - \frac{1}{4}$ of this time

$\Omega - 50$ years (now):
Information revolution
Small computers with 1 brain power

- 12 years
- 3 years
- 9 months
- 10 weeks

http://www.idsia.ch/~juergen/history.html
Deep Learning is a half century old although recent “tabloid science” stories claim it is a recent thing

888 references, 88 pages:
http://www.idsia.ch/~juergen/deep-learning-overview.html
Critique (also at Google+) of paper by self-proclaimed “deep learning conspiracy” (LeCun & Bengio & Hinton) who cite each other but not the pioneers of the field: http://www.idsia.ch/~juergen/deep-learning-conspiracy.html
Father of Deep Learning
Ivakhnenko et al, since 1965
Deep multilayer perceptrons with polynomial activation functions
Incremental layer-wise training by regression analysis - learn numbers of layers and units per layer - prune superfluous units 8 layers already back in 1971 still used in the 2000s
Who introduced the term “deep learning” to Machine Learning and Neural Networks?

- Dechter, 1986 (ML)
- Aizenberg et al, 2000 (NNs)
- Gomez & Schmidhuber (2005): first NN paper with word combination “learn deep” in title
who invented backpropagation?

http://www.idsia.ch/~juergen/who-invented-backpropagation.html
Supervised Backpropagation (BP)

The deepest NNs:
RNNs are general computers
Learn program = weight matrix

http://www.idsia.ch/~juergen/rnn.html
1991: SEPP HOCHREITER’S ANALYSIS OF THE FUNDAMENTAL DEEP LEARNING PROBLEM

\[
\left\| \frac{\partial e(t-q)}{\partial e(t)} \right\| = \left\| \prod_{m=1}^{q} \left\| WF'(Net(t-m)) \right\| \right\|
\]

\[
\leq (\left\| W \right\| \max_{Net} \left\{ \left\| F'(Net) \right\| \right\})^q
\]

COMPARE: HOCHREITER & BENGIO & FRASCONI & SCHMIDHUBER, 2001

http://www.idsia.ch/~juergen/fundamentaldeeplearningproblem.html

http://www.idsia.ch/~juergen/firstdeelearner.html
With Hochreiter, Gers, Graves, Fernandez, Gomez, Bayer...

1997-2007. Since 2015 on your phone! Google, Microsoft, IBM, others, all use LSTM now
Today’s LSTM RNNs shaped by:

**Ex-PhD students** (TUM & IDSIA):
- Sepp Hochreiter (PhD 1999)
- Felix Gers (PhD 2001)
- Alex Graves (PhD 2008)
- Daan Wierstra (PhD 2010)
- Justin Bayer (2009), others

**Postdocs** at IDSIA (2000s):
- Fred Cummins
- Santiago Fernandez
- Faustino Gomez
- Others
Connectionist Temporal Classification (CTC): Graves, Fernandez, Gomez, Schmidhuber ICML 2006

No pre-segmented data; RNN maximises probability of training set label sequences

\[ O_{ML}(S) = - \sum_{(x,z) \in S} \ln(p(z|x)) \]
LSTM: First RNN to win contests: 3 ICDAR 2009 connected handwriting competitions

E.g., Graves & Schmidhuber NIPS 2010

http://www.idsiach/~juergen/handwriting.html
LSTM for speech: 2003 as good as HMMs, 2007: LSTM stack gets best results on keyword spotting in a large corpus (vs HMMs). Today: best large vocabulary speech recognition ...
BAIDU’s DeepSpeech uses our CTC-based RNNs for end-to-end speech recognition without any HMMs / GMMs (Hannun et al., Baidu, 2014); broke Switchboard benchmark record.

A dozen of the many 2014/2015 benchmark records with LSTM RNNs, often at major IT companies:

1. Large vocabulary speech recognition (Sak et al., Google, Interspeech 2014)
2. English to French translation (Sutskever et al., Google, NIPS 2014)
3. Text-to-speech synthesis (Fan et al., Microsoft, Interspeech 2014)
4. Prosody contour prediction (Fernandez et al., IBM, Interspeech 2014)
5. Google Voice improved by 49% (Sak et al, 2015, now for >1 billion users)
6. Syntactic parsing for NLP (Vinyals et al., Google, 2014)
7. Photo-real talking heads (Soong and Wang, Microsoft, ICASSP 2015)
8. Social signal classification (Brueckner & Schulter, ICASSP 2014)
9. Arabic handwriting recognition (Bluche et al., DAS 2014)
10. Image caption generation (Vinyals et al., Google, 2014)
11. Keyword spotting (Chen et al., Google, ICASSP 2015)
12. Video to textual description (Donahue et al., 2014; Li Yao et al., 2015)

http://www.idsia.ch/~juergen/rnn.html
1993: Gradient-based meta-RNNs that can learn to run their own weight change algorithm: J. Schmidhuber. A self-referential weight matrix. ICANN 1993

This was before LSTM. In 2001, however, Hochreiter taught a meta-LSTM to learn a learning algorithm for quadratic functions that was faster than backprop
MNIST: 60,000 digits for training, 10,000 for testing, 7 layer MLP; >12m weights; train 200 days on CPU = 5 on GPU; >$10^{15}$ weight updates, 5B/s, 2010: new world record 0.35% (Ciresan et al.) Since then: decline of unsupervised pre-training for FNNs, like in the 1990s for RNNs

Two old ideas: backprop (3-5 decades old), training pattern deformations (Baird, 1990, 2 decades old)
<table>
<thead>
<tr>
<th>Unsupervised ➔ Supervised</th>
</tr>
</thead>
<tbody>
<tr>
<td>1990s: Trend from unsupervised to supervised RNNs</td>
</tr>
<tr>
<td>Both trends driven by our team</td>
</tr>
</tbody>
</table>
Our Deep GPU-Based Max-Pooling CNNs (IJCAI 2011)
e.g., http://www.idsia.ch/~juergen/deeplearning.html

ICDAR 2011 offline Chinese handwriting recognition contest (4000 classes):
1st & 2nd rank

http://www.idsia.ch/~juergen/handwriting.html
Ensembles of deep sparse CNNs + Max-Pooling + MLP on top: 1 year on CPU = 1 week on GPU. 2011-2012: first human-competitive MNIST result: 0.2% (after almost a decade of ~0.4%).

Ciresan, Meier, Masci, Gambardella, Schmidhuber, IJCAI 2011, IJCNN 2011, CVPR 2012
Traffic Sign Contest, Silicon Valley, 2011:
Our GPU-MPCNN was twice better than humans
3 times better than closest artificial competitor
6 times better than best non-neural thing: FIRST

SUPERHUMAN VISUAL PATTERN RECOGNITION

http://www.idsia.ch/~juergen/superhumanpatternrecognition.html
IJCNN 2011 traffic sign recognition competition, Silicon Valley, 2011: 
1\textsuperscript{ST} (0.56\% ERROR) 
2\textsuperscript{ND} HUMANS (1.16\%) 
3\textsuperscript{RD} (1.69\%) 
4\textsuperscript{TH} (3.86\%) 
Ciresan, Meier, Masci, Schmidhuber, IJCNN 2011, Neural Networks, 2012

Very similar GPU-MPCNNs later used for ImageNet (Krizhevsky & Hinton 2012, Zeiler & Fergus 2013, …)
Ernst Dickmanns, the robot car pioneer, Munich, 80s

1995: Munich to Denmark and back on public Autobahns, up to 180 km/h, no GPS, passing other cars

2014: 20 year anniversary of self-driving cars in highway traffic

http://www.idsia.ch/~juergen/robotcars.html
Our Deep Learner Won ISBI 2012 Brain Image Segmentation Contest:
First feedforward Deep Learner to win an image segmentation competition
(but compare deep recurrent LSTM 2009: segmentation & classification)

http://www.idsia.ch/~juergen/deeplearningwinsbraincontest.html
DEEP LEARNING WINS
MICCAI 2013 GRAND CHALLENGE
ON MITOSIS DETECTION

http://www.idsia.ch/~juergen/deeplearningwinsMICCAIgrandchallenge.html
Thanks to Dan Ciresan & Alessandro Giusti
Some of Our Deep Learning “Firsts”

- First recurrent NN to win contests (2009)
- First NN to win connected handwriting contests (2009)
- First outperformance of humans in a computer vision contest (2011)
- First deep NN to win Chinese handwriting contest (2011)
- European handwriting (MNIST): old error record almost halved (2011)
- First deep NN to win image segmentation contest (2012)
- First deep NN to win object detection contest (2012)
- First deep NN to win medical imaging contest (2012)
- First RNN controller that reinforcement learns from raw video (2013)
- …
Image caption generation with LSTM RNNs translating internal representations of CNNs (Vinyals, Toshev, Bengio, Erhan, Google, 2014)
Best Segmentation with PyramMiD-LSTM (NIPS 2015)
LSTM learns knot-tying tasklets: Mayr Gomez Wierstra Nagy Knoll Schmidhuber, IROS’06
Reinforcement Learning in Partially Observable Worlds

Finds Complex Neural Controllers with a Million Weights – RAW VIDEO INPUT!
Faustino Gomez, Jan Koutnik, Giuseppe Cuccu, J. Schmidhuber, GECCO 2013
Octopus-arm control: 82 in, 32 out, 3'680 weights, only 20 DCT coefficients, compression 1:184

Octopus-arm with low-level vision, 32x32 in, 32 out, 33'824 weights, 160 DCT, compression 1:211

TORCS driving video game, low-level vision, 64x64 in, 3 out, 1'115'139 weights, 200 DCT, compression 1:5575

http://www.idsia.ch/~juergen/compressednetworksearch.html
The first 4 members of DeepMind include 2 former PhD students of my lab. But I am not happy with their Nature paper, although 3 of its authors were trained here, because others at IDSIA published Reinforcement Learning with high-dimensional video input earlier.
No new NN winter, because physics dictates that future hardware will be 3D-RNN-like: many processors connected by many short and few long wires

http://www.idsia.ch/~juergen/rnn.html
IJCNN 1990, NIPS 1991: Reinforcement Learning with Recurrent Controller & Recurrent World Model

A bit like AIXI, but with feasible local search
IJNS 1991: R-Learning of Visual Attention on 1,000,000 times slower computers

Fig. 1. A typical visual scene. The diameters of the receptive fields of the retina’s input units are indicated by circles.

Fig. 2. An artificial fovea provides inputs for a control network which is able to move the fovea around. A model network is trained to predict the next input from the current input and the current controller action.
1991: current goal = extra fixed input
2015: all of this is coming back!

Fig. 1. Translations: Examples of fovea trajectories leading from various start positions to different targets.

Fig. 5. One controller for various targets specified by an additional constant input: Examples of fovea trajectories leading from various start positions to different targets. The first target is near the left corner of the triangle. The second target is near the lower corner.
RoboCup World Champion 2004, Fastest League, 5m/s

Lookahead expectation & planning with neural networks (Schmidhuber, IEEE INNS 1990): successfully used for RoboCup by Alexander Gloye-Förster (went to IDSIA)
http://www.idsia.ch/~juergen/learningrobots.html

Alex @ IDSIA, led FU Berlin’s RoboCup World Champion Team 2004
Formal theory of fun & novelty & surprise & attention & creativity & curiosity & art & science & humor

Maximize Future Fun(Data X,O(t))~ \[ \frac{\partial \text{CompResources}(X,O(t))}{\partial t} \]

http://www.idsia.ch/~juergen/creativity.html
PowerPlay not only solves but also continually invents problems at the borderline between what's known and unknown - training an increasingly general problem solver by continually searching for the simplest still unsolvable problem.
True Artificial Intelligence Will Change Everything

Jürgen Schmidhuber
The Swiss AI Lab IDSIA
Univ. Lugano & SUPSI
http://www.idsia.ch/~juergen

NNAISENSE
Next: build small animal-like AI that learns to think and plan hierarchically like a crow or a capuchin monkey

Evolution needed billions of years for this, then only a few more millions for humans
neural networks-based artificial intelligence

THE DAWN OF AI
Open Source Neural Networks Library by my PhD students K Greff and R Srivastava

http://people.idsia.ch/~juergen/brainstorm.html