DevCon School

Технологии будущего

DevCon School
Технологии будущего

Microsoft Cognitive Toolkit – инструментарий для проектирования и обучения нейронных сетей

Елизавета Лаврова
Технологический евангелист, Microsoft

Цели

Определение

ИИ, когнитивность, нейронные сети

Демократизация ИИ

Готовый набор инструментов для применения нейросетевых технологий

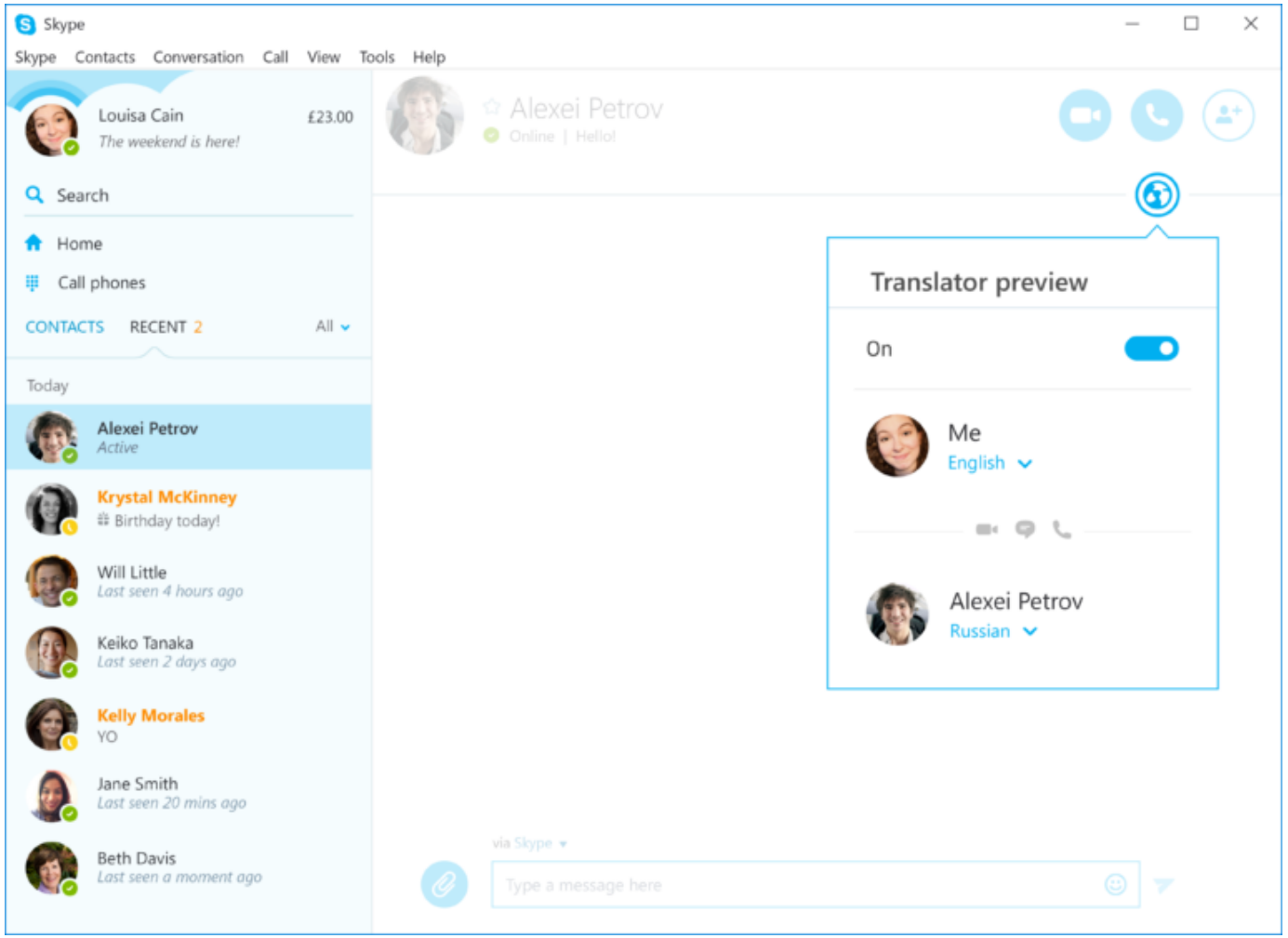
Практика

Решение задачи с использованием нейросетевого инструментария

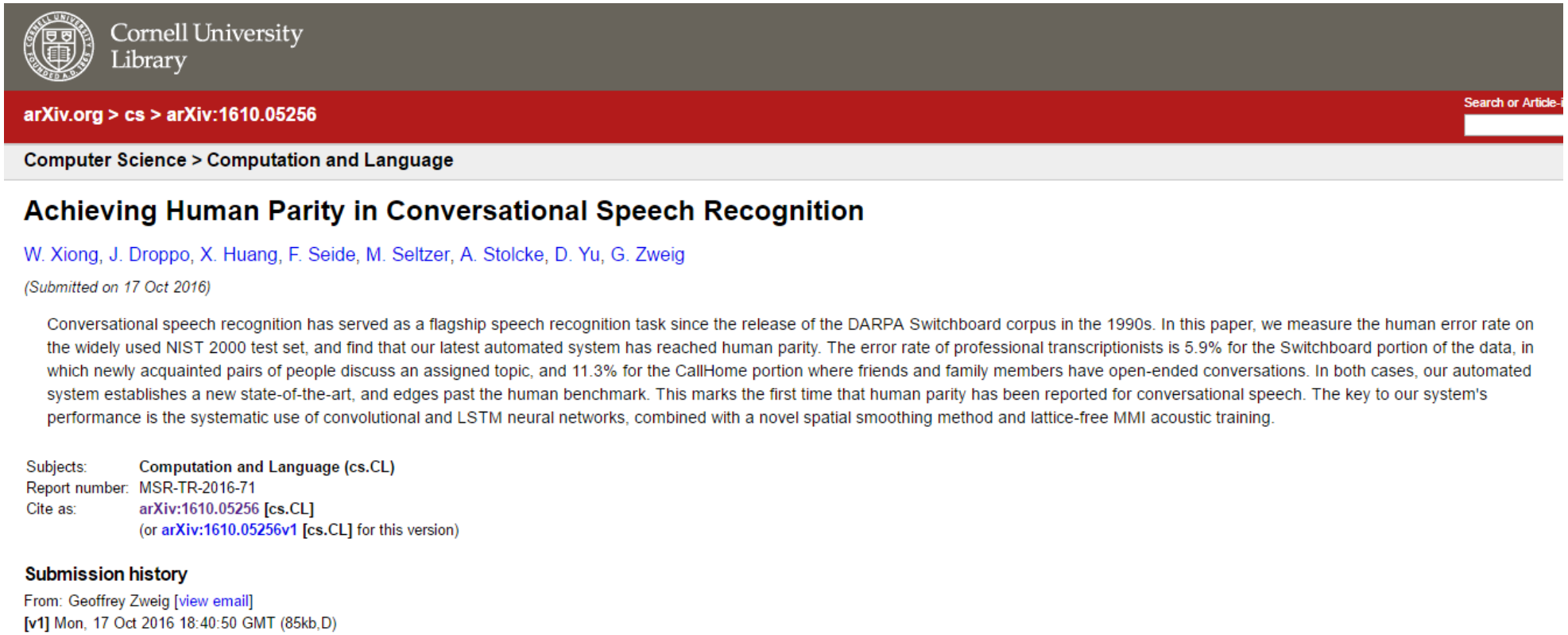
ИИ вокруг нас

ИИ вокруг нас

ИИ вокруг нас



ИИ вокруг нас



Cornell University Library

arXiv.org > cs > arXiv:1610.05256

Search or Article-

Computer Science > Computation and Language

Achieving Human Parity in Conversational Speech Recognition

W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, G. Zweig

(Submitted on 17 Oct 2016)

Conversational speech recognition has served as a flagship speech recognition task since the release of the DARPA Switchboard corpus in the 1990s. In this paper, we measure the human error rate on the widely used NIST 2000 test set, and find that our latest automated system has reached human parity. The error rate of professional transcriptionists is 5.9% for the Switchboard portion of the data, in which newly acquainted pairs of people discuss an assigned topic, and 11.3% for the CallHome portion where friends and family members have open-ended conversations. In both cases, our automated system establishes a new state-of-the-art, and edges past the human benchmark. This marks the first time that human parity has been reported for conversational speech. The key to our system's performance is the systematic use of convolutional and LSTM neural networks, combined with a novel spatial smoothing method and lattice-free MMI acoustic training.

Subjects: **Computation and Language (cs.CL)**
Report number: MSR-TR-2016-71
Cite as: arXiv:1610.05256 [cs.CL]
(or arXiv:1610.05256v1 [cs.CL] for this version)

Submission history

From: Geoffrey Zweig [view email]
[v1] Mon, 17 Oct 2016 18:40:50 GMT (85kb,D)

Глубинное обучение в продуктах Microsoft

Когнитивные сервисы

https://how-old.net
http://www.captionbot.ai

Переводчик Skype

Bing

Cortana
реклама
релевантность
мультимедиа
…

HoloLens

Microsoft Research

речь, изображения, текст

Введение в ИНС

#msdevcon

DevCon School
Технологии будущего

Задачи машинного обучения

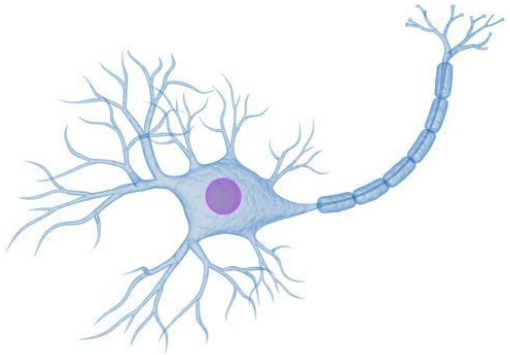
Когда мы используем нейронные сети

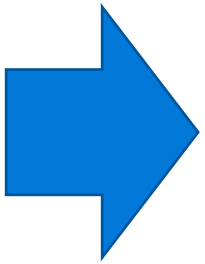
Объем данных

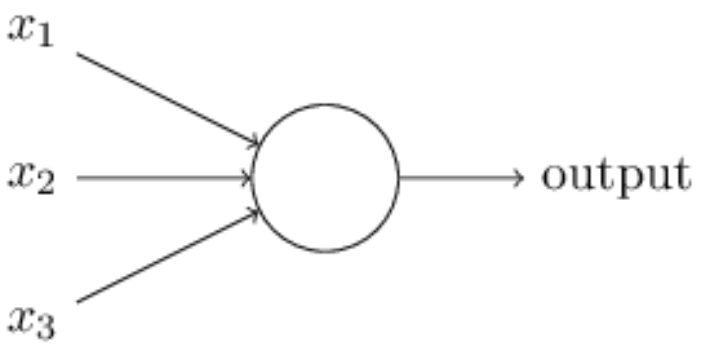
Вычислительные ресурсы

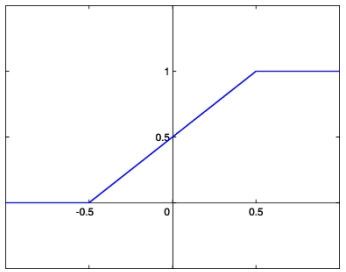
Алгоритмы

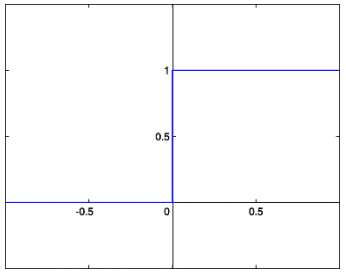
Формальный нейрон

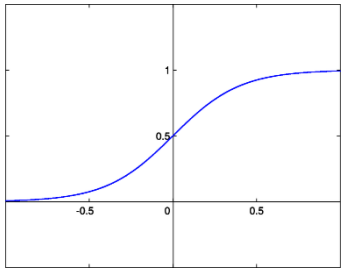




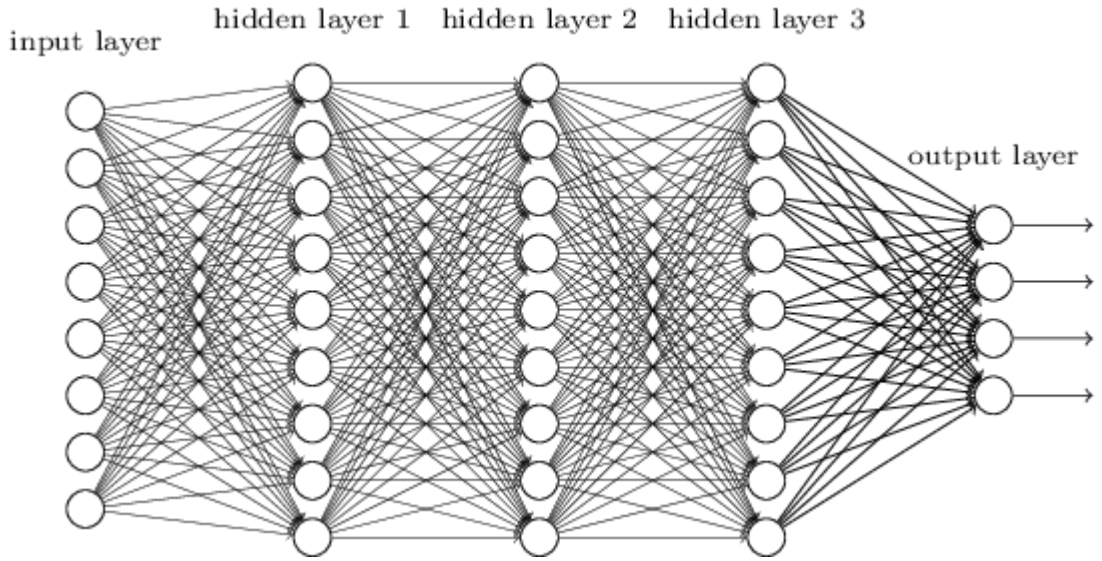








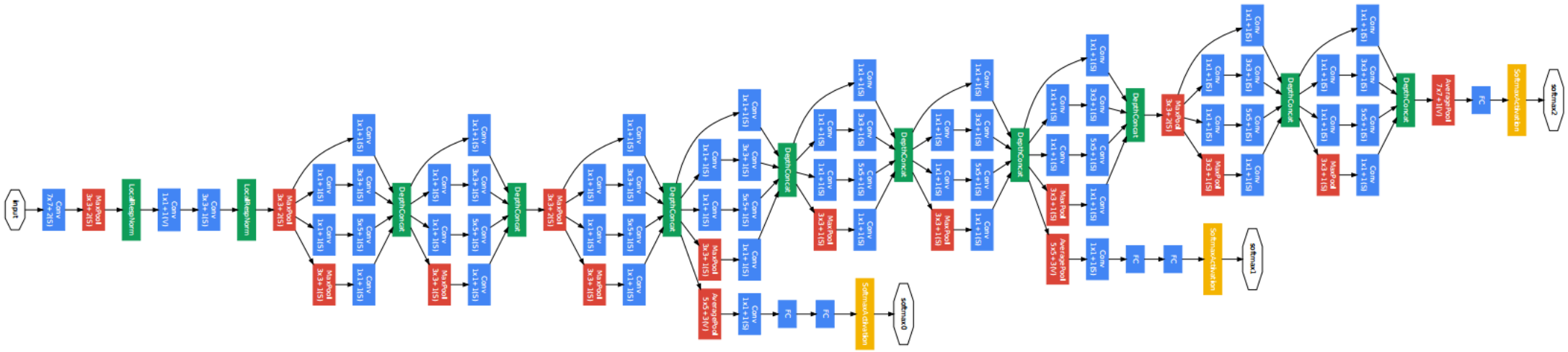
Нейронная сеть



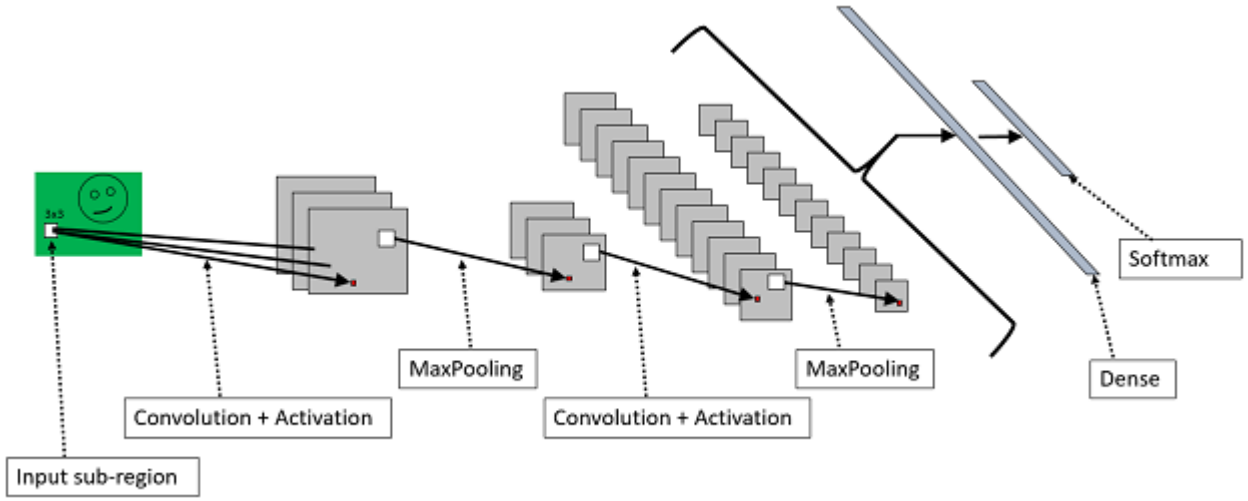
$$\Delta = F_{out} - F_{ref}$$

$$\frac{\partial \Delta}{\partial W_n} = \frac{\partial \Delta(F_n)}{\partial F_n} \cdot \frac{\partial F_n(F_{n-1}, W_n)}{\partial W_n}$$

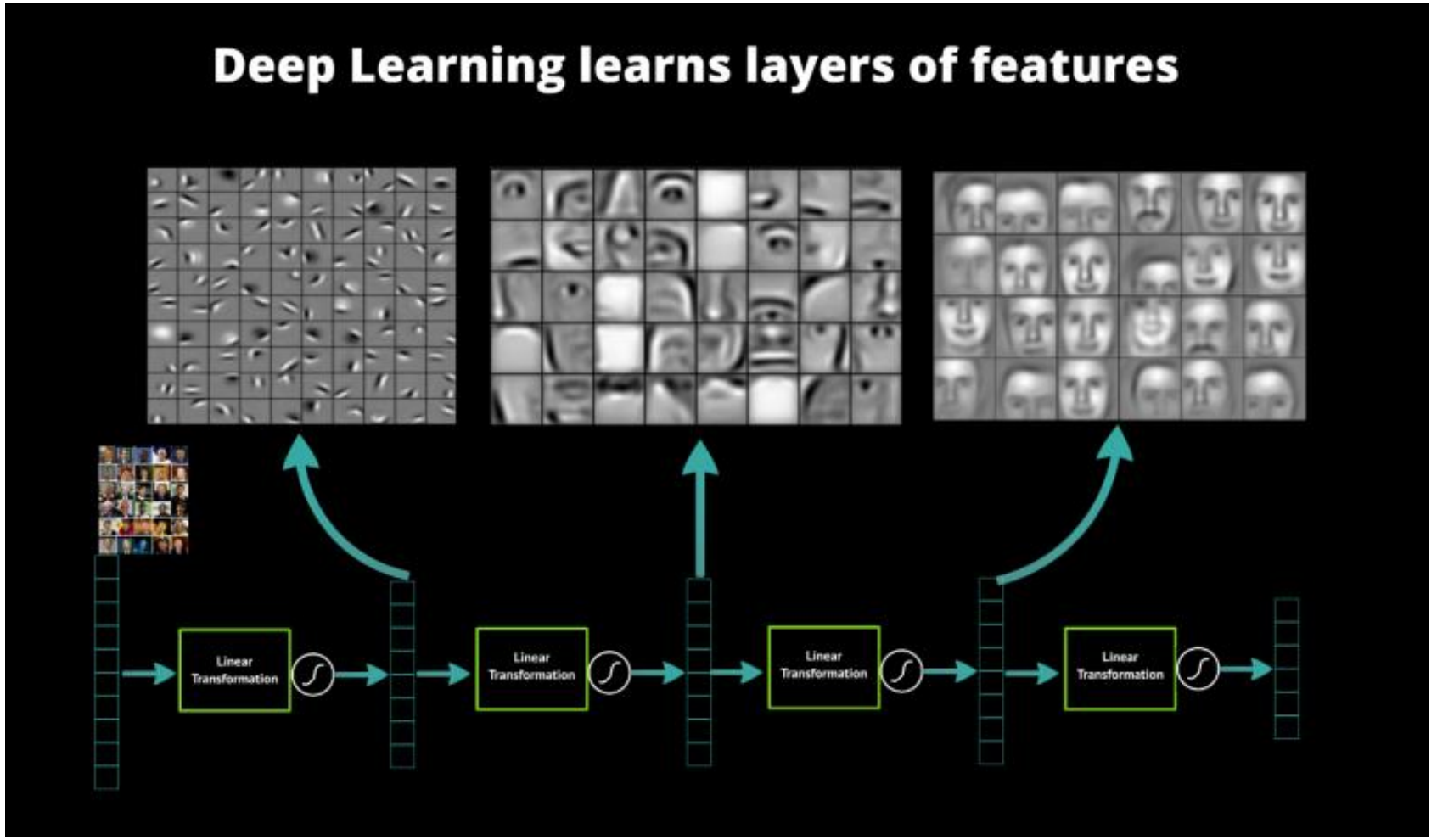
Глубинное обучение



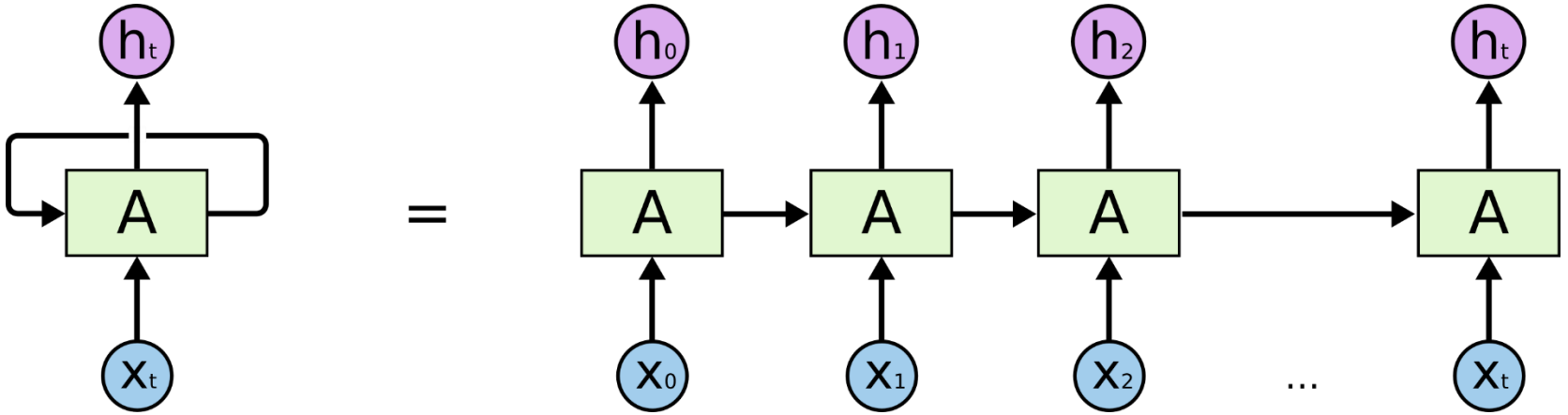
Сверточная нейронная сеть



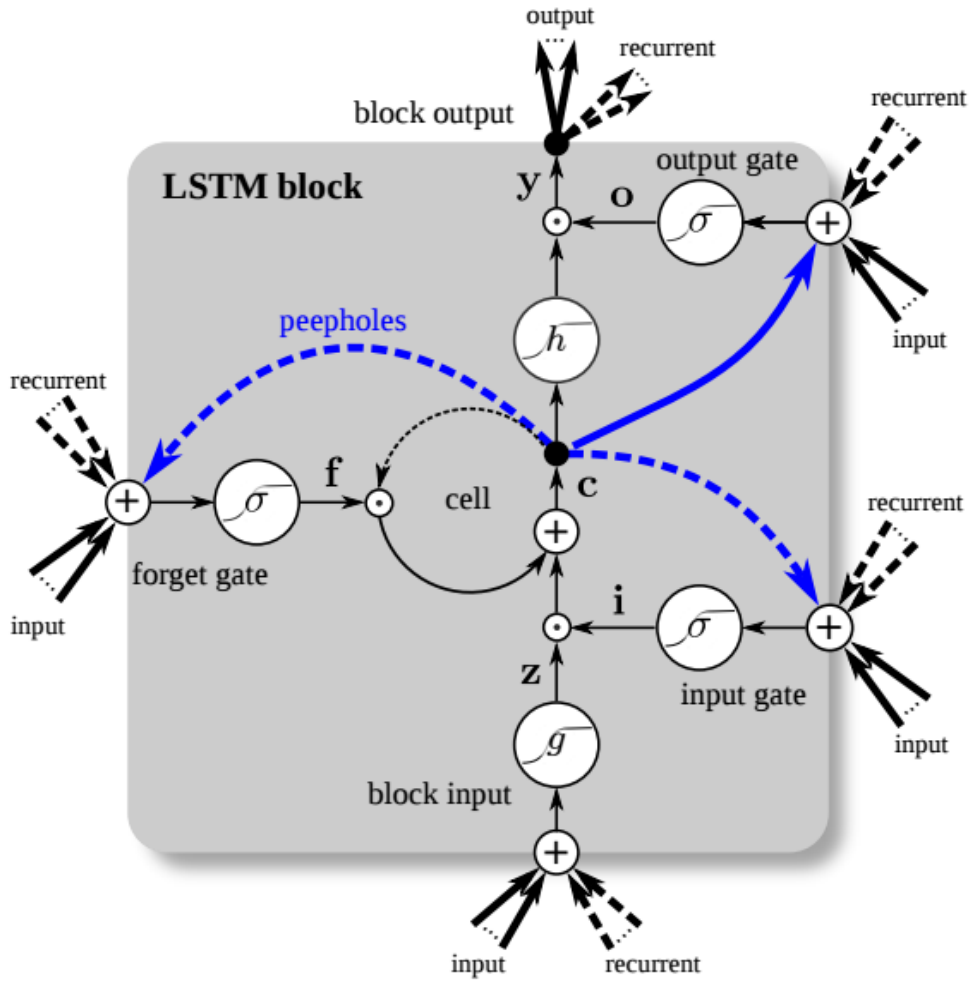
Обучение нейронной сети



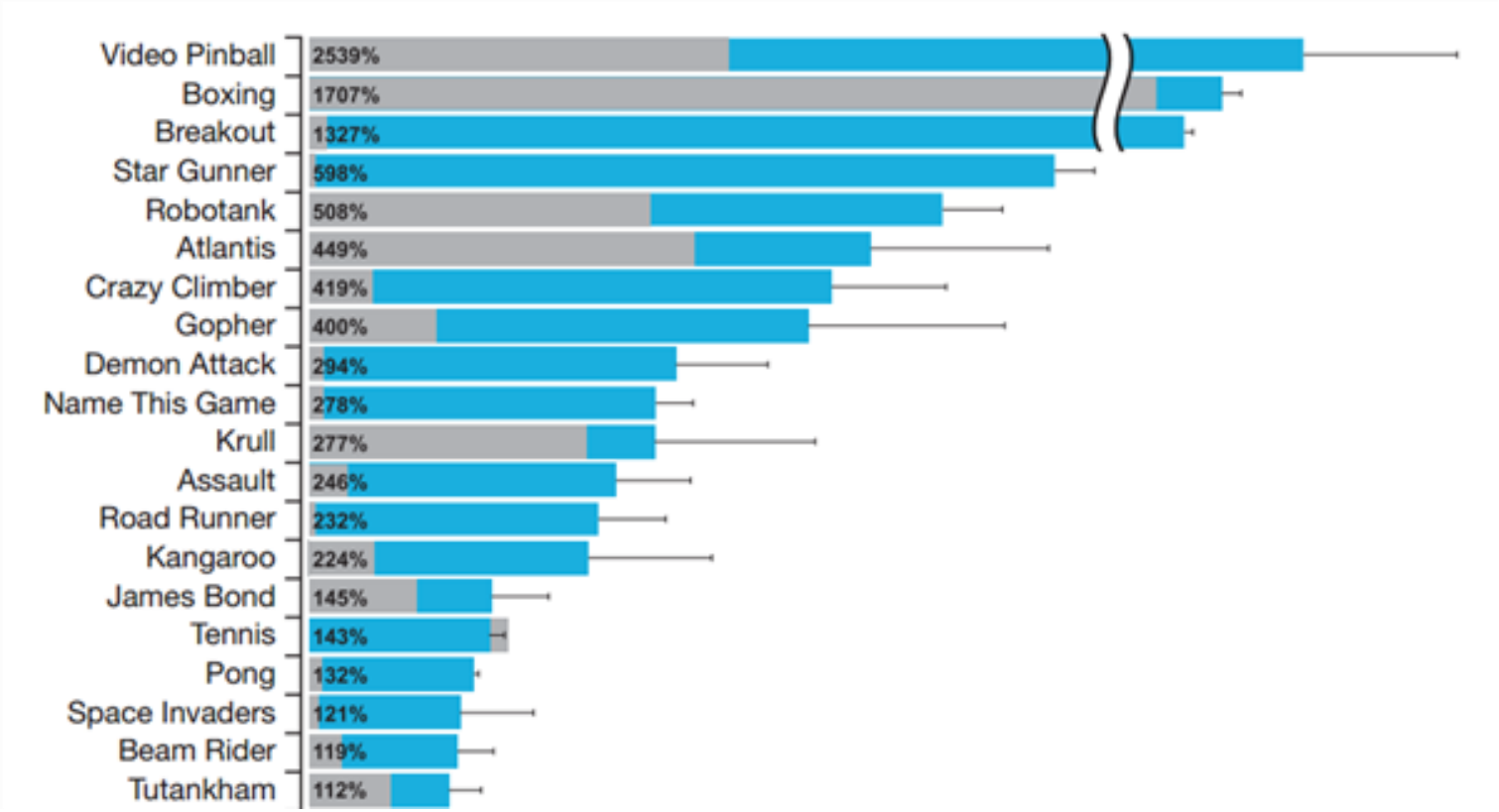
Рекуррентная нейронная сеть



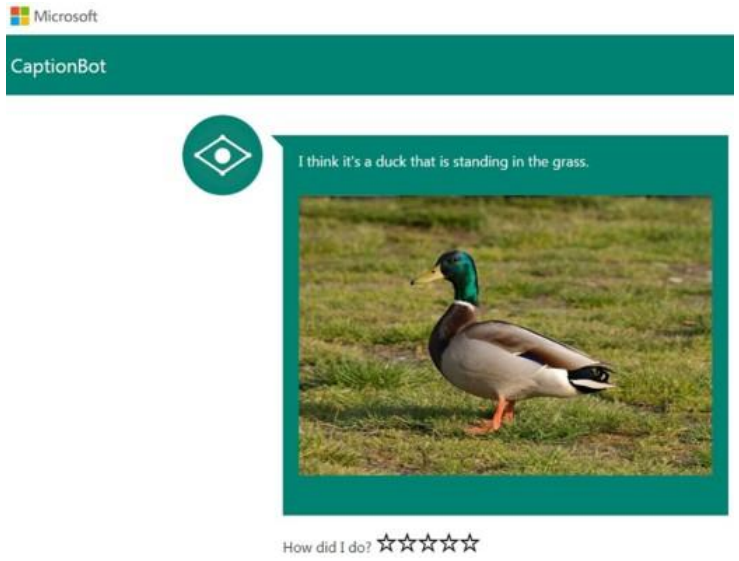
LSTM

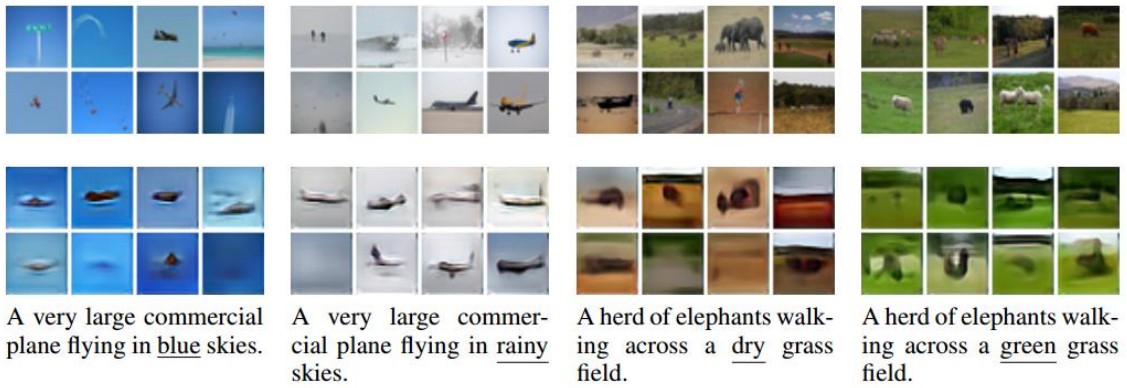

output
recurrent
block output
recurrent
output gate
LSTM block
y
o
peepholes
h
input
recurrent
recurrent
f
cell
c
forget gate
input
recurrent
i
input gate
z
input
g
block input
input
recurrent

Обучение с подкреплением



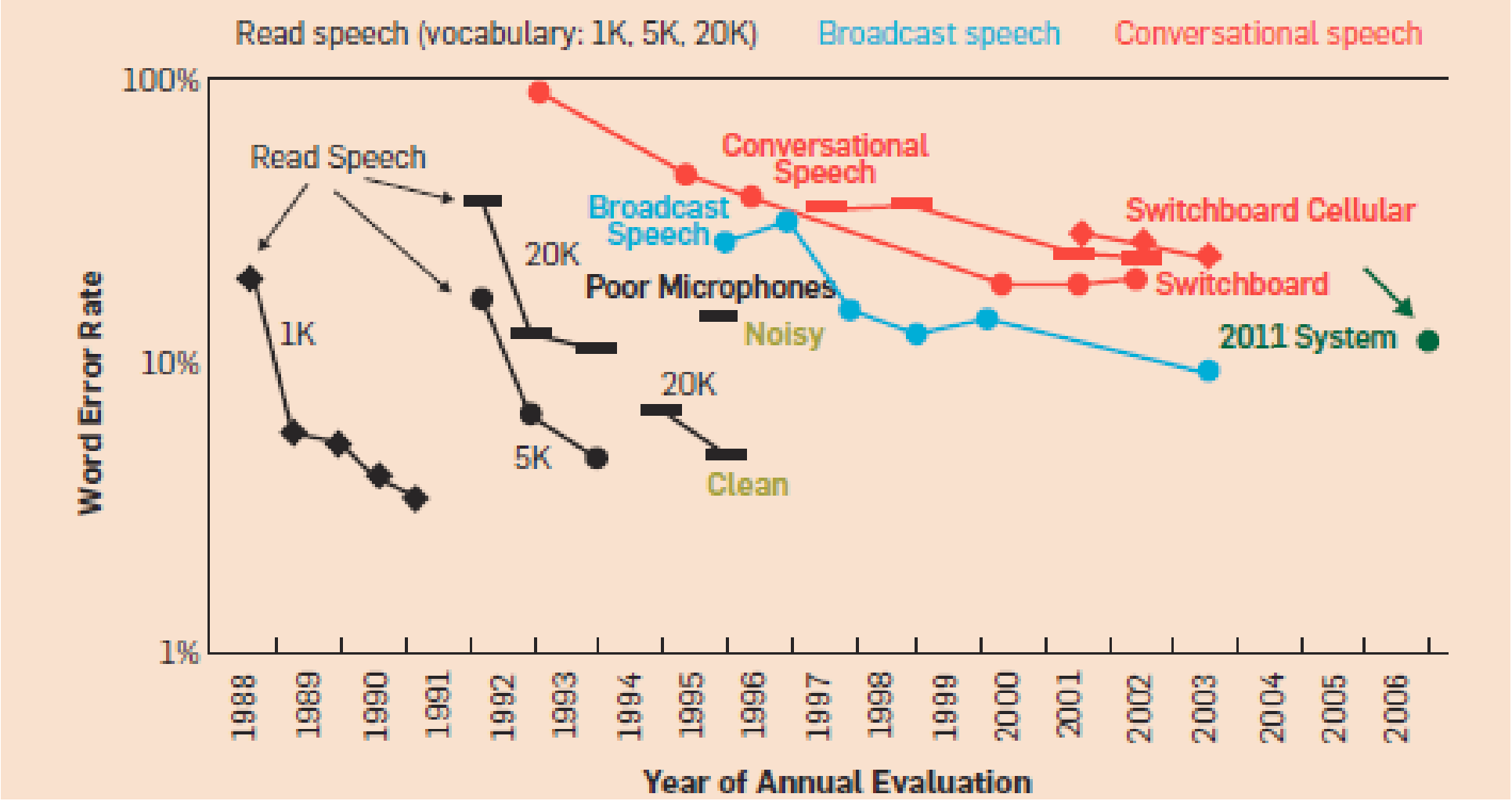
Комбинированная архитектура



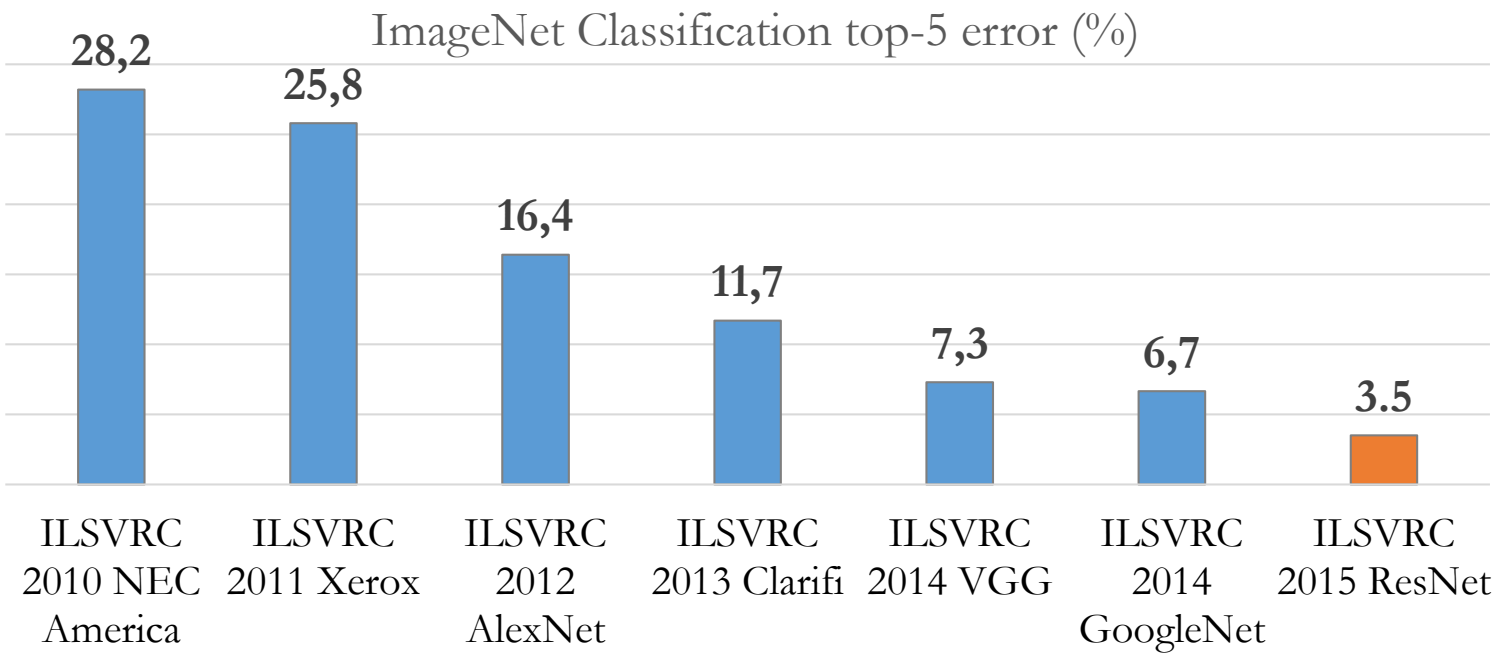


Что такое Microsoft Cognitive Toolkit

Хронология уменьшения относительной частоты ошибок распознавания речи при возрастании сложности задачи



ImageNet: Microsoft 2015 ResNet



CNTK → Cognitive Toolkit

- Кроссплатформенный набор инструментов от Microsoft с открытым исходным кодом, предназначенный для проектирования и обучения нейронных сетей глубинного обучения
- Позволяет создавать нейронные сети практически любой аритектуры, совмещая элементарные нейросетевые блоки
- Готов к интеграции в реальные проекты: качество работы соответствует уровню современного развития технологии, поддерживается работа на мульти-GPU/мультисервере

Кроссплатформенный набор инструментов от Microsoft с открытым исходным кодом, предназначенный для проектирования и обучения нейронных сетей глубинного обучения

Открытый исходный код

разработан исследователями Microsoft Speech (Dong Yu et al.) 4 года назад;
опубликован (CodePlex) в начале 2015

выложен на GitHub с января 2016 под пермиссивной лицензией

Используется в группах продуктов Microsoft

Поддержка Linux, Windows, .Net, docker

Позволяет создавать нейронные сети практически любой аритектуры, совмещая элементарные нейросетевые блоки

пример: ИНС прямого распространения с 2 скрытыми слоями

$$h_1 = \sigma(W_1 x + b_1)$$
$$h_2 = \sigma(W_2 h_1 + b_2)$$
$$P = \mathrm{softmax}(W_{\mathrm{out}} h_2 + b_{\mathrm{out}})$$

```
h1 = Sigmoid (W1   * x  + b1)
h2 = Sigmoid (W2   * h1 + b2)
P  = Softmax (Wout * h2 + bout)
```

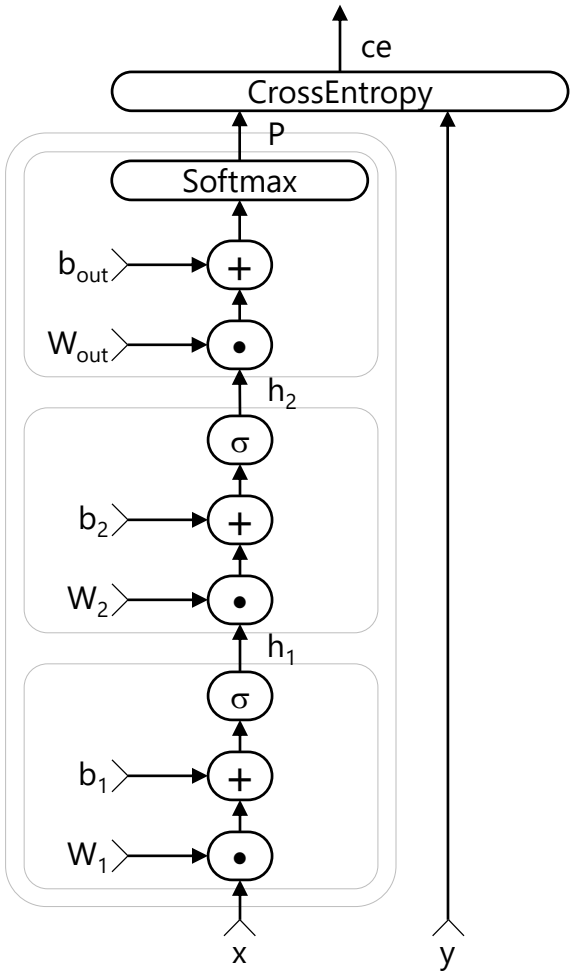
входной вектор $x \in \mathrm{R}^M$ и унитарный выходной вектор $y \in \mathrm{R}^J$

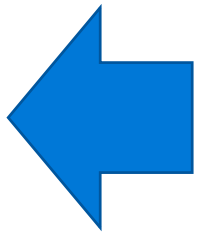
обучение по критерию кросс-энтропии

$$ce = y^{\mathrm{T}} \log P$$
$$\sum_{\mathrm{corpus}} ce = \max$$

```
ce = CrossEntropy (y, P)
```

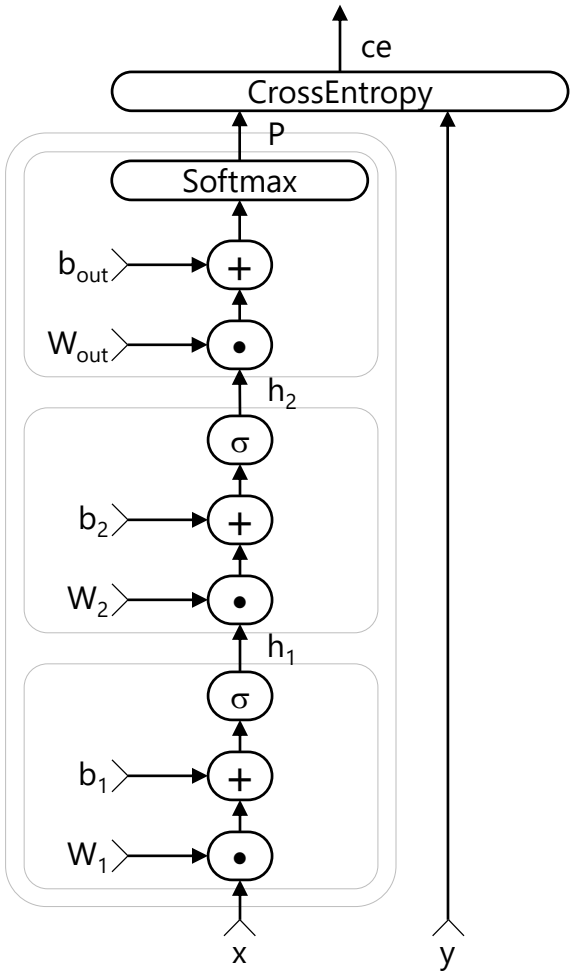
Позволяет создавать нейронные сети практически любой аритектуры, совмещая элементарные нейросетевые блоки





```
h1 = Sigmoid (W1   * x  + b1)
h2 = Sigmoid (W2   * h1 + b2)
P  = Softmax (Wout * h2 + bout)
ce = CrossEntropy (y, P)
```

Позволяет создавать нейронные сети практически любой аритектуры, совмещая элементарные нейросетевые блоки



Узлы: элементарные функции

Связи: веса

Автоматический поиск градиента

$\partial\mathcal{F} / \partial \text{in} = \partial\mathcal{F} / \partial \text{out} \cdot \partial \text{out} / \partial \text{in}$

Отложенное вычисление

Редактируемая конфигурация

Позволяет создавать нейронные сети практически любой аритектуры, совмещая элементарные нейросетевые блоки

Взаимозаменяемость блоков позволяет проектировать НС различного типа

Глубинные НС
Рекуррентные НС
Сверточные НС
LSTM
Обучение с подкреплением

Модели применимы для распознавания

Речи
Изображений
Текста

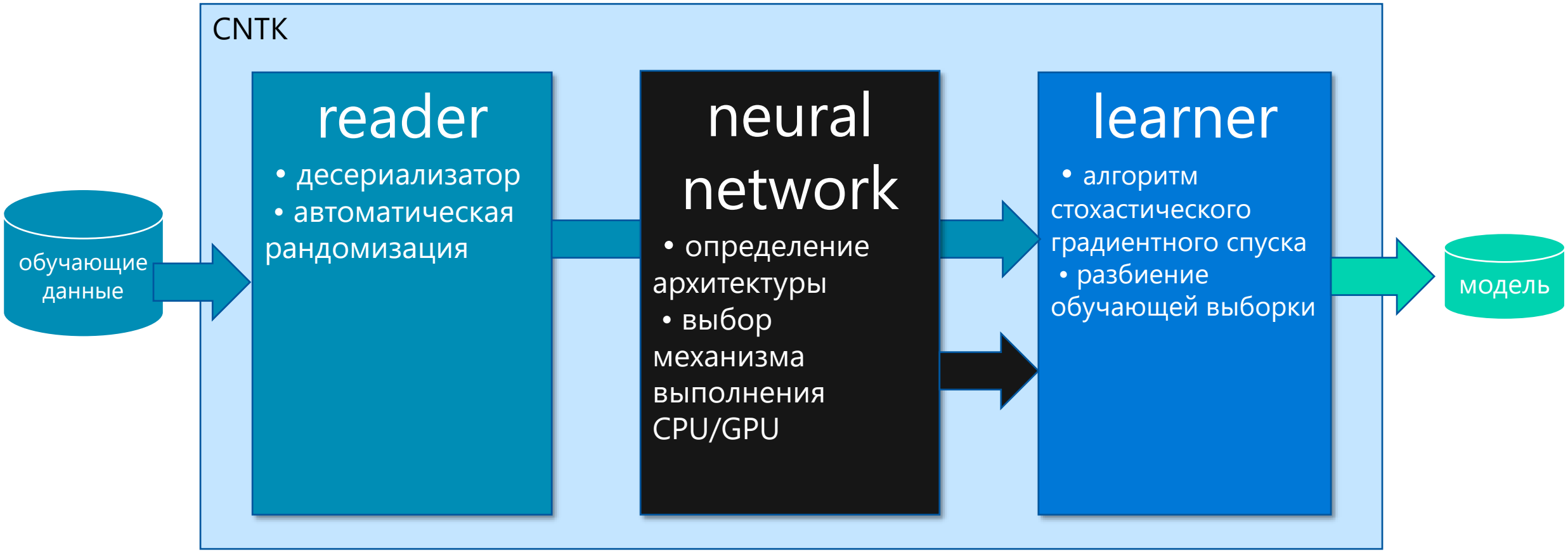
Готов к интеграции в реальные проекты

Качество работы соответствует уровню современного развития технологии как при решении тестовых задач, так и в реальных продуктах

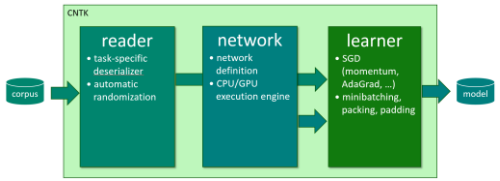
Оптимизация для работы с GPU

Поддержка параллельного обучения на мульти-GPU/мультисервере

Архитектура



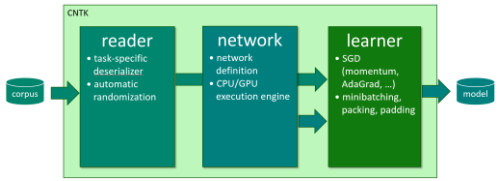
top-level configuration




```
cntk configFile=yourConfig.cntk command="train:eval" root="exp-1"
```

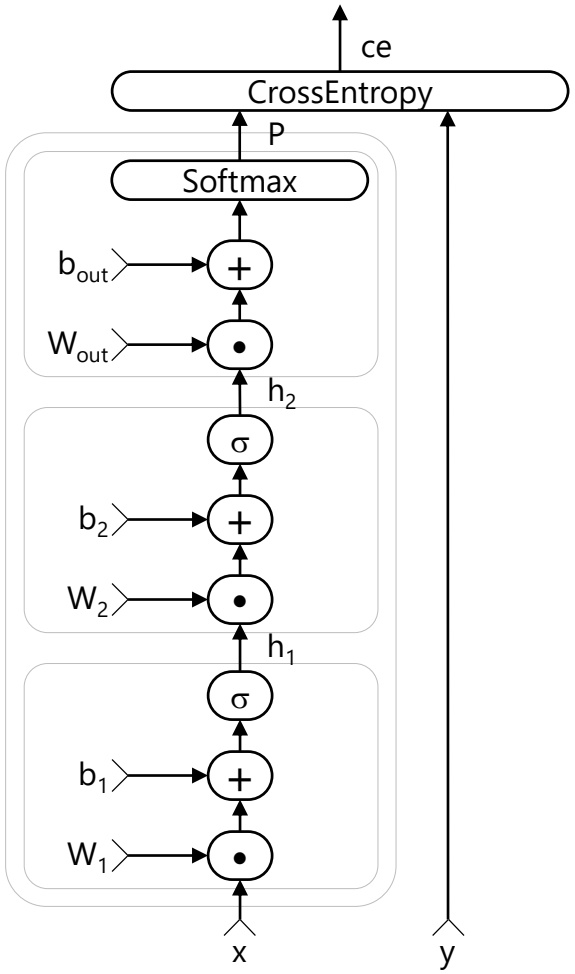
```
# content of yourConfig.cntk:
train = {
    action = "train"
    deviceId = "auto"
    modelPath = “$root$/models/model.dnn"

    reader = { … }
    BrainScriptNetworkBuilder = { … }
    SGD = { … }
}
eval = { … }
```

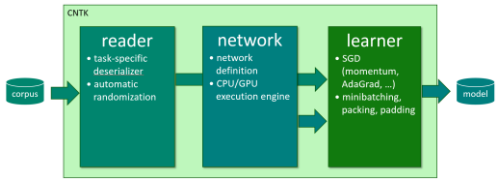
reader




```
reader = {
    readerType = "ImageReader"
    file = "$ConfigDir$/train_map.txt"
    randomize = "auto"
    features = { width=224; height=224; channels=3; cropRatio=0.875 }
    labels = { labelDim=1000 }
}
```

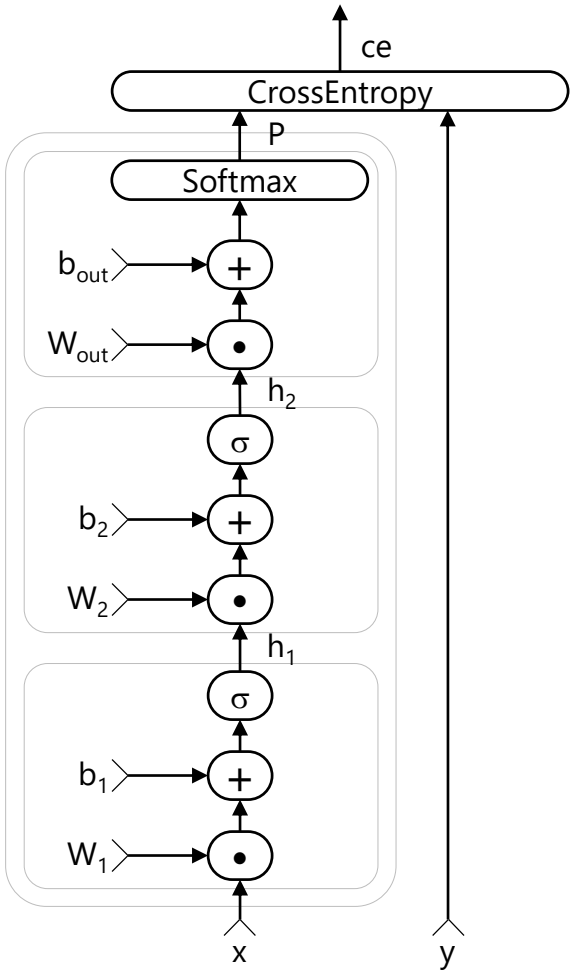
network

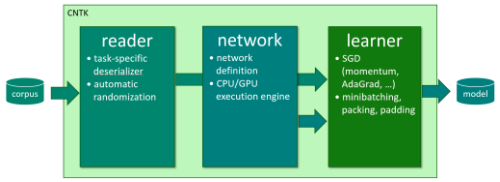





```
M = 40 ; N = 512 ; J = 9000 // feat/hid/out dim
x = Input{M} ; y = Input{J} // feat/labels
W1   = Parameter{N, M} ; b1   = Parameter{N}
W2   = Parameter{N, N} ; b2   = Parameter{N}
Wout = Parameter{J, N} ; bout = Parameter{J}

h1 = Sigmoid(W1   * x  + b1)
h2 = Sigmoid(W2   * h1 + b2)
P  = Softmax(Wout * h2 + bout)
ce = CrossEntropy(y, P)
```

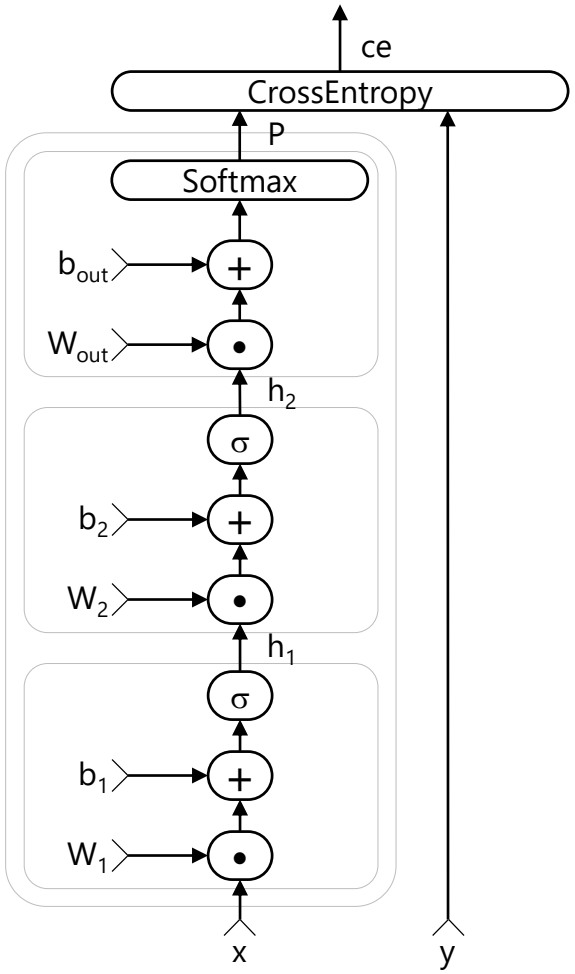
network

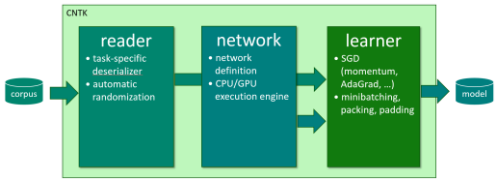





```
M = 40 ; N = 512 ; J = 9000 // feat/hid/out dim
x = Input{M} ; y = Input{J} // feat/labels
Layer (x, out, in, act) = {   // reusable block
    W = Parameter{out, in} ; b = Parameter{out}
    h = act(W * x + b)
}.h
h1 = Layer(x,  N, M, Sigmoid)
h2 = Layer(h1, N, N, Sigmoid)
P  = Layer(h2, J, N, Softmax)
ce = CrossEntropy(y, P)
```

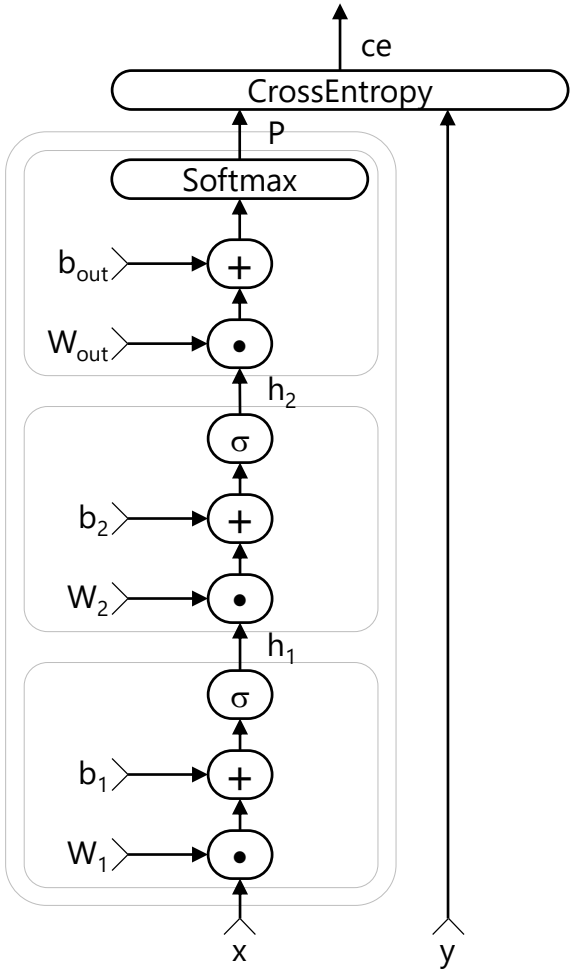
network

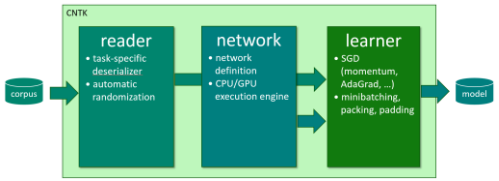





```
M = 40 ; N = 512 ; J = 9000 // feat/hid/out dim
x = Input{M} ; y = Input{J} // feat/labels
DenseLayer {Nout, activation=Identity} = {
    W = Parameter(Nout, 0) ; b = Parameter(Nout)
    apply(x) = activation(W * x + b)
}.apply
h1 = DenseLayer{N, activation=Sigmoid}(x)
h2 = DenseLayer{N, activation=Sigmoid}(h1)
P  = DenseLayer{J, activation=Softmax}(h2)
ce = CrossEntropy(y, P)
```

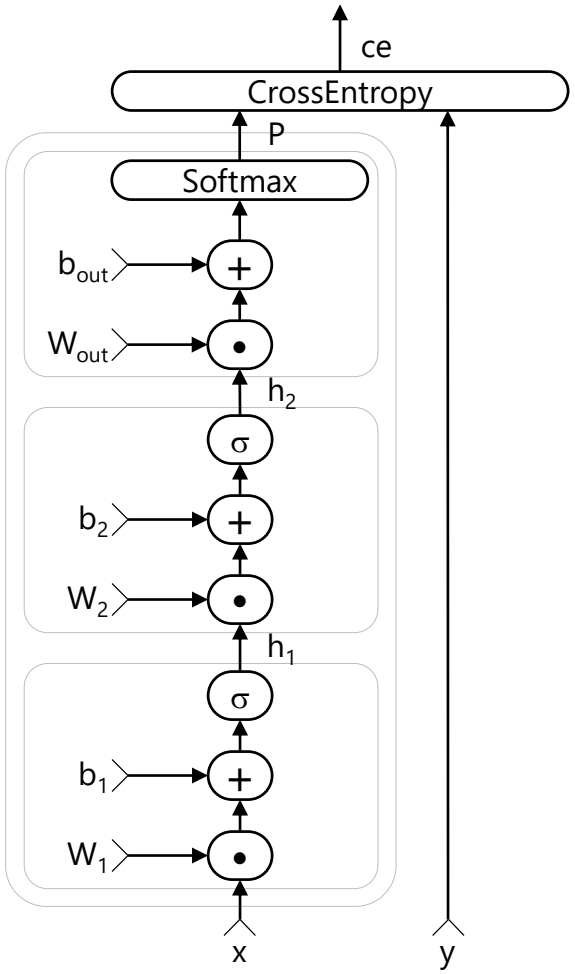
network

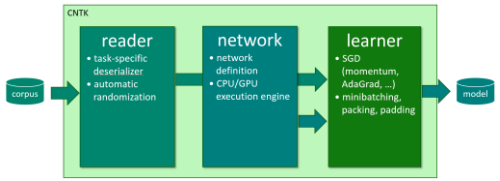





```
M = 40 ; N = 512 ; J = 9000 // feat/hid/out dim
x = Input{M} ; y = Input{J} // feat/labels
DenseLayer {out, activation=Identity} = { … }
Sequential (fnArray) = { … }
model = Sequential (
    DenseLayer{N, activation=Sigmoid} :
    DenseLayer{N, activation=Sigmoid} :
    DenseLayer{J, activation=Softmax}
)
P = model (x)
ce = CrossEntropy(y, P)
```

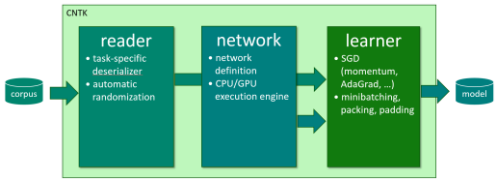
network





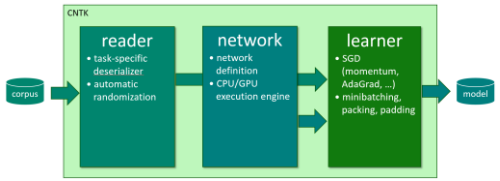

```
M = 40 ; N = 512 ; J = 9000 // feat/hid/out dim
x = Input{M} ; y = Input{J} // feat/labels
DenseLayer {out, activation=Identity} = { … }
Sequential (fnArray) = [ … ]
model = Sequential (
    DenseLayer{N} : Sigmoid :
    DenseLayer{N} : Sigmoid :
    DenseLayer{J} : Softmax
)
P = model (x)
ce = CrossEntropy(y, P)
```

learner




```
SGD = {
    maxEpochs = 50
    minibatchSize = $mbSizes$
    learningRatesPerSample = 0.007*2:0.0035
    momentumAsTimeConstant = 1100
    AutoAdjust = { … }
    ParallelTrain = { … }
}
```

Организация процесса



Конфигурация (reader, network, learner)

Обучение и оценка

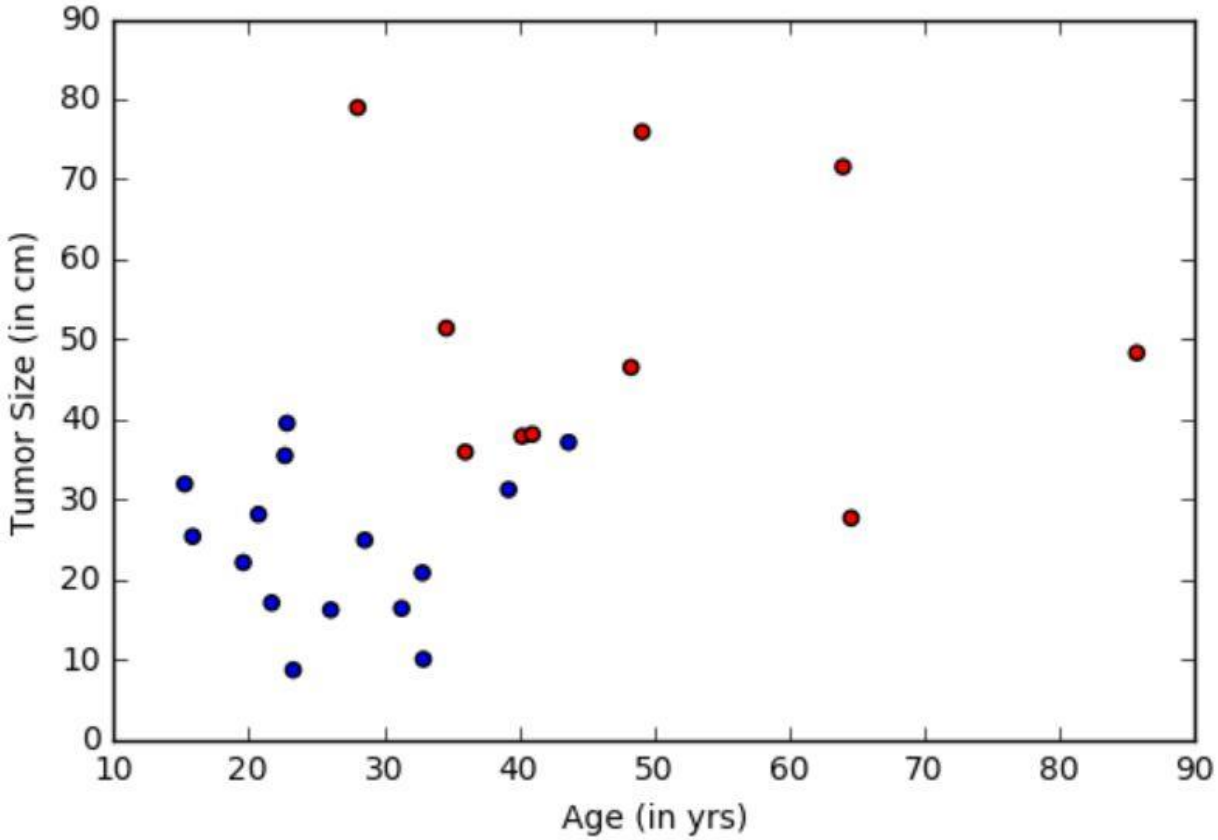
Модификация модели

Использование готовой модели в проекте с EvalDll.dll/.so (C++) или EvalWrapper.dll (.Net)

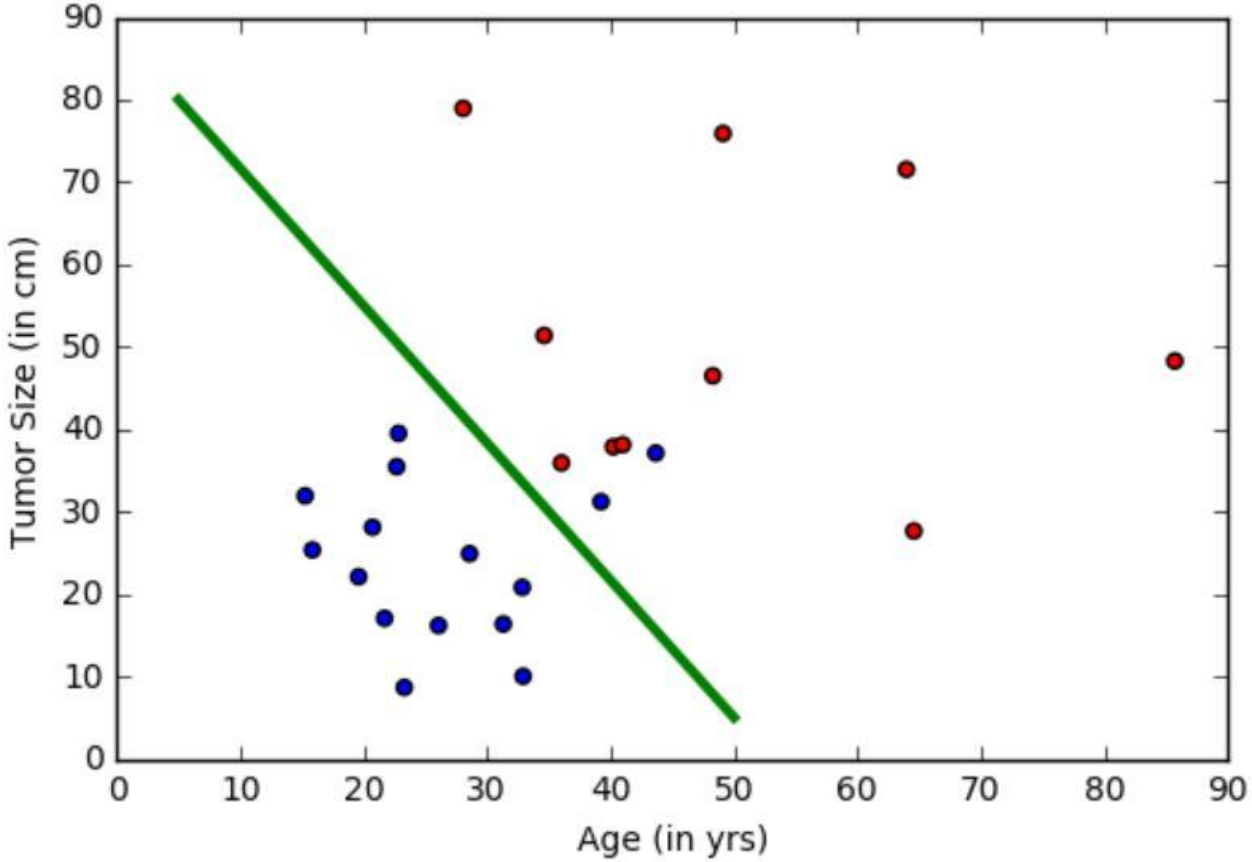
Демонстрация

Логистическая регрессия

Проблема


Tumor Size (in cm)
Age (in yrs)
0
10
20
30
40
50
60
70
80
90
10
20
30
40
50
60
70
80
90

Цель



Порядок действий

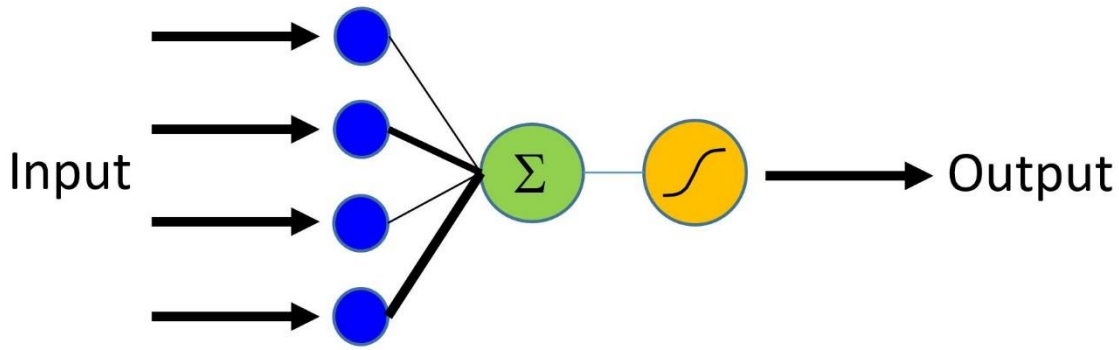
Сбор данных

Предобработка данных

Создание модели

Обучение модели

Оценка качества



Полезные ссылки для дальнейшего изучения

Web site https://www.microsoft.com/en-us/research/product/cognitive-toolkit/

Github https://github.com/microsoft/cntk

Documentation https://github.com/microsoft/cntk/wiki

Data Science Virtual Machine https://azure.microsoft.com/en-us/marketplace/partners/microsoft-ads/standard-data-science-vm/

Microsoft Azure

SALES 8-800-100-6347 | MY ACCOUNT | PORTAL | Search

Why Azure | Solutions | Products | Documentation | Pricing | Partners | Blog | Resources | Support

FREE ACCOUNT

Deep Learning toolkit for the DSVM

by Microsoft

Deploy

The data science virtual machine (DSVM) on Azure, based on Windows Server 2012, contains popular tools for data science modeling and development activities such as Microsoft R Server Developer Edition, Anaconda Python, Jupyter notebooks for Python and R, Visual Studio Community Edition with Python and R Tools, Power BI desktop, SQL Server Developer edition, and many other data science and ML tools. Use the DSVM to jump-start modeling and development for your data science project.

This deep learning toolkit provides GPU versions of mxnet and CNTK for use on Azure GPU N-series instances. These GPUs use discrete device assignment, resulting in performance that is close to bare-metal, and are well-suited to deep learning problems that require large training sets and expensive computational training efforts. The deep learning toolkit also provides a set of sample deep learning solutions that use the GPU, including image recognition on the CIFAR-10 database and a character recognition sample on the MNIST database. GPU instances are currently only available in the South Central US.

By continuing to create and use this toolkit you are accepting the following license agreements.

Deploying this toolkit requires access to Azure GPU NC-class instances.

After provisioning the deep learning toolkit, see the README file in C:\dsvm\deep-learning, or on the desktop, for more information.

VERSION: 1.0.3

ИИ становится продуктивным

Повсеместно внедряется в программные продукты

Microsoft снижает планку

Используя технологии Microsoft, вы сможете применять у себя ИИ-технологии без глубокого погружения в теорию

Круглый стол

Приходите вечером на круглый стол

Q&A

Microsoft Cognitive Toolkit – инструментарий для проектирования и обучения нейронных сетей

Елизавета Лаврова
t-ellavr@microsoft.com

#msdevcon

DevCon School
Технологии будущего

Отзывы

Помогите нам стать лучше!

На вашу почту отправлена индивидуальная ссылка на электронную анкету. 2 ноября в 23:30 незаполненная анкета превратится в тыкву.

Заполните анкету и подходите к стойке регистрации за приятным сюрпризом!

#msdevcon

Оставляйте отзывы в социальных сетях. Мы все читаем. Спасибо вам! ☺

DevCon School
Технологии будущего