Deep Learning: a mesterséges intelligencia hajtóereje

Gyires-Tóth Bálint

BME TMIT
SmartLab
Intelligent Interactions
Nemzeti Kiválóság Program

Bemutatkozás

Bálint Gyires-Tóth, Ph.D.

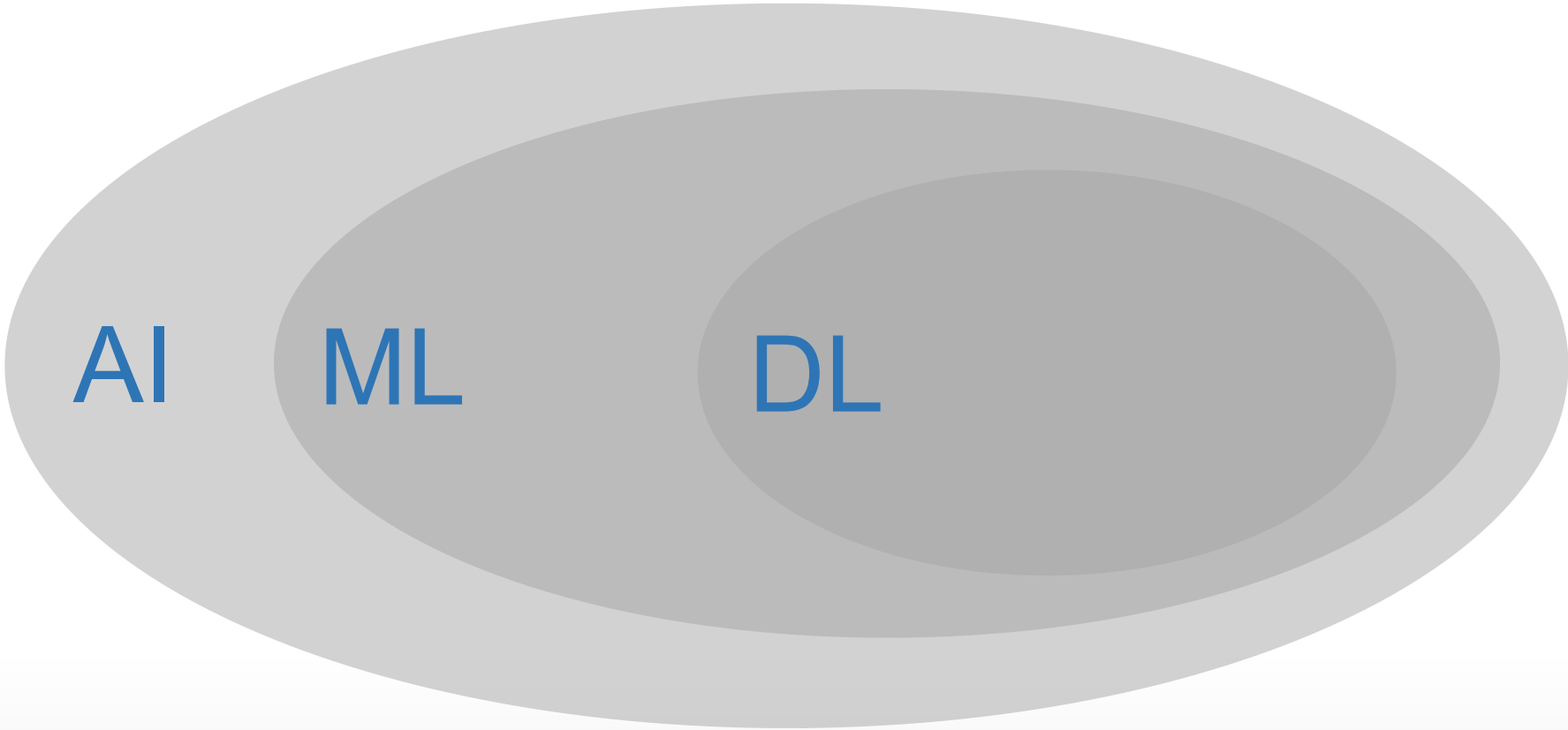
SmartLab @ BME TMIT

- Jelfeldolgozás
- Idősor modellezés
- Gépi tanulás és mély tanulás 2007 óta

NVidia Deep Learning Institute

Certified Instructor & University Ambassador

AI, Machine Learning, Deep Learning



De mi is az a deep learning?

Gépi tanuló algoritmusok struktúrált összesége, ahol több rétegen keresztül próbáljuk az adatok különböző szintű absztrakcióit megtanulni és modellezni.

Gyakorlatban a deep learning kifejezést elsősorban a mély neuronhálókkal kapcsolatban használják.

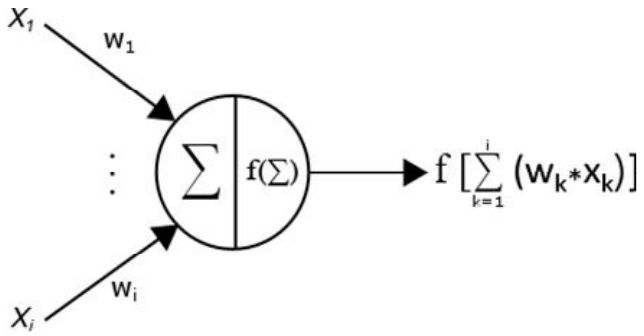
Miben új?

Új algoritmusok + nagymennyiségű adat + GPU

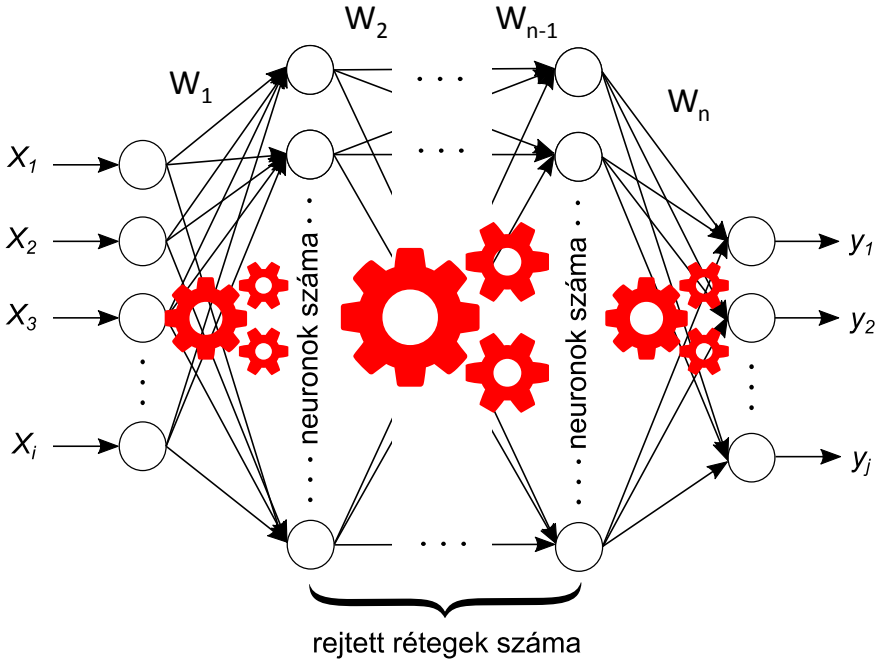
+ egyre alacsonyabb „belépési küszöb”

+ erősen open-source orientált nyílt kutatói közösség („democratizing AI”)

Előrecsatolt, teljes összeköttetésű mély neuronháló



f(∑): nemlináris függvény
pl. f(∑)=max(0, ∑)



backpropagation

TANULÁS: súlyok „hangolása”

A **deep learning** alapú AI a következő nagy lehetőség a növekedésre!

Számos területen már bizonyított…

~ Hallás

~ Beszéd

~ Látás

~ Döntés

Skálázhatóság

Hardver architektúra – „consumer grade”

1-8 GPU workstation (min. 64 GB RAM, 2-4 TB HDD)

NVIDIA TITAN X, 1080Ti, TITAN Xp, Titan V, 2080Ti

- Titan V: 12 teraFLOP (FP32)
- Tensor Core: 110 Tflops (FP16, deep learning)
- Max. 12 GB GPU RAM

Cloud Services

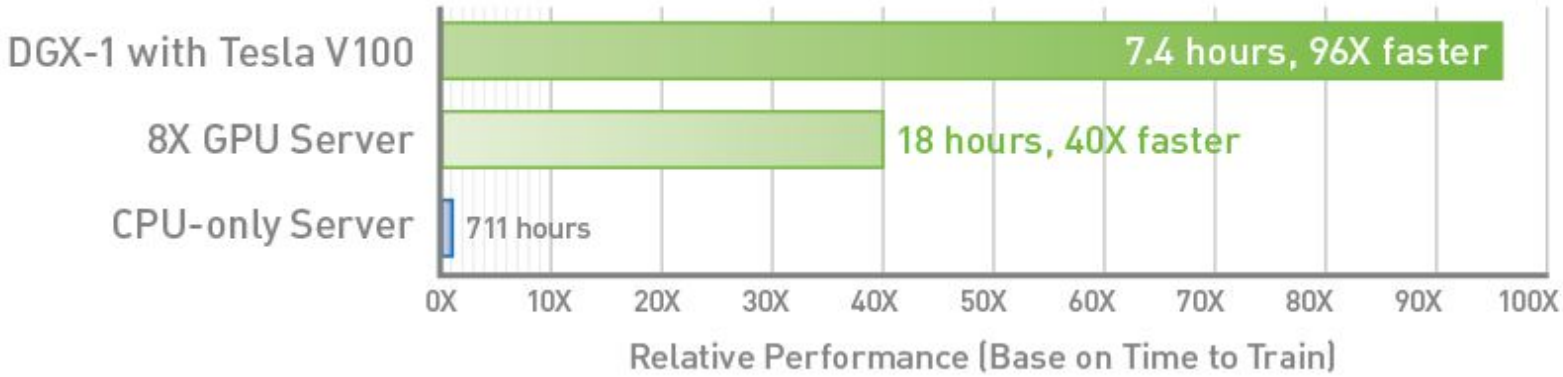
Hardver architektúra – „Server grade”

DGX Station (28 FP32, 56 FP16, 480 Tflop DL), $69.000

DGX-1 V100 (56 FP32, 112 FP16, 960 Tflop DL), $149.000

DGX-2 V100 (112 FP32, 224 FP16, 2 Pflops DL), $399.000

NVIDIA DGX-1 Delivers 96X Faster Training



Workload: ResNet50, 90 epochs to solution | CPU Server: Dual Xeon E5-2699 v4, 2.6GHz


DGX SATURNV
NVIDIA


summit
IBM

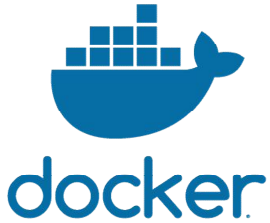
Szoftver architektúra – DL keretrendszerek

Python alapú fejlesztés

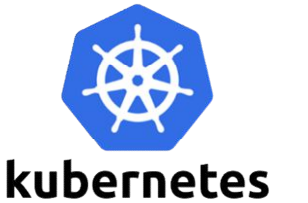
- TensorFlow (Google)
- Keras (Google)
- PyTorch (Facebook AI Research, Uber „Pyro” probabilistic programming)
- Microsoft CNTK, Chainer, Theano, Matlab DL Toolbox, etc.
- Caffe, Kaldi

Szoftver architektúra - Skálázhatóság

Docker: Container alapú fejlesztés az egyszerűbb fejlesztés, tesztelés, éles futtatás és skálázhatóság céljából.



Kubernetes: a container alapú rendszerek automatizált menedzsmentje.



Apache Spark 2.3 + Kubernetes 1.7+: az elosztott adatfeldolgozás és modellezés menedzsmentje. Az MLLib könyvtárban van deep learning modul.

Rapids AI

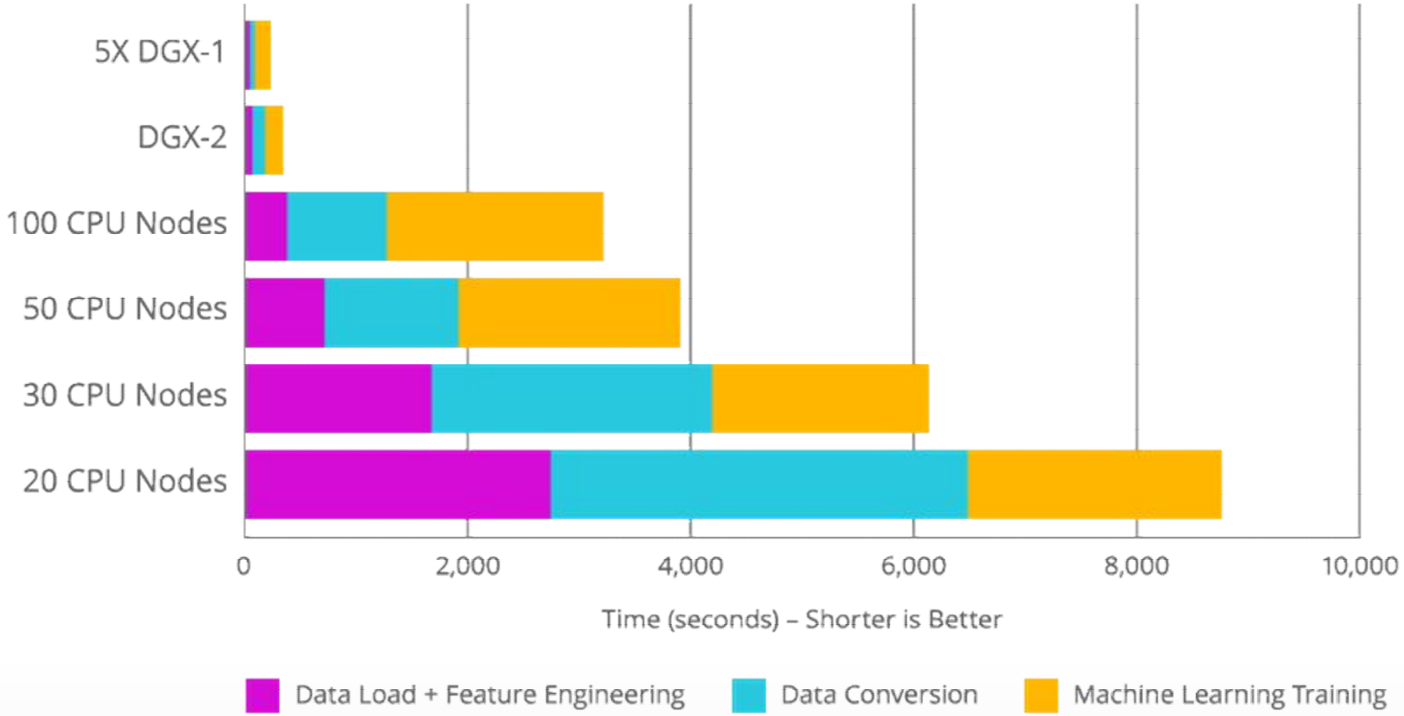
Pandas → Dask → Rapids AI

100 millió adatpont lineáris transzformációja

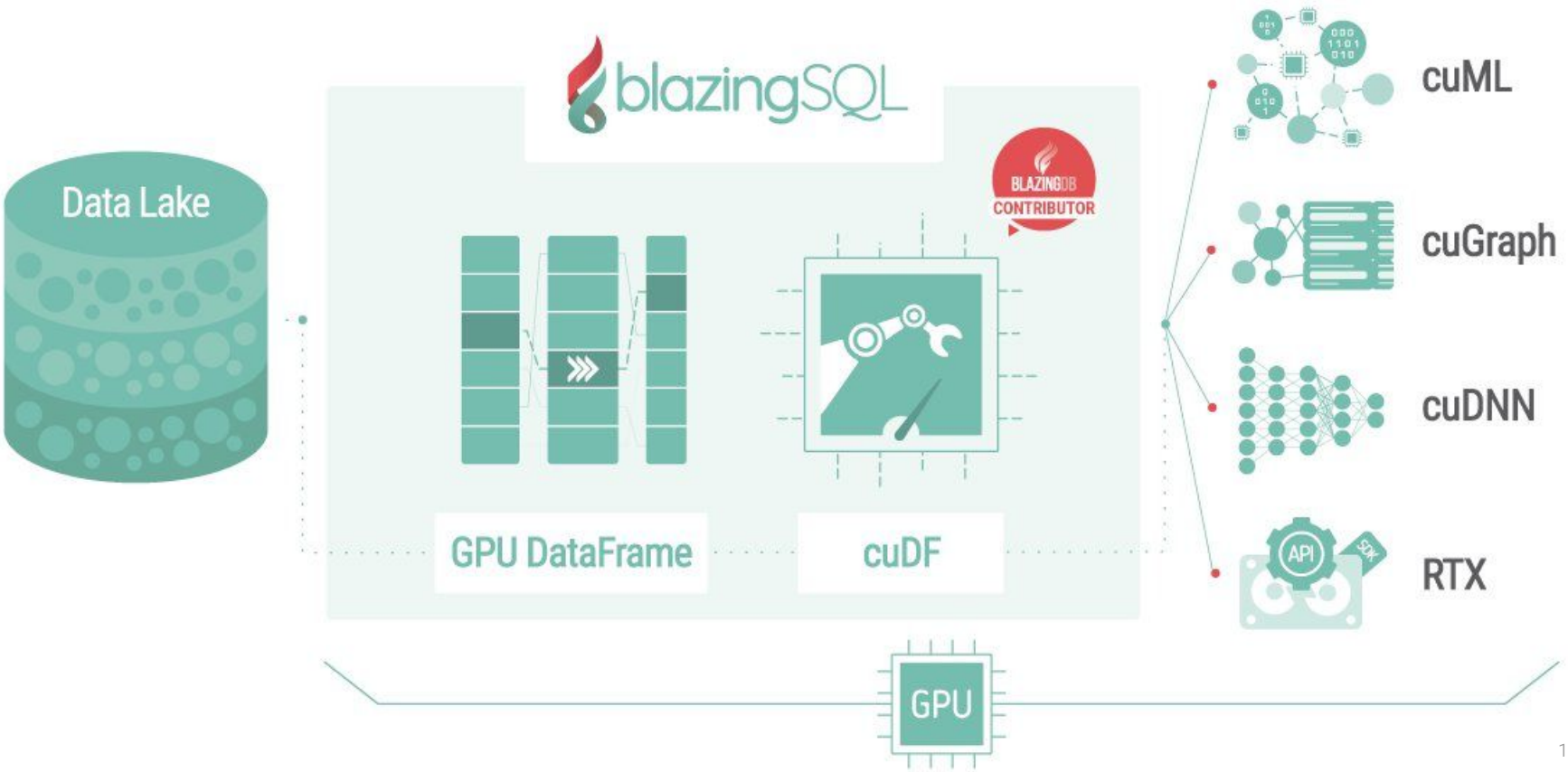
Techonológia	Sebesség
Pandas (CPU)	3957.1 sec
Dask (CPU paralell)	788.55 sec
Rapids (GPU)	2.103 sec

CPU: Dell T7910 Xeon E5-2640V4x2 | GPU: NVIDIA Titan Xp

Rapids AI


5X DGX-1
DGX-2
100 CPU Nodes
50 CPU Nodes
30 CPU Nodes
20 CPU Nodes
0
2,000
4,000
6,000
8,000
10,000
Time (seconds) – Shorter is Better
Data Load + Feature Engineering
Data Conversion
Machine Learning Training

Rapids AI + Blazing SQL



Alkalmazások

Alkalmazási területek

- Szekvenciális adatok, idősorok, többskálás modellezés
- Beszédfelismerés, beszédszintézis
- Természetes nyelvfeldolgozás (NLP), gépi fordítás
- Képfelismerés, objektum felismerés, objektum szegmentáció
- Ajánló rendszerek
- Vállalati adatok
- Neurális művészet
- Stb.

Alkalmazás 1

Szekvenciális adatok, idősor klasszifikáció.

CÉL:

Okotstelefon alapú gesztus felismerés

BEMENET:

200 x 7 adatpont (1-2 seconds)

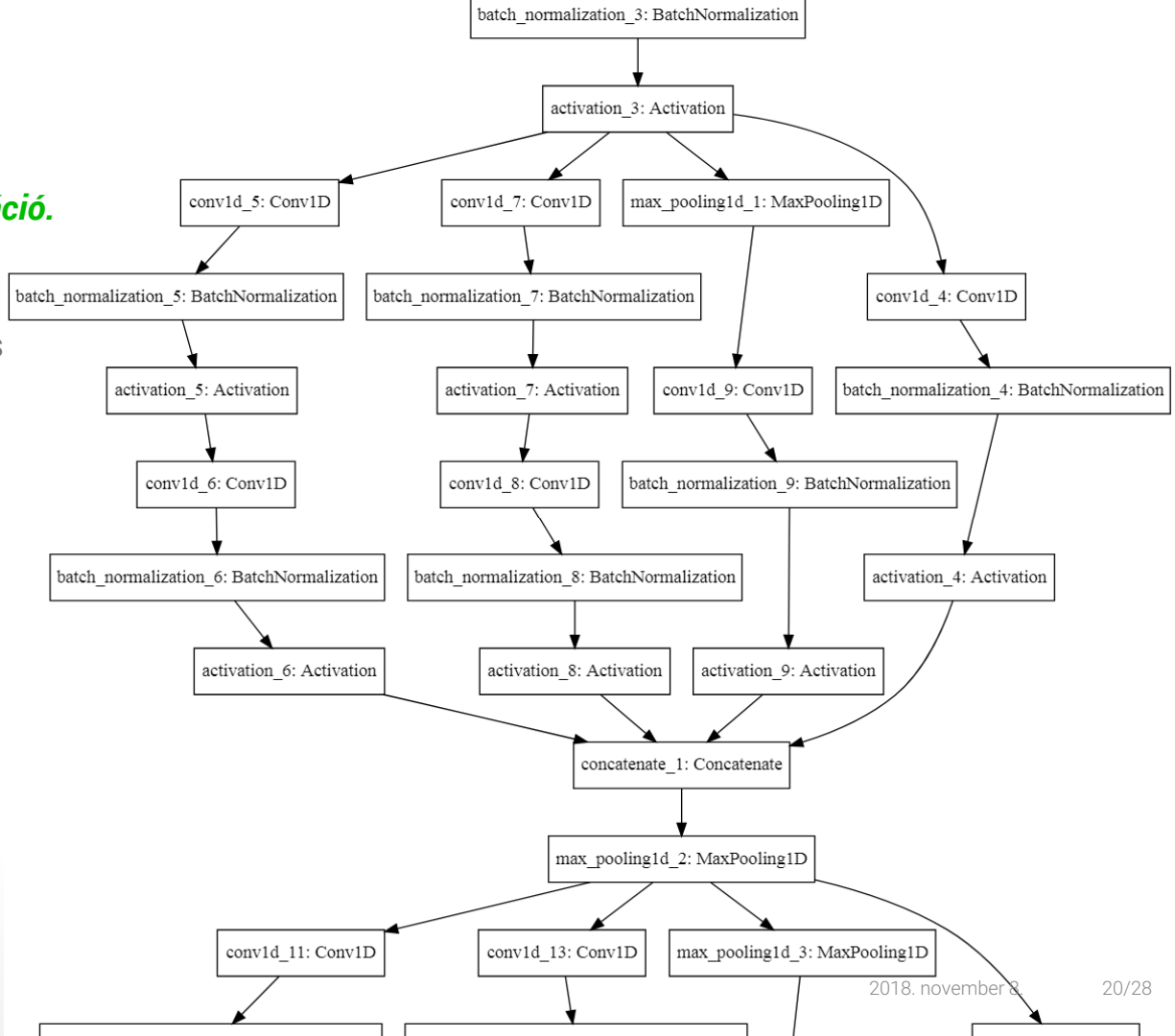
KIMENET:

Bináris osztályozás

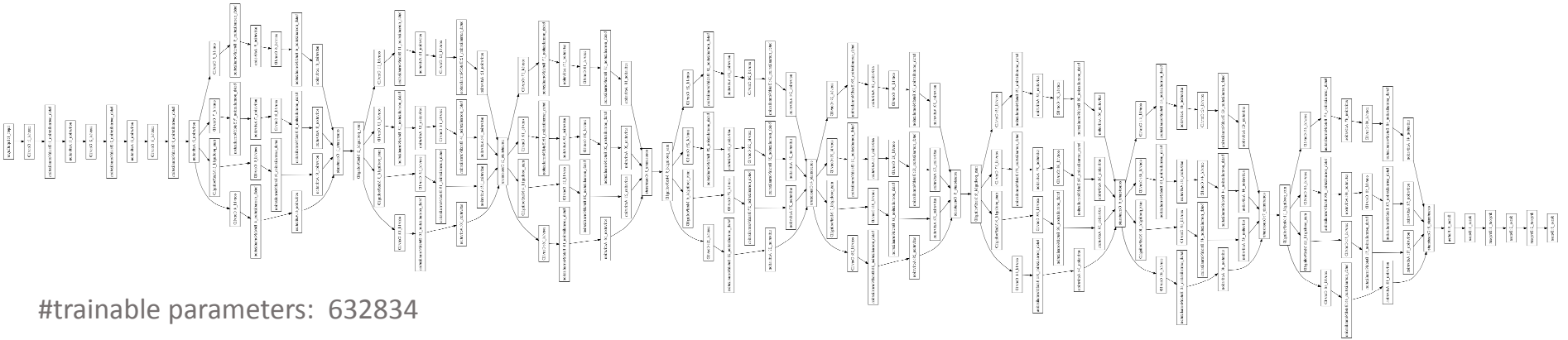
>99% pontosság.

~1000 óra adat 100 embertől

Kb. 26000 GPU óra (~3 GPU év) 12 Tflops-os GPU-kon.



Alkalmazás 1



#trainable parameters: 632834

#trainable parameters: 286498

Alkalmazás 1 - számok

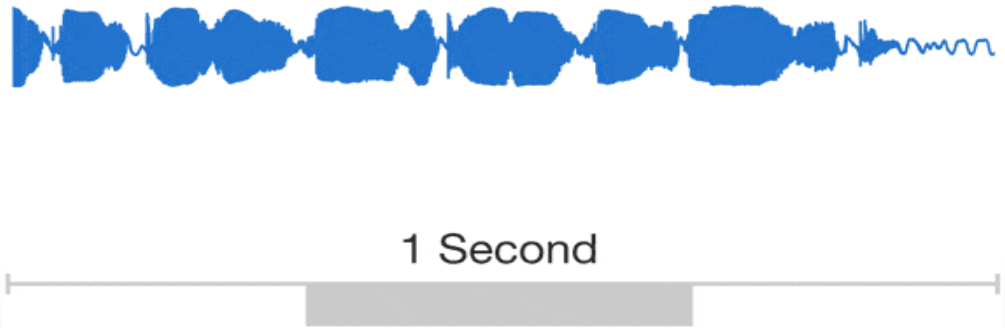
Fizetett adatgyűjtés.

2017 január:	~60 óra adat 12 embertől
2017 június:	~420 óra adat 50 embertől
2017 december:	~1000 óra adat 100 embertől

Kb. 26000 GPU óra (~3 GPU év) 12 Tflops-os GPU-kon.

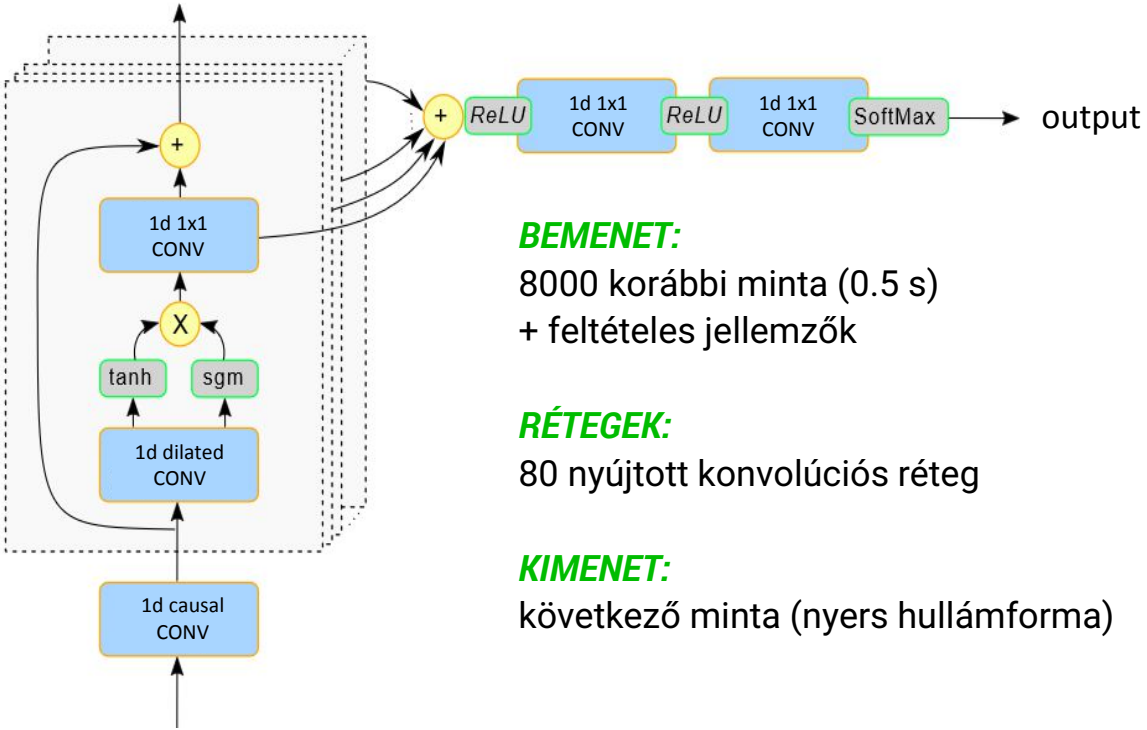
Alkalmazás 2 - WaveNet

- Google DeepMind, 2016 szeptember (később: Deep Voice, Baidu)
- Új típusú Text-to-Speech eljárás
- Szekvencia/idősor generálás több skálán



Source: https://deepmind.com/blog/wavenet-generative-model-raw-audio/

Alkalmazás 2 - WaveNet



(DeepMind WaveNet alapján)

BEMENET:

8000 korábbi minta (0.5 s)
+ feltételes jellemzők

RÉTEGEK:

80 nyújtott konvolúciós réteg

KIMENET:

következő minta (nyers hullámforma)

#trainable parameters: 1.2 million

Alkalmazás 2 - WaveNet hangminták

Google DeepMind – angol

SmartLab– magyar

Alkalmazási területek

Google Duplex, call center-ek, vasútállomás, képernyő felolvasók, okostelefonok, automotive, stb.

Alkalmazás 2 - WaveNet számokban

Tanítás

25 óra stúdiófelvétel

4 hét 12 Tflops GPU-n

Predikció

1 minta generálása 7.5 ms

1 másodperc generálása 12 másodperc

Összefoglalás

Deep Learning jövőbemutató eredményei

Feketedoboz működéstől való félelem

Skálázhatóság

Alkalmazott kutatás, ipari alkalmazások

Köszönöm a figyelmet!

Gyires-Tóth Bálint
toth.b@tmit.bme.hu

Jelen kutatás az Emberi Erőforrások Minisztériuma által az Új Nemzeti Kiválóság Program keretében meghírdetett Bolyai+ Felsőoktatási Fiatal Oktatói, Kutatói Ösztöndíj támogatásával készült (ÚNKP-18-4-BME-394).

Nemzeti Kiválóság Program

BME TMIT

SmartLab
Intelligent Interactions