Deep learning in MATLAB Workshop CloudUT

15.03.2021

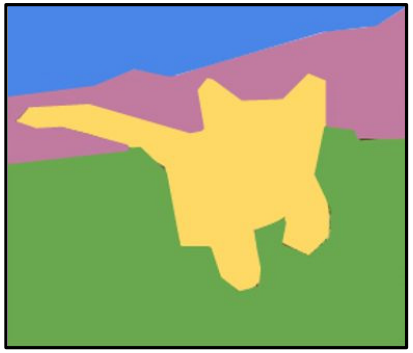
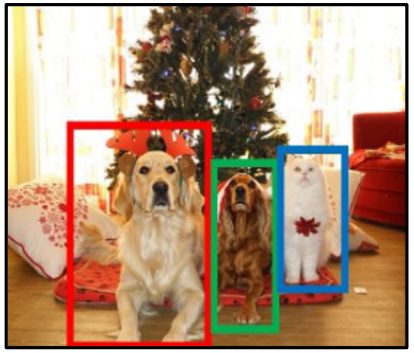
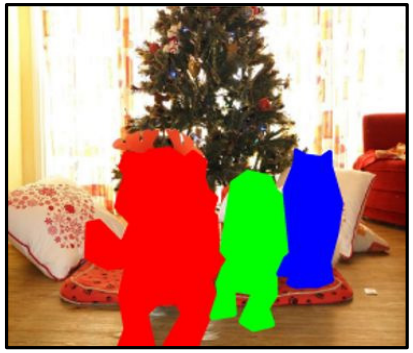
Ion Giosan, Cristian-Cosmin Vancea
Departamentul Calculatoare
Universitatea Tehnică din Cluj-Napoca
Ion.Giosan@cs.utcluj.ro, Cristian.Vancea@cs.utcluj.ro

Cuprins

- Context
- Obiective
- Metodologia de lucru
- Dezvoltare în MATLAB și CloudUT
- Aplicații demonstrative
- Concluzii

Context

Aplicații din domeniul Viziunii Artificiale

Clasificare	Segmentare semantică	Detecție de obiecte	Detecție de instanțe
			
Cat	Grass, Cat, Tree, Sky	Dog, Dog, Cat	Dog, Dog, Cat
Fără informații spațiale	Fără obiecte, doar pixeli	Obiecte multiple	

Sursa: http://cs231n.stanford.edu/slides/2020/lecture_12.pdf

Soluții: Abordări cu grad ridicat de complexitate bazate pe rețele neuronale.

Obiective

Utilizarea infrastructurii CloudUT pentru aplicații care necesită:

- Calcul paralel pe CPU sau GPU pentru rezolvarea unor probleme complexe. Ex: antrenare sau predicție pentru aplicații bazate pe învățare profundă.
- Spațiu de stocare. Ex: procesele de antrenare, validare și testare utilizează colecții mari de date.

Metodologia de lucru

Dezvoltarea și rularea aplicației pe mașina locală în MATLAB

- Se proiectează modelul de rețea și se stabilesc detaliile de antrenare: numărul de epoci, dimensiunea loturilor (*batch size*), rata de învățare și alți hiperparametrii.

Portarea în infrastructura CloudUT

1. Cerere adresată inginerului de sistem: cuprinde o aproximare estimativă a resurselor hardware necesare.
2. În urma unei analize inginerul de sistem pune la dispoziție o mașină virtuală configurată adecvat accesibilă prin VPN.
3. Se copiază aplicația și datele în infrastructura CloudUT (FTP, RDP).
4. Se rulează secțiunile de calcul intens în cloud (permite scalabilitate pentru rezolvarea unor probleme complexe).

Dezvoltare în MATLAB

MATLAB

- Oferă suport pentru numeroase domenii: calcul numeric, simulări, procesări de imagini, inteligență și viziune artificială, etc.
- Necesită resurse computaționale pentru calcul paralel și distribuit: memorie, CPU, GPU.
- Mediu de dezvoltare extrem de utilizat de comunitatea științifică.

Învățarea profundă în MATLAB

- MATLAB oferă suport pentru definirea și antrenarea de rețele neuronale precum și pentru predicție cu modelele antrenate.
- Procesele de antrenare/predicție sunt scalabile în raport cu resursele de calcul disponibile.
- Creșterea vitezei de lucru în MATLAB este proporțională capacitatea de lucru a mașinii.

MATLAB în cloud

Infrastructura cloud dispune de resurse de calcul substanțiale => timpi de execuție îmbunătățiți pentru aplicațiile MATLAB.

Soluții cloud pentru MATLAB

- MathWorks Cloud – web-based în cloud-ul MathWorks.
- Azure, AWS – suport MATLAB în cloud-uri publice.
- Docker Container – container-izare prin NVIDIA GPU Cloud.
- Cloud Services – acces din MATLAB la servicii cloud publice.
- Hosting Providers – suport Matlab în alte cloud-uri: RONIN, Nimbix, Rescale, UberCloud.

MATLAB în cloud privat

Posibilități de lucru cu MATLAB în cloud privat

- Crearea unei imagini și execuția unui container MATLAB folosind *docker*.
- Utilizarea MATLAB Deep Learning Container oferit de nVidia pentru accelerarea aplicațiilor de învățare profundă – necesită *NVIDIA Container Runtime* și *docker* (implementat pe AWS și sistemele nVidia-DGX).
- Utilizarea unei mașini virtuale echipate - în funcție de necesități - cu o parte din resursele disponibile în cloud – varianta CloudUT.

  Presupune:
  - Acces la mașina virtuală prin VPN și Remote Desktop Protocol (RDP).
  - Instalarea inițială a mediului MATLAB pe mașina virtuală.
  - Portarea aplicațiilor proprii și a datelor de prelucrare.
  - Execuție facilă direct din mediul MATLAB.
  - Licență *Parallel Computing Toolbox* pentru paralelizare și exploatare eficientă a resurselor de calcul oferite.

  Notă: Nu sunt necesare cunoștințe tehnice suplimentare specifice infrastructurii cloud sau de programare în cloud.

MATLAB în CloudUT

Scopuri și Avantaje

- Pentru colectivele de cercetare din UTCN.
- Permite execuția performantă a aplicațiilor folosind infrastructura CloudUT prin alocarea de resurse corespunzătoare aplicațiilor care urmează a fi executate.
- Reducerea timpilor de execuție a aplicațiilor pe care le dețin colectivele de cercetare.
- Posibilitatea testării unor aplicații care nu ar fi putut fi rulate pe un calculator personal cu set restrâns de resurse de calcul și/sau spațiu de stocare insuficient.
- Acces rapid în rețeaua UTCN + confidențialitate.

Aplicații demonstrative

1) Recunoașterea cifrelor scrise de mână (clasificare)

Intrare

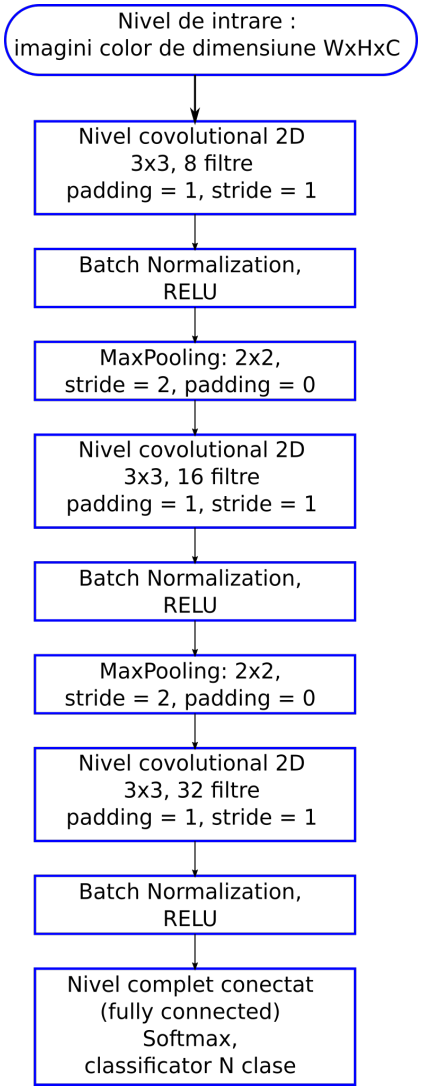
2) Segmentarea semantică a imaginilor color

Intrare Rezultat urmărit

Recunoașterea cifrelor

Rețea neuronală CNN cu arhitectură liniară compusă din straturi de:

- *Convoluție* - dimensiune 3x3 - cu număr variat de filtre (8, 16, 32).
- *Batch Normalization* – pentru a normaliza volumul de date astfel: ieșirea unui strat este obținută din valorile de intrare din care se scade media lor și apoi se împarte cu deviația standard a lor.
- *Rectified Linear Unit (ReLU)* – funcții de activare în cadrul rețelei având la bază funcția de maxim R(z) = max(0,z).
- *MaxPooling* – reduce dimensiunea spațială a volumului de date și consideră maximul din ferestre de 2x2 pixeli.



Recunoașterea cifrelor

Setul de date MNIST (28x28x1)

Sursa: https://en.wikipedia.org/wiki/MNIST_database

Parametrii de antrenare

- numărul de filtre;
- numărul de epoci;
- dimensiunea unui lot (*batch size*).

Recunoașterea cifrelor

Scenariul de lucru

1. Pregătire aplicaţie pe mașina locală și rulare (cu resurse limitate).
2. Portare aplicaţie și rulare în CloudUT.
3. Comparare resurse utilizate (memorie, timp) și a acurateței modelelor antrenate la rulare pe mașina locală față de CloudUT.

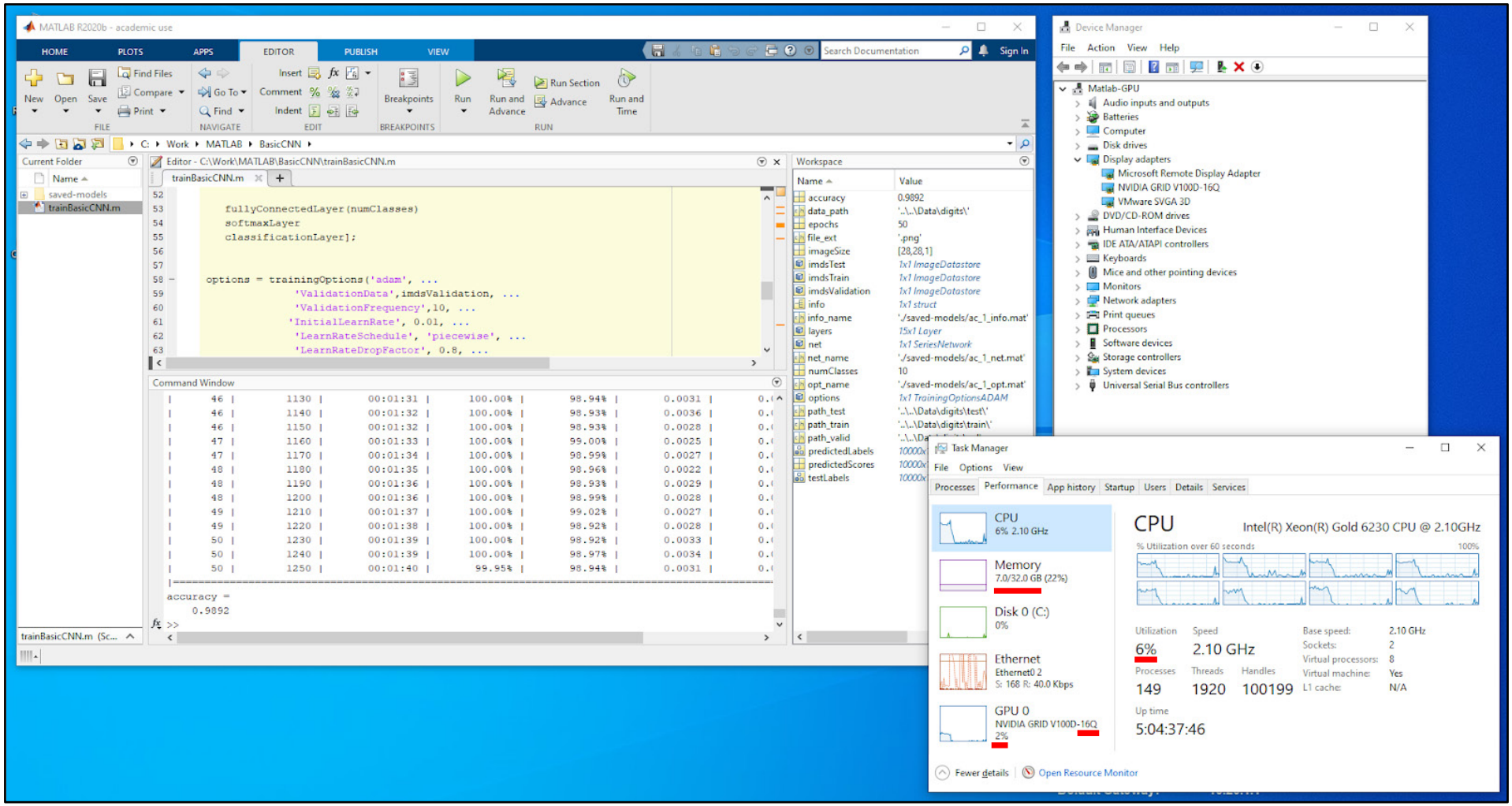
Recunoașterea cifrelor

Configurații de testare

1. stație locală: Windows 10
   - **GPU** NVIDIA GeForce RTX 2080 Ti/PCIe/SSE2 12GB memorie
   - **CPU** Intel i7-3770K@3.5GHz (8 fire de execuție)
   - 16GB **RAM**
2. mașină virtuală 1, în CloudUT: Windows 10
   - **GPU** NVIDIA V100 cu 16GB memorie
   - **CPU** Intel Xeon Gold 6230@2.1GHz (8 nuclee de procesare)
   - 32GB **RAM**
3. mașină virtuală 2, în CloudUT: Windows 10
   - **GPU** NVIDIA V100 cu 32GB memorie
   - **CPU** Intel Xeon Gold 6230@2.1GHz (8 nuclee de procesare)
   - 128GB **RAM**

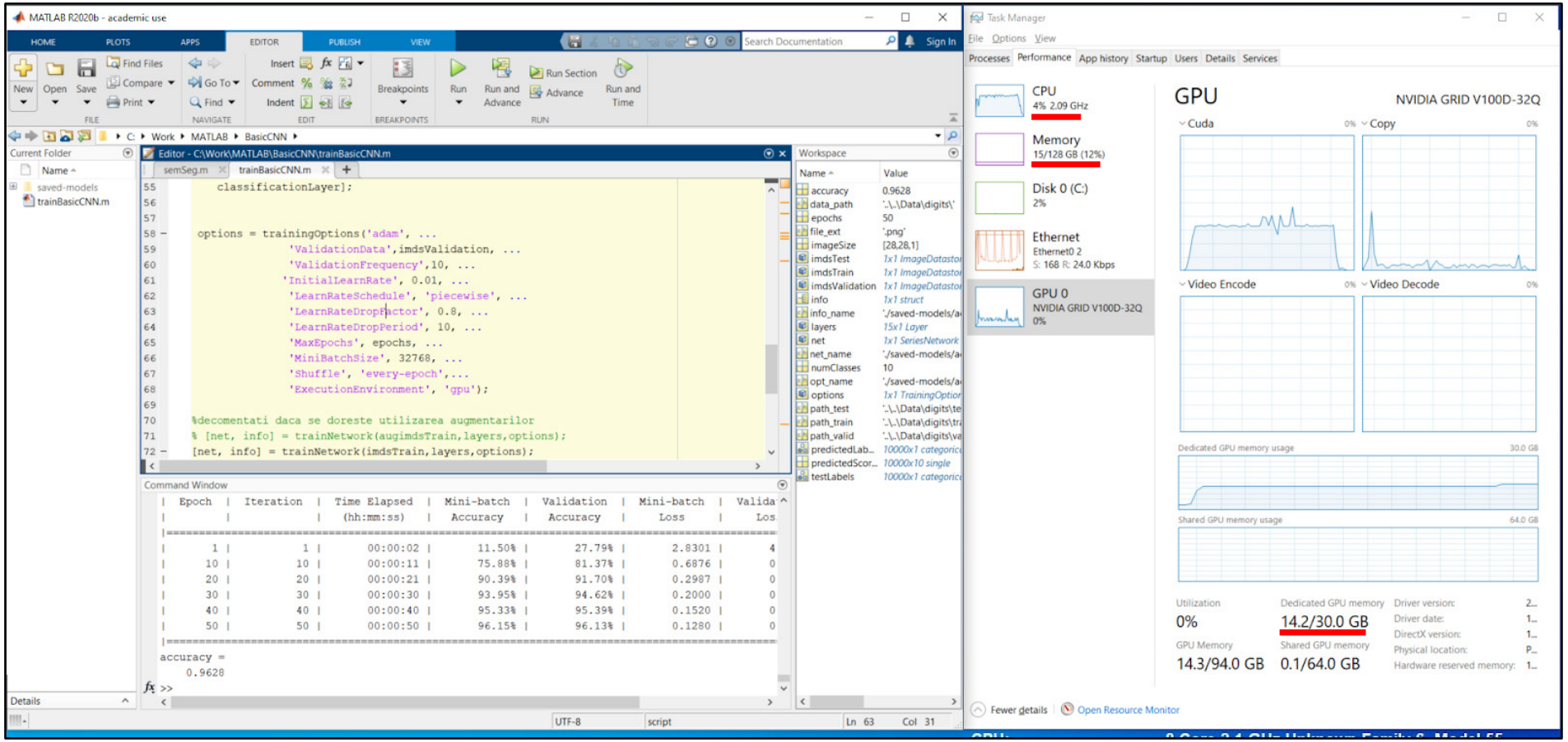
Recunoașterea cifrelor

Sesiune de antrenare în CloudUT cu resursele disponibile și utilizarea acestora pentru mașina virtuală 1



Recunoașterea cifrelor

Sesiune de antrenare în CloudUT cu resursele disponibile și utilizarea acestora pentru mașina virtuală 2



Recunoașterea cifrelor

Rezultate comparative obținute la antrenarea rețelei neuronale

Parametrii			a) Rulare pe stația locală		b) Rulare în CloudUT pe mașina virtuală 1		c) Rulare în CloudUT pe mașina virtuală 2	
Epoci	Dimensiune Imagini	Batch Size	Acuratețe	Timp antrenare (hh:mm:ss)	Acuratețe	Timp antrenare (hh:mm:ss)	Acuratețe	Timp antrenare (hh:mm:ss)
50	28x28x1	1024	0.9899	0:05:42	0.9887	0:02:39	0.9898	0:02:44
50	28x28x1	2048	0.989	0:04:22	0.9892	0:01:40	0.989	0:01:39
50	28x28x1	4096	0.9898	0:03:46	0.99	0:01:13	0.9898	0:01:12
50	28x28x1	8192	0.9876	0:03:29	0.9875	0:01:07	0.9875	0:01:00
50	28x28x1	16384	0.9832	0:03:37	0.9829	0:00:58	0.9829	0:01:00
50	28x28x1	32768	0.9628	0:03:32	0.9628	0:01:06	0.9628	0:00:50

Recunoașterea cifrelor

Analiza comparativă a timpilor de execuție

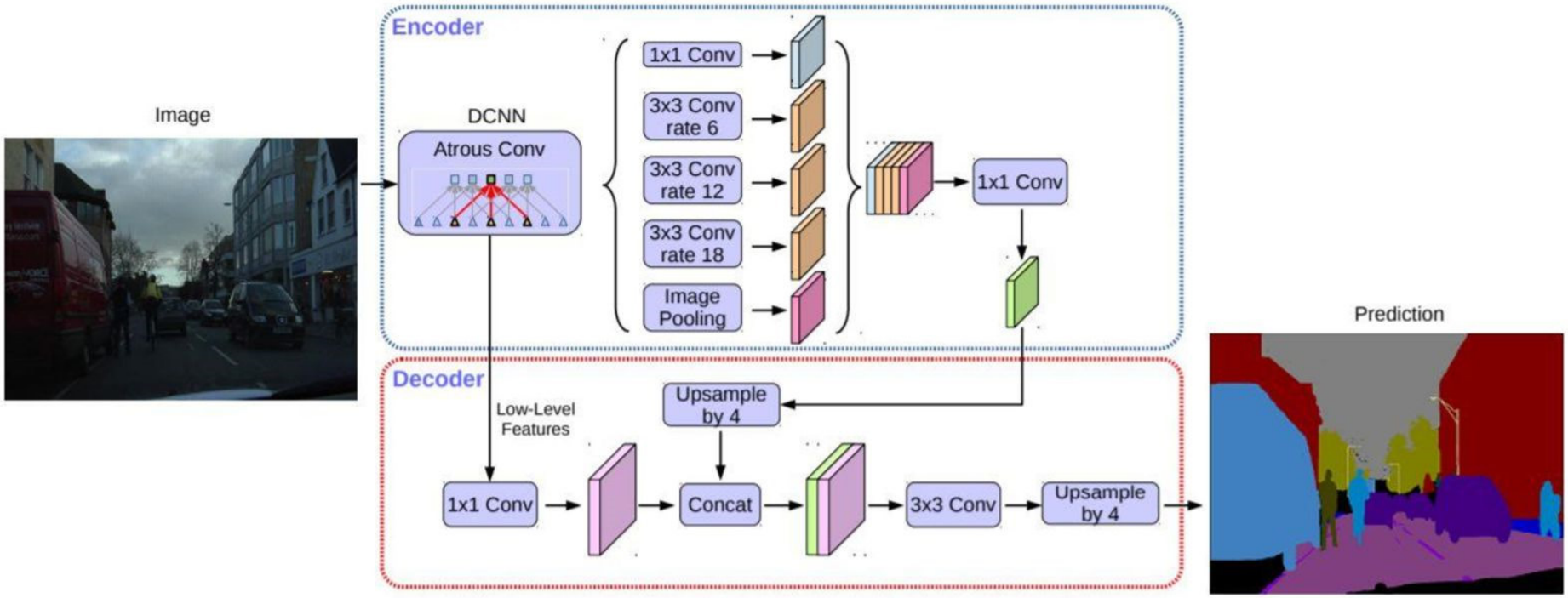
- Pe mașina virtuală 1 din CloudUT față de stația locală se constată o reducere a acestora în medie de cca. 65% (min. 53%, max. 73%).
- Pe mașina virtuală 2 din CloudUT, cu un lot de imagini dimensiune mare (32 768 imagini), se observă o reducere a timpului de execuție de cca. 24% față de timpul obținut de mașina virtuală 1.
  - Se datorează în special calculului intensiv realizat pe procesorul grafic, determinat de dimensiunea mare a lotului de imagini.
  - Calculul s-a finalizat mai repede datorită cantității de memorie disponibilă de 32GB în cazul mașinii virtuale 2 vs. 16GB în cazul mașinii virtuale 1.

Segmentarea semantică

Arhitectura rețelei neuronale

Topologie **DeepLabv3+** [1]: Encoder (Resnet-18) + Decoder

- Straturi utilizate: *Convoluție*, *Batch Normalization*, *ReLU*, *Image Pooling*
- Număr de niveluri: 100
- Resnet-18 [2] (1000 clase): pre-antrenată pe 1 mil. imagini din setul Imagenet (sursa: http://www.image-net.org)

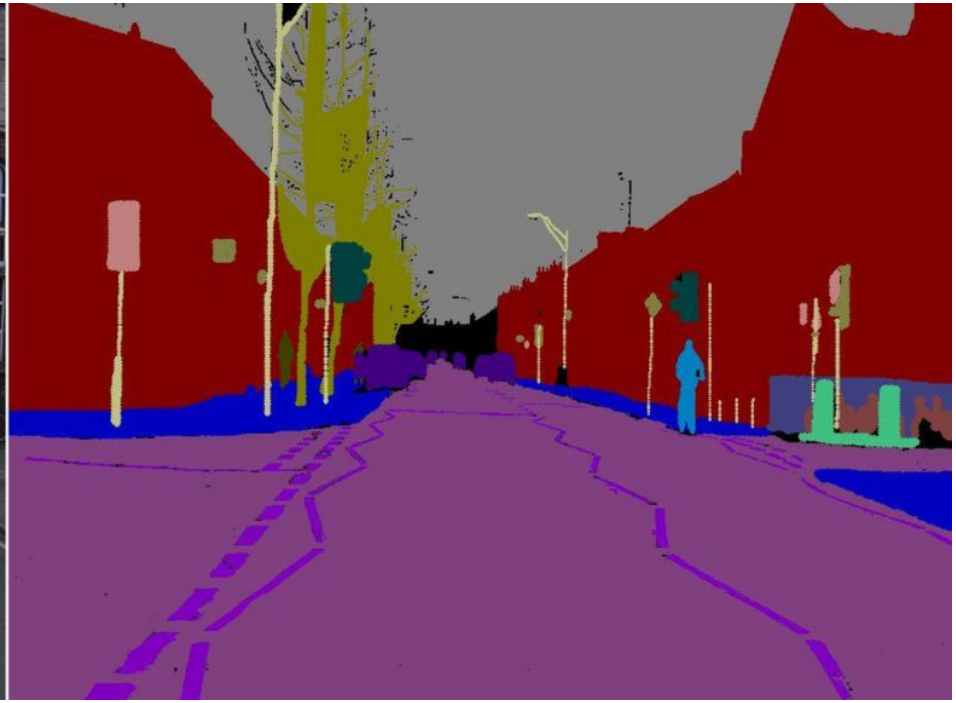


Segmentarea semantică

Setul de date CamVid (701 imagini 960x720x3) – 32 clase

Sursa: http://web4.cs.ucl.ac.uk/staff/g.brostow/MotionSegRecData

Imagine de trafic

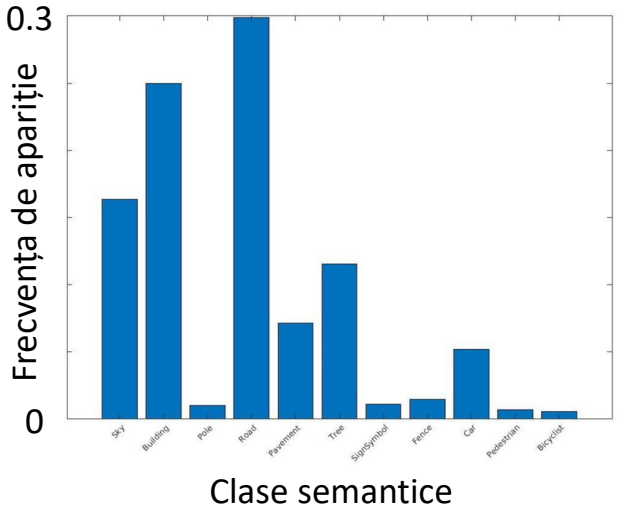


Imagine etichetată

Segmentarea semantică

Pași de procesare

1. Încărcarea datelor + partiționare: **60%** (antrenare) – **20%** (validare) – **20%** (testare).
2. Restrângerea numărului de clase: 32 -> 11 (cer, clădire, piloni, drum, pavaj, copac, semn de circulație, autovehicul, biciclist, pieton, gard).
3. Definirea rețelei DeepLabv3+.
4. Determinarea incidenței claselor în setul de date și ajustarea ponderilor în funcția de eroare la antrenare. → *cross-entropy*
5. Augmentarea setului de antrenare (reflexii, translații).
6. Antrenarea rețelei.
7. Evaluarea performanței pe setul de testare.



Segmentarea semantică

Parametrii de antrenare

- Se utilizează algoritmul de regresie *Stochastic Gradient Descent with Momentum*
- Rata de învățare: 0,003 (inițial), scade cu un factor de 0.3 la fiecare 10 epoci.
- Numărul de epoci: variabil
- Dimensiunea unul lot de antrenare (*batch size*): variabil

Metrici de evaluare

- Acuratețe
- Intersection over Union (IoU)
- Scorul F1 pe contururi (BFScore)

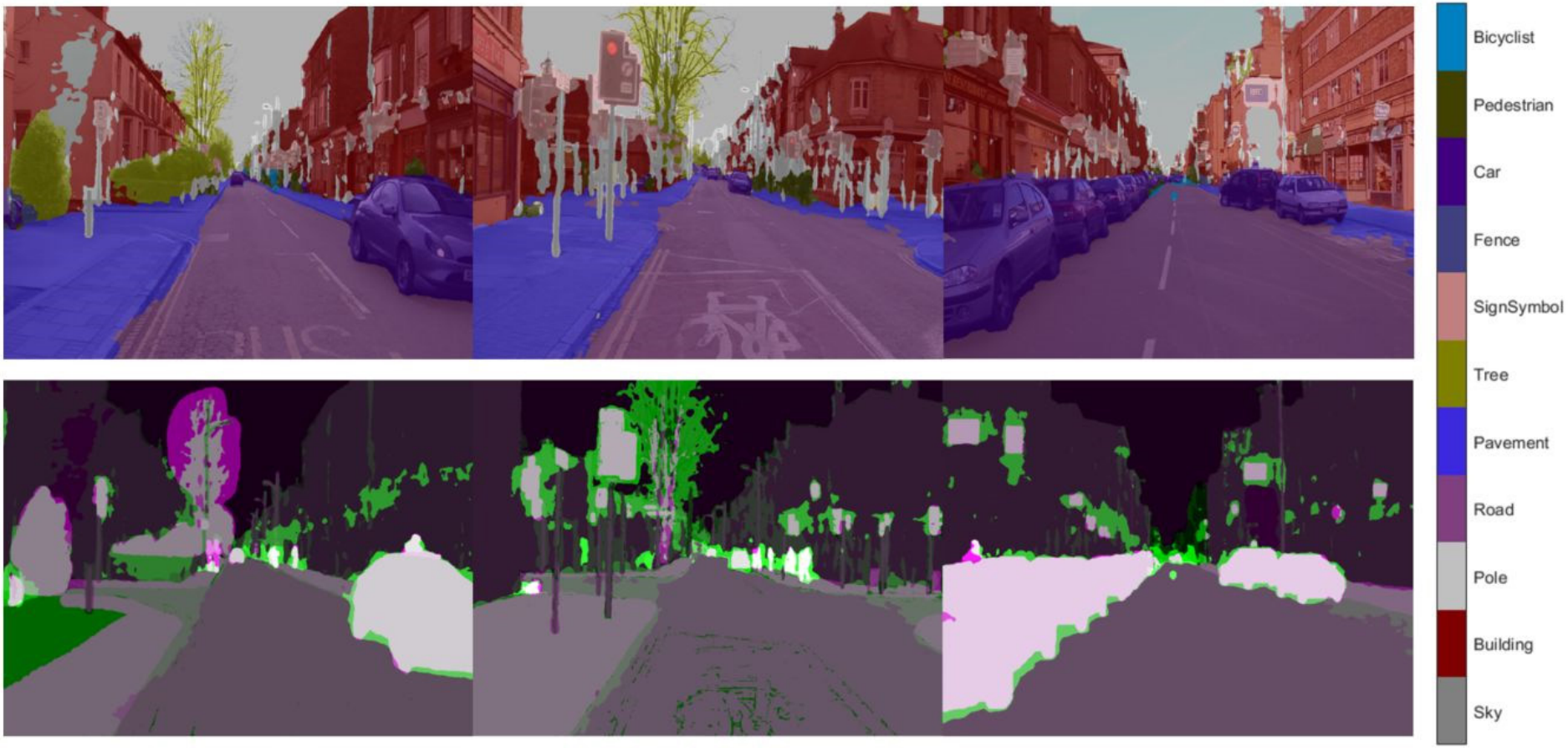
Segmentarea semantică

Configurații de testare

1. stație locală: Windows 10
   - **GPU** NVIDIA GeForce RTX 2060 SUPER 8GB memorie
   - **CPU** Intel i7-3770K@3.5GHz (8 fire de execuție)
   - 16GB **RAM**
2. server local: Windows 10
   - **GPU** NVIDIA GeForce GTX 1080Ti cu 12GB memorie
   - **CPU** Intel i9-7900X@4GHz (20 nuclee de procesare)
   - 128GB **RAM**
3. mașină virtuală 1, în CloudUT: Windows 10
   - **GPU** NVIDIA V100 cu 16GB memorie
   - **CPU** Intel Xeon Gold 6230@2.1GHz (8 nuclee de procesare)
   - 32GB **RAM**

Segmentarea semantică

Rezultate experimentale (mașina virtuală 1, CloudUT)



Batch Size = 8, Antrenare: 1550 iterații (30 epoci)

Segmentarea semantică

Rezultate experimentale (mașina virtuală 1, CloudUT) **pe setul de testare**

Batch Size = 8
Antrenare: 1550 iterații (30 epoci)

Acuratețe globală	Acuratețe medie	IoU mediu	IoU ponderat	BFScore mediu
0.89658	0.86185	0.66822	0.83346	0.70117

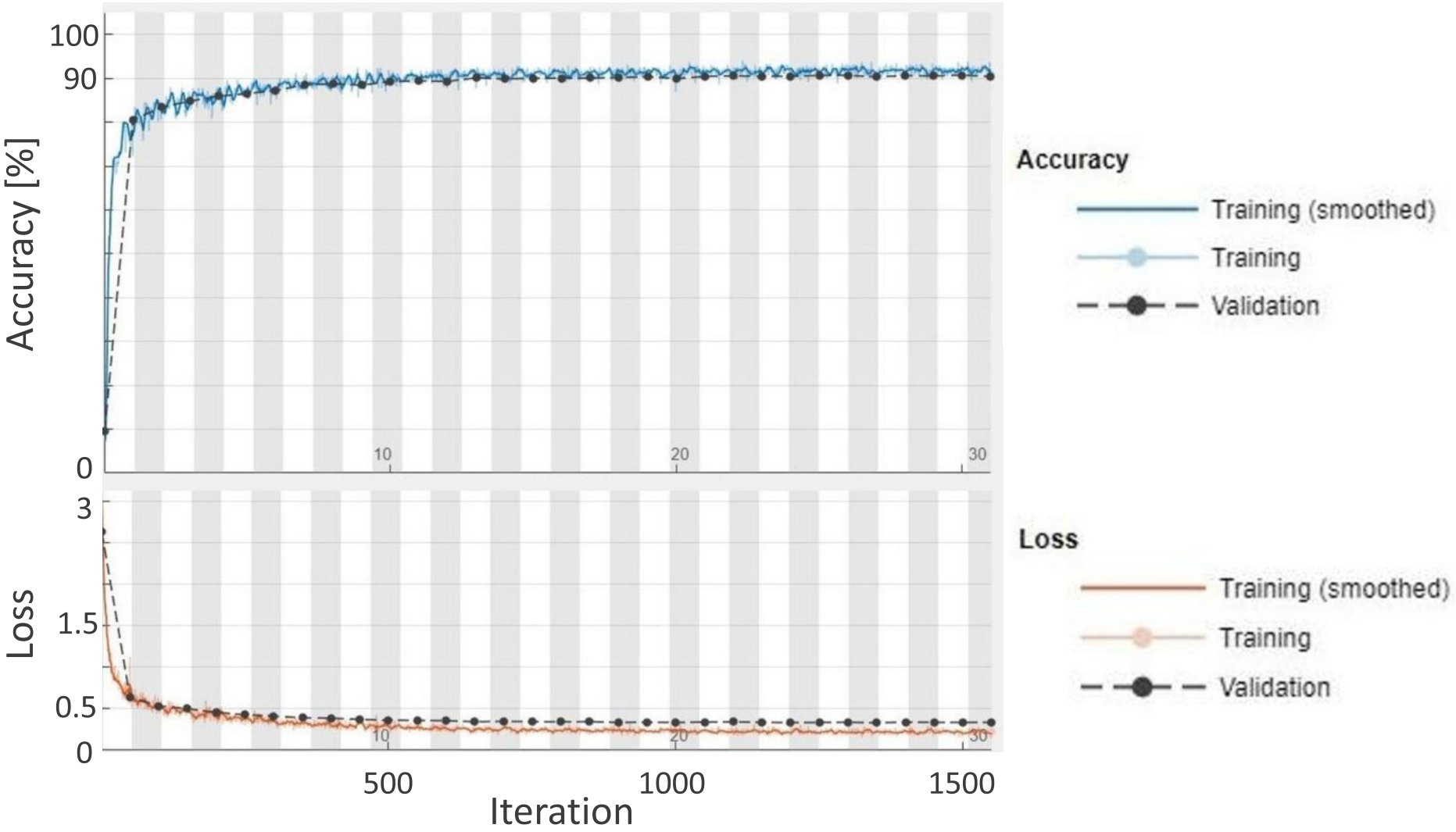
Batch Size > 8 => Eroare: RAM insuficient

Obs: Eficiență mai mică pentru obiectele de dimensiune redusă în setul de antrenare.

Clasă	Setul de testare		
	IoU	Acuratețe	BFScore
Cer	0.90804	0.93887	0.90636
Clădire	0.80341	0.82986	0.66167
Pilon	0.25494	0.74593	0.59589
Drum	0.93057	0.94582	0.81708
Pavaj	0.7472	0.8924	0.76095
Copac	0.77623	0.88676	0.73405
Semn	0.43684	0.75522	0.55289
Gard	0.59367	0.81132	0.59107
Mașină	0.79434	0.92213	0.75142
Pieton	0.4633	0.86498	0.6267
Biciclist	0.64193	0.88703	0.5652

Segmentarea semantică

Rezultate experimentale la antrenare (mașina virtuală 1, CloudUT)



Segmentarea semantică

Rezultate experimentale comparative la antrenare

Parametrii				Rulare în CloudUT pe mașina virtuală$_1$		Rulare pe server local		Rulare pe stație locală	
Epoci	Iterații	Dimensiune Imagini	Batch Size	Acuratețe	Timp (hh:mm:ss)	Acuratețe	Timp (hh:mm:ss)	Acuratețe	Timp (hh:mm:ss)
1	1	960x720x3	8	0.0721	00:00:45	0.0966	00:00:26	0.0721	00:00:49
2	104	960x720x3	8	0.8375	00:07:44	0.8366	00:06:59	0.8364	00:16:07
4	208	960x720x3	8	0.8475	00:14:43	0.8463	00:13:34	0.8468	00:31:01
8	400	960x720x3	8	0.9042	00:27:34	0.8871	00:25:21	-	-
16	800	960x720x3	8	0.9178	00:55:12	0.9177	00:51:30	-	-
25	1300	960x720x3	8	0.9244	01:27:38	0.9246	01:23:15	-	-
30	1550	960x720x3	8	0.9079	01:43:53	0.9083	01:39:09	-	-

Obs: $V_{m\_virtuală1} \approx V_{server} = 2.2 \times V_{st\_locală}$ Resurse insuficiente

Segmentarea semantică

Configurații de testare

1. server local: Windows 10
   - **GPU** NVIDIA GeForce GTX 1080Ti cu 12GB memorie
   - **CPU** Intel i9-7900X@4GHz (20 nuclee de procesare)
   - 128GB **RAM**
2. mașină virtuală 2, în CloudUT: Windows 10
   - **GPU** NVIDIA V100 cu 32GB memorie
   - **CPU** Intel Xeon Gold 6230@2.1GHz (8 nuclee de procesare)
   - 128GB **RAM**

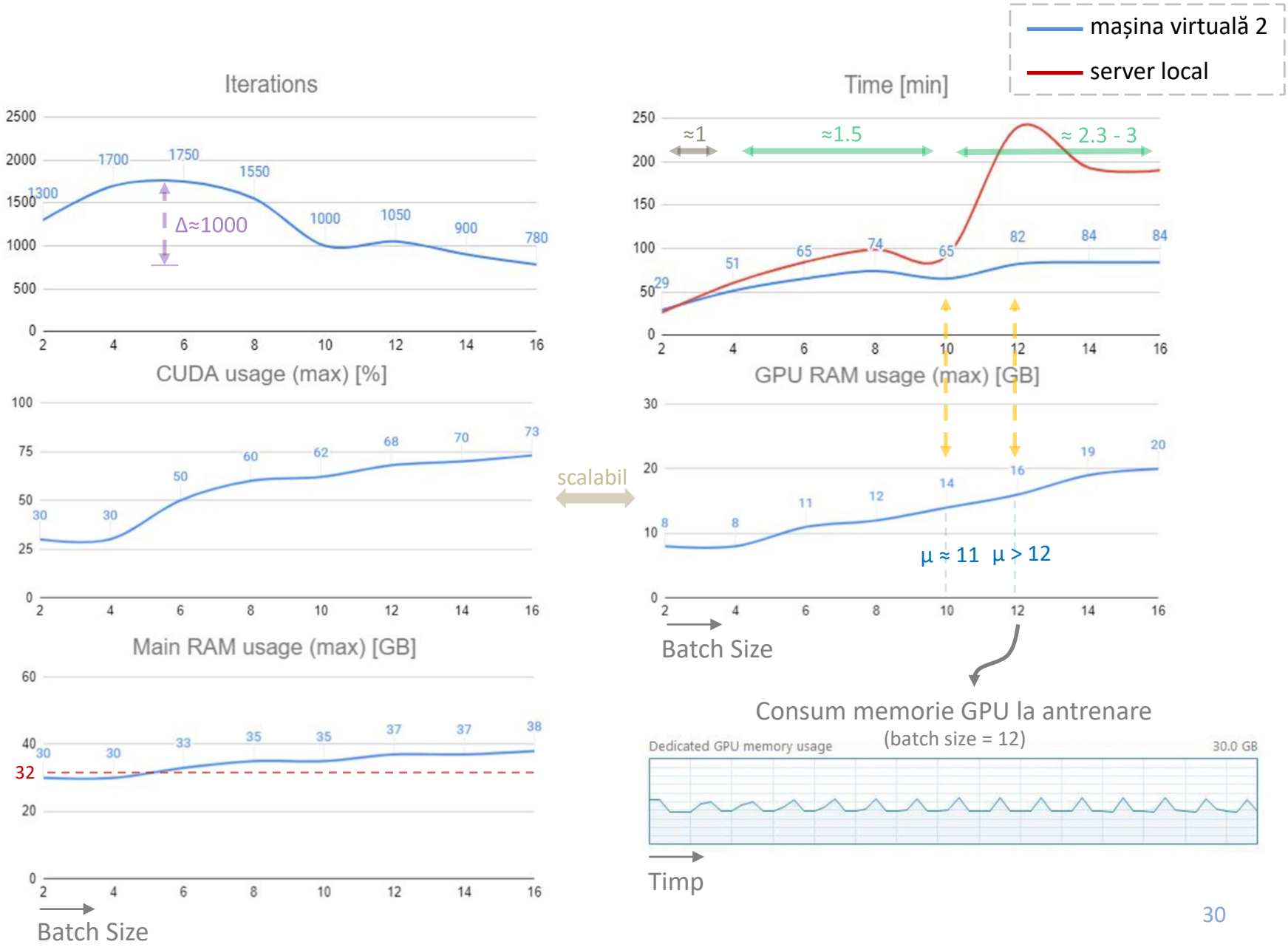
Segmentarea semantică

Consumul de timp și resurse la antrenare (max. 30 epoci)

Batch size	Iterații necesare	CloudUT - mașina virtuală 2 -		Server local		CloudUT (valori maxime) - mașina virtuală 2 -		
		Timp (hh:mm)	Performanță viteză [%]	Timp (hh:mm)	Performanță viteză [%]	CUDA [%]	Memorie GPU (GB)	RAM (GB)
2	1300	00:29	355	00:26	396	30	8	30
4	1700	00:51	202	01:00	172	30	8	30
6	1750	01:05	158	01:24	123	50	11	33
8	1550	01:14	139	01:39	104	60	12	35
10	1000	01:05	158	01:31	113	62	14	35
12	1050	01:22	126	04:00	43	68	16	37
14	900	01:24	123	03:13	53	70	19	37
16	780	01:24	123	03:10	54	73	20	38

Performanță = Timp / $\text{Timp}_{\text{m\_virtuală1(batch size=8)}}$

Rezultate experimentale comparative la antrenare (max. 30 epoci)



Segmentarea semantică

Analiza comparativă

- MATLAB utilizează algoritmi eficienți de paralelizare => scalabilitate a performanței în raport cu resursele disponibile.
- Batch Size redus –> consum redus de GPU (memorie/CUDA)
  => performanțe de viteză asemănătoare cu îmbunătățire până la 150% pentru mașina virtuală 2.
- Batch Size ridicat -> consum ridicat de GPU (memorie/CUDA)
  => performanțe de viteză crescute până la 230-300% pentru mașina virtuală 2.
- Calculul paralel realizat pe GPU permite o stabilizare a timpului de lucru pentru loturi mari de date în raport cu loturi mici atunci când resursele de calcul sunt suficiente.
- Pentru utilizarea unor modele complexe cantitatea RAM alocată este indicat să depășească 32G.

Concluzii

Beneficiile CloudUT pentru aplicații MATLAB de învățare profundă

- Timp de execuţie redus în raport cu complexitatea problemei.
- Posibilitatea antrenării cu loturi mai mari de date poate avea efect pozitiv asupra acurateței modelului de rețea.

Paşi de utilizare a infrastructurii CloudUT de către colectivele de cercetare

- Dezvoltarea integrală a aplicației (MATLAB) pe maşina locală.
- Stabilirea resurselor hardware necesare pentru rulare mai eficientă.
- Cerere adresată administratorilor CloudUT => alocarea unei mașini.
- Instalarea aplicațiilor și portarea datelor pe maşina alocată.
- Rularea aplicației și efectuarea experimentelor care necesită calcul intens.

Referințe

- [1] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, Hartwig Adam. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. Proceedings of 15th European Conference on Computer Vision (ECCV2018), pp 833-851, September 8–14, 2018.
- [2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. Deep Residual Learning for Image Recognition. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770-778, June 2016, DOI: 10.1109/CVPR.2016.90.

Mulțumim pentru atenție!

Ion Giosan, Cristian-Cosmin Vancea
Departamentul Calculatoare
Universitatea Tehnică din Cluj-Napoca
Ion.Giosan@cs.utcluj.ro
Cristian.Vancea@cs.utcluj.ro