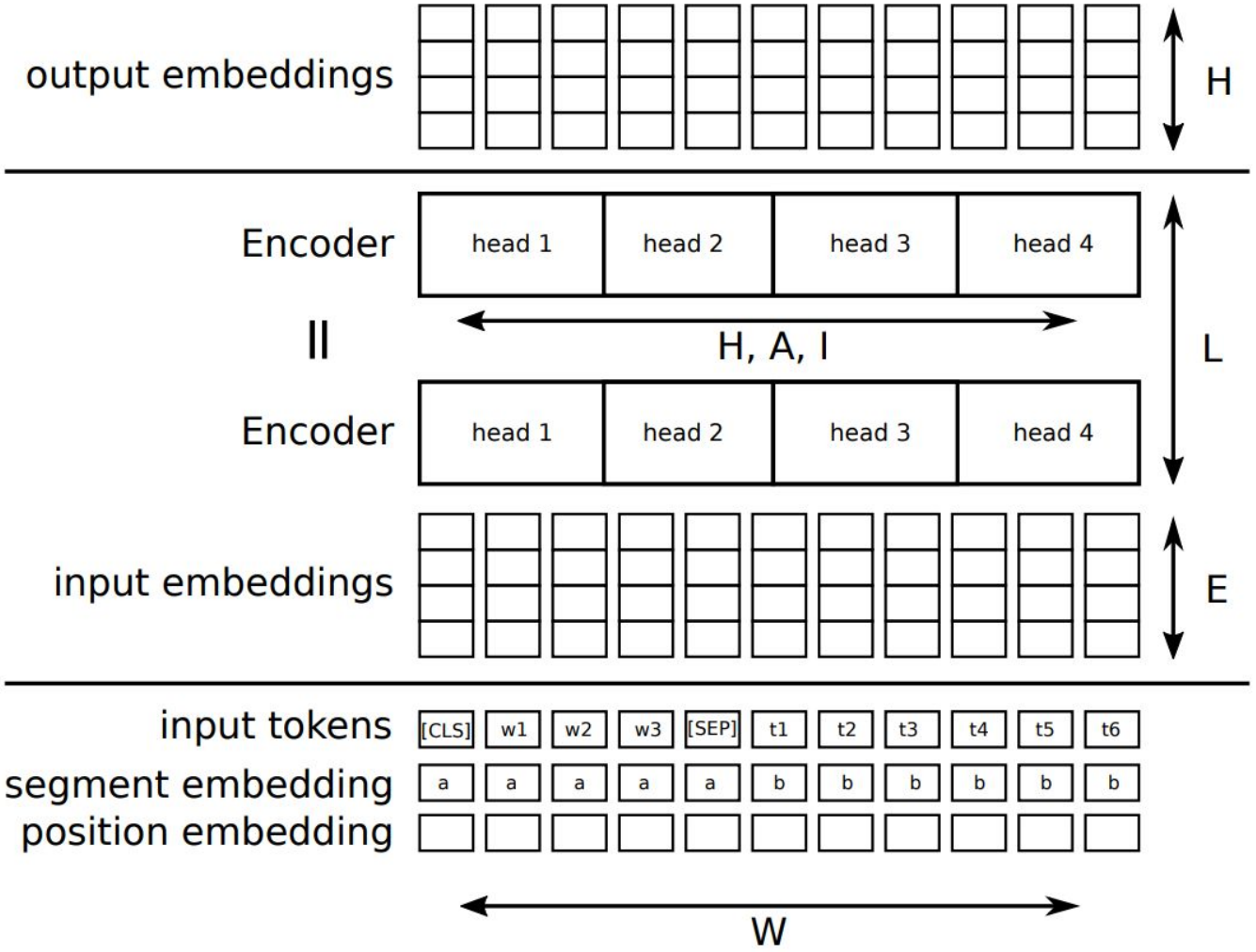
ALBERT Speeds

Petr Zelina
469366

Parameters

ALBERT base

- vocab: 30K
- **W** width: 512
- **L** layers: 12
- **A** heads: 12
- **H** hidden: 768
- **I** forward: 3072
- **E** embed: 128



Implementations – TensorFlow

- based on the Google research ALBERT
- primarily meant for TPU cloud
- TF 1.5 (old dependencies)
- supports single GPU
- no mixed precision
- 2 step pipeline

Google research ALBERT: https://github.com/google-research/albert
csalbert: https://github.com/ZepZep/csalbert

Implementations – PyTorch

- much more readable
- missing evaluation
- slow all in one pipeline
- possible mixed precision
- possible multiGPU

ALBERT Pytorch: https://github.com/graykode/ALBERT-Pytorch
Czech version: https://github.com/ZepZep/ALBERT-Pytorch/tree/csalbert

Machines

- **1060 6GB**
  - 1 280 CUDA cores 120 W 5 years old
  - home PC
  - $200, 4 300 CZK (bought for 8 000 CZK)
- **T4 16 GB**
  - 2 560 CUDA cores 70 W 3 years old
  - Metacentrum adan / NLP MUNI apollo
  - $3 000, 60 000 CZK
- **A100 40 GB**
  - 6 912 CUDA cores 400 W less than a year old
  - BostonLabs
  - $9 600, 205 000 CZK

TF results

- did not manage to run on A100

GPU	framework	mixed	vocab	width	layers	hidden	forward	batch_size	memory	batch/s	mem/ex	ex/s
TensorFlow												
1060	TF	FALSE	30 000	512	12	768	3 072	6	5 515	1,13	919,17	6,78
T4	TF	FALSE	30 000	512	12	768	3 072	16	15 109	0,61	944,31	9,76

PyTorch – from the box

- relatively simple installation

GPU	framework	mixed	vocab	width	layers	hidden	forward	batch_size	memory	batch/s	mem/ex	ex/s
From the box												
1060	torch	FALSE	30 522	512	12	768	3 072	4	3 540	1,44	884,95	5,76
T4	torch	FALSE	30 522	512	12	768	3 072	15	12 724	0,62	848,25	9,32
A100	torch	FALSE	30 522	512	12	768	3 072	40	33 604	1,19	840,09	47,60

PyTorch – optimized pipeline

- with bigger batch sizes GPU waited for training example creation
- low / unstable GPU utilization

  -> multiprocessing + prefetching

- 1.05 x 1060 speedup
- 1.05 x T4 speedup
- 1.60 x A100 speedup

PyTorch – mixed precision

- For some parts of the NN it is not necessary to use 32 bit floats
  -> 16 bit floats instead
- automatic support in PyTorch
  - ```
    with torch.cuda.amp.autocast(enabled=True):
        # backpropagation
    ```
- Enables bigger batch size

- 1.05 x 1060 speedup
- 2.20 x T4 speedup
- 1.22 x A100 speedup
    ```

PyTorch – multiGPU

- only managed to test T4
- possible to use on MetaCentrum
- data parallelism
  - different batches on each device - each calculates gradient
  - average gradient across all devices
  - sync weights

- 2 x T4 GPUs on apollo
- 1.85 x T4 speedup

GPU	framework	mixed	vocab	width	layers	hidden	forward	batch_size	memory	batch/s	mem/ex	ex/s	mem	vs 1060	vs T4	vs A100
From the box																
1060	torch	FALSE	30 522	512	12	768	3 072	4	3 540	1,44	884,95	5,76	1,00	0,95	0,58	0,07
T4	torch	FALSE	30 522	512	12	768	3 072	15	12 724	0,62	848,25	9,32	0,96	1,53	0,95	0,12
A100	torch	FALSE	30 522	512	12	768	3 072	40	33 604	1,19	840,09	47,60	0,95	7,82	4,84	0,62
Optimized pipeline																
1060	torch	FALSE	30 522	512	12	768	3 072	4	3 540	1,521	884,95	6,08	1,00	1,00	0,62	0,08
T4	torch	FALSE	30 522	512	12	768	3 072	15	12 724	0,656	848,25	9,84	0,96	1,62	1,00	0,13
A100	torch	FALSE	30 522	512	12	768	3 072	40	33 604	1,931	840,09	77,24	0,95	12,70	7,85	1,00
Mixed precision																
1060	torch	TRUE	30 522	512	12	768	3 072	6	4 096	1,066	682,74	6,40	0,77	1,05	0,65	0,08
T4	torch	TRUE	30 522	512	12	768	3 072	20	13 160	1,08	658,00	21,60	0,74	3,55	2,20	0,28
A100	torch	TRUE	30 522	512	12	768	3 072	48	31 198	1,97	649,95	94,56	0,73	15,54	9,61	1,22
MultiGPU																
2 x T4	torch	TRUE	30 522	512	12	768	3 072	36	25 429	1,116	706,36	40,18	0,80	6,60	4,08	0,52

Notes

- pipeline optimization needed for faster GPUS
- Mixed precision
  - does not help for 1060
  - semi-automatic mixed precision with A100
- T4 has good multiGPU scaling
- ALBERT base Training times (512 M examples)

1060	2 years 8 months
T4	9 months
2 x T4	5 months
13 x T4	23 days
A100	2 months
4 x A100	16 days
64 x Google TPU	3 days

TF vs PyTorch

- TF vs PyTorch optimized pipeline
  - comparable on T4
  - TF slightly faster on 1060
- PyTorch Mixed + multiGPU -> 4x speedup, ease of use

GPU	framework	mixed	vocab	width	layers	hidden	forward	batch_size	memory	batch/s	mem/ex	ex/s	mem	vs 1060	vs T4	vs A100
Optimized pipeline																
1060	torch	FALSE	30 522	512	12	768	3 072	4	3 540	1,521	884,95	6,08	1,00	1,00	0,62	0,08
T4	torch	FALSE	30 522	512	12	768	3 072	15	12 724	0,656	848,25	9,84	0,96	1,62	1,00	0,13
A100	torch	FALSE	30 522	512	12	768	3 072	40	33 604	1,931	840,09	77,24	0,95	12,70	7,85	1,00
Mixed precision																
1060	torch	TRUE	30 522	512	12	768	3 072	6	4 096	1,066	682,74	6,40	0,77	1,05	0,65	0,08
T4	torch	TRUE	30 522	512	12	768	3 072	20	13 160	1,08	658,00	21,60	0,74	3,55	2,20	0,28
A100	torch	TRUE	30 522	512	12	768	3 072	48	31 198	1,97	649,95	94,56	0,73	15,54	9,61	1,22
MultiGPU																
2 x T4	torch	TRUE	30 522	512	12	768	3 072	36	25 429	1,116	706,36	40,18	0,80	6,60	4,08	0,52
TensorFlow																
1060	TF	FALSE	30 000	512	12	768	3 072	6	5 515	1,13	919,17	6,78	1,04	1,11	0,69	0,09
T4	TF	FALSE	30 000	512	12	768	3 072	16	15 109	0,61	944,31	9,76	1,07	1,60	0,99	0,13

Parameter effects

- effect of reducing certain parameters on
  - memory
  - speed

Example for A100

framework	mixed	vocab	width	layers	hidden	forward	batch_size	memory	batch/s	mem/ex	ex/s	mem	speed	mul
torch	TRUE	30522	512	12	768	3072	48	31 198	1,97	649,95	94,56	0,87	16,68	1,00
torch	TRUE	30522	512	12	768	2048	48	28 865	2,106	601,35	101,09	0,80	17,83	1,07
torch	TRUE	30522	512	12	512	3072	48	23 672	2,4	493,17	115,20	0,66	20,32	1,22
torch	TRUE	30522	512	8	768	3072	48	21 548	2,649	448,93	127,15	0,60	22,43	1,34
torch	TRUE	30522	256	12	768	3072	48	13 120	3,098	273,33	148,70	0,37	26,24	1,57
torch	TRUE	20000	512	12	768	3072	48	30 594	2,058	637,37	98,78	0,85	17,43	1,04
torch	FALSE	30522	512	12	768	3072	40	33 604	1,931	840,09	77,24	1,12	13,63	0,82

Parameter effects – summary

	1060		1060 relative		T4		T4 relative		A100		A100 relative	
model	mem/ex	ex/s	mem/ex	ex/s	mem/ex	ex/s	mem/ex	ex/s	mem/ex	ex/s	mem/ex	ex/s
ALBERT base	748,12	5,67	1,00	1,00	658,00	21,60	1,00	1,00	649,95	94,56	1,00	1,00
forward 2/3	687,67	6,54	0,92	1,15					601,35	101,09	0,93	1,07
hidden 2/3	554,87	7,67	0,74	1,35					493,17	115,20	0,76	1,22
layers 2/3	531,11	7,95	0,71	1,40					448,93	127,15	0,69	1,34
width 1/2	371,70	10,15	0,50	1,79					273,33	148,70	0,42	1,57
vocab 2/3	730,54	5,84	0,98	1,03					637,37	98,78	0,98	1,04
no mixed	884,95	5,76	1,18	1,02	848,25	9,84	1,29	0,46	840,09	77,24	1,29	0,82