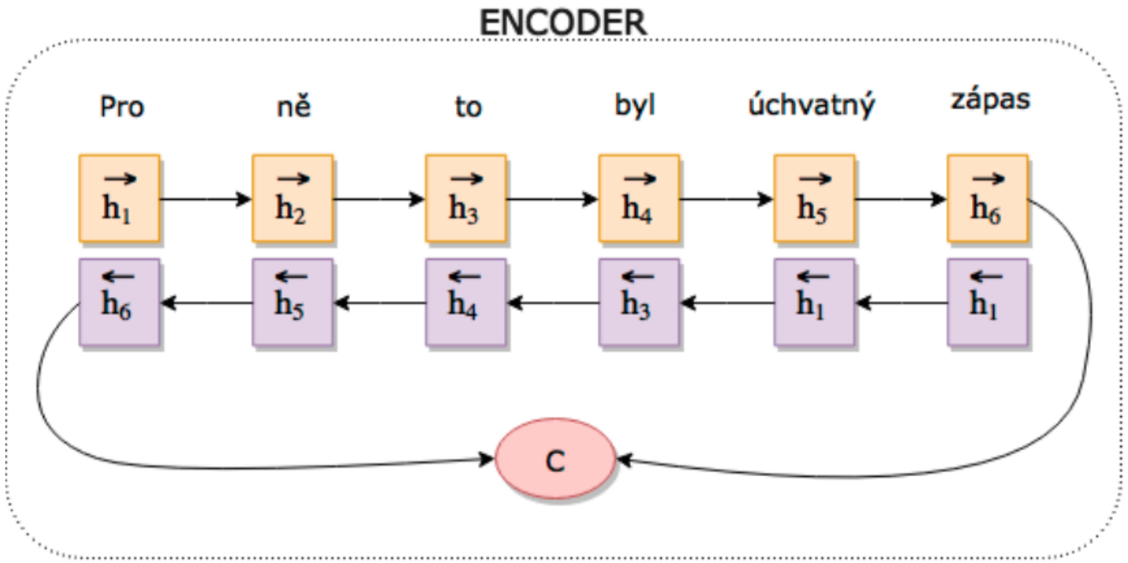
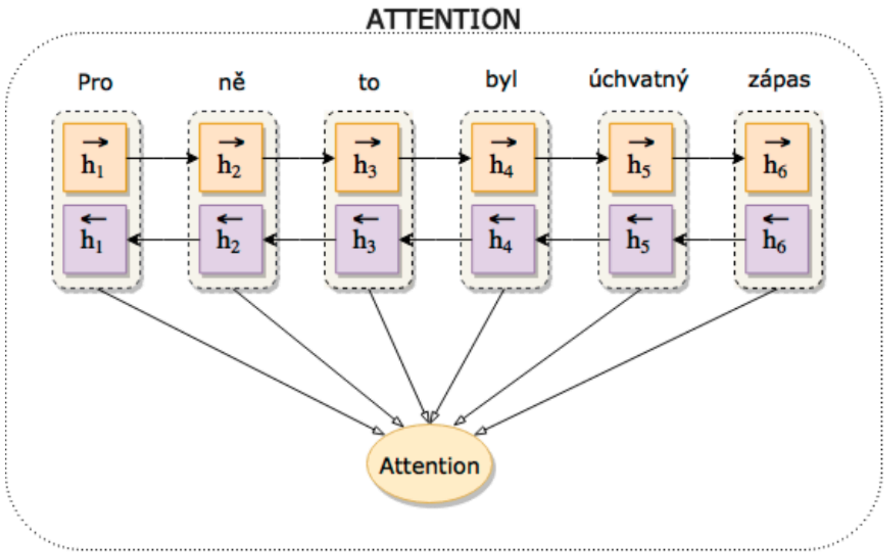
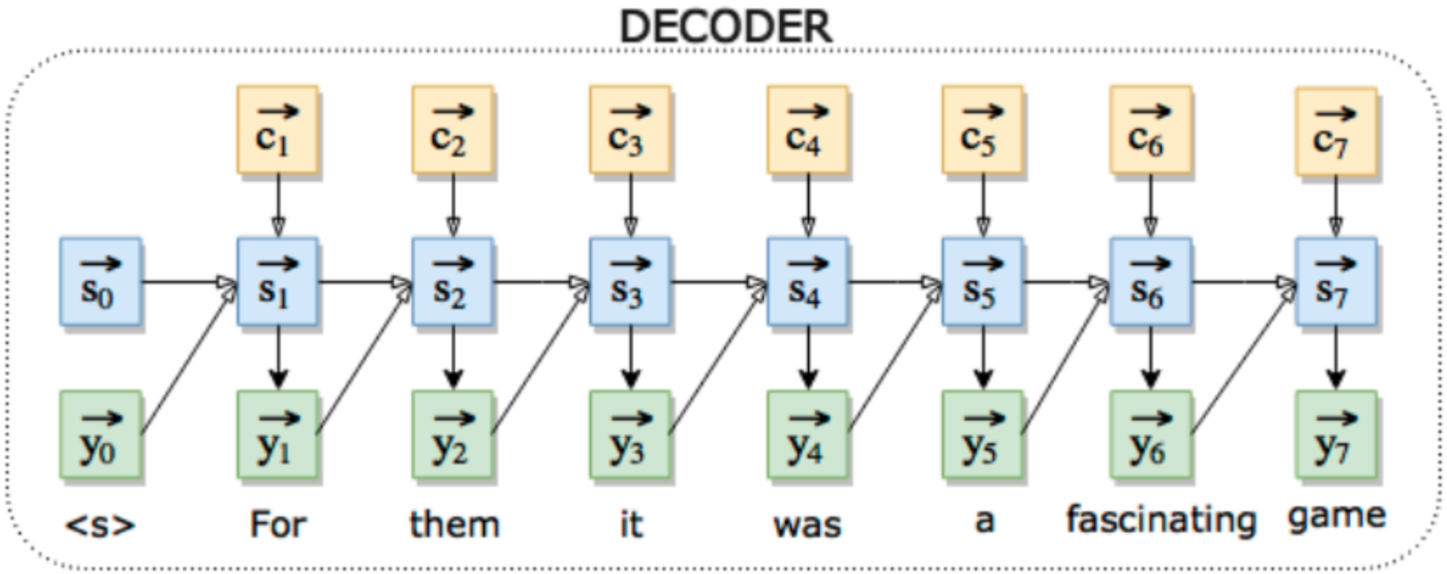
Fast decoding in neural machine translation with Ray

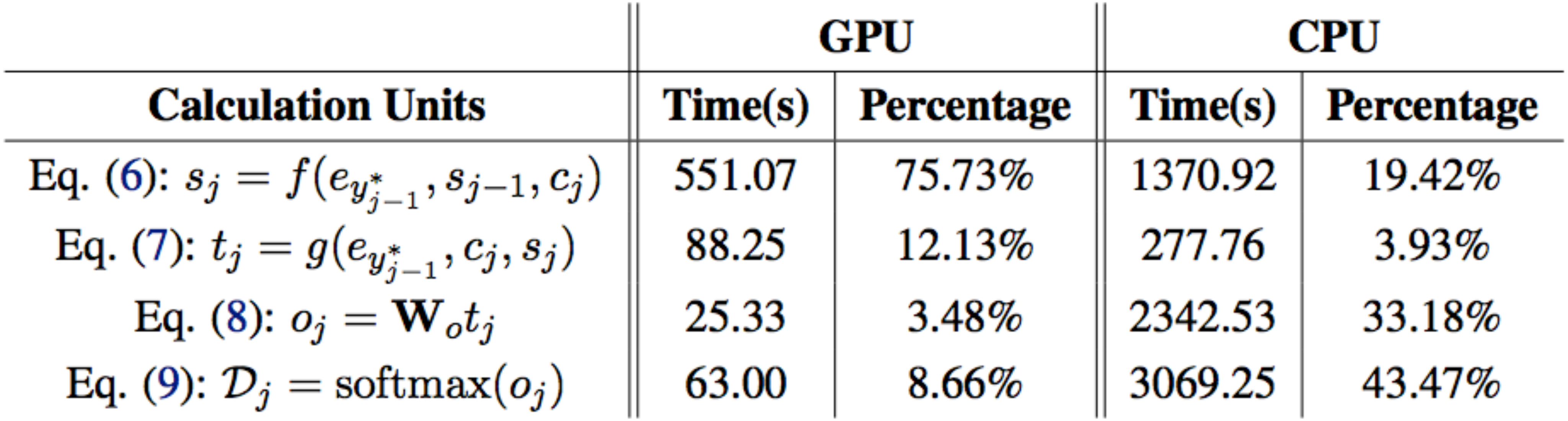
MAREK STRELEC


ENCODER
Pro
ně
to
byl
úchvatný
zápas
$\overrightarrow{h}_1$
$\overrightarrow{h}_2$
$\overrightarrow{h}_3$
$\overrightarrow{h}_4$
$\overrightarrow{h}_5$
$\overrightarrow{h}_6$
$\overleftarrow{h}_6$
$\overleftarrow{h}_5$
$\overleftarrow{h}_4$
$\overleftarrow{h}_3$
$\overleftarrow{h}_1$
$\overleftarrow{h}_1$
C


ATTENTION
Pro
ně
to
byl
úchvatný
zápas
$\overrightarrow{h}_1$
$\overrightarrow{h}_2$
$\overrightarrow{h}_3$
$\overrightarrow{h}_4$
$\overrightarrow{h}_5$
$\overrightarrow{h}_6$
$\overleftarrow{h}_1$
$\overleftarrow{h}_2$
$\overleftarrow{h}_3$
$\overleftarrow{h}_4$
$\overleftarrow{h}_5$
$\overleftarrow{h}_6$
Attention


DECODER
$\overrightarrow{c}_1$
$\overrightarrow{c}_2$
$\overrightarrow{c}_3$
$\overrightarrow{c}_4$
$\overrightarrow{c}_5$
$\overrightarrow{c}_6$
$\overrightarrow{c}_7$
$\overrightarrow{s}_0$
$\overrightarrow{s}_1$
$\overrightarrow{s}_2$
$\overrightarrow{s}_3$
$\overrightarrow{s}_4$
$\overrightarrow{s}_5$
$\overrightarrow{s}_6$
$\overrightarrow{s}_7$
$\overrightarrow{y}_0$
$\overrightarrow{y}_1$
$\overrightarrow{y}_2$
$\overrightarrow{y}_3$
$\overrightarrow{y}_4$
$\overrightarrow{y}_5$
$\overrightarrow{y}_6$
$\overrightarrow{y}_7$
<s>
For
them
it
was
a
fascinating
game

Time cost statistics for decoding



	GPU		CPU	
Calculation Units	Time(s)	Percentage	Time(s)	Percentage
Eq. (6): $s_j = f(e_{y^*_{j-1}}, s_{j-1}, c_j)$	551.07	75.73%	1370.92	19.42%
Eq. (7): $t_j = g(e_{y^*_{j-1}}, c_j, s_j)$	88.25	12.13%	277.76	3.93%
Eq. (8): $o_j = \mathbf{W}_o t_j$	25.33	3.48%	2342.53	33.18%
Eq. (9): $\mathcal{D}_j = \mathrm{softmax}(o_j)$	63.00	8.66%	3069.25	43.47%

Zhang, Wen, et al. (2018)

Ray

- “A flexible, high-performance distributed execution framework”
- Implements a dynamic task graph computation model
- Global Control Store
- Bottom-up distributed scheduler
- Actor abstraction

Steps

- □ Implement an NMT model in TensorFlow
- □ Train the model on a subset of parallel data (Europarl)
- □ Experiments
  - □ Distributed batched translation
  - □ Distributed Beam Search
  - □ Dynamic Beam Search
  - □ Heterogeneous environment
- □ Compare times and BLEU score

Thank you!