Presentation - PSL week - Green AI

Erwan Fagnou

Who are we?

Erwan Fagnou (me) - PhD student

Alexandre Allauzen - Researcher & professor

Paul Caillon - PostDoc

→ Miles team, LAMSADE lab, Dauphine University

What do I do?

Frugal deep learning…

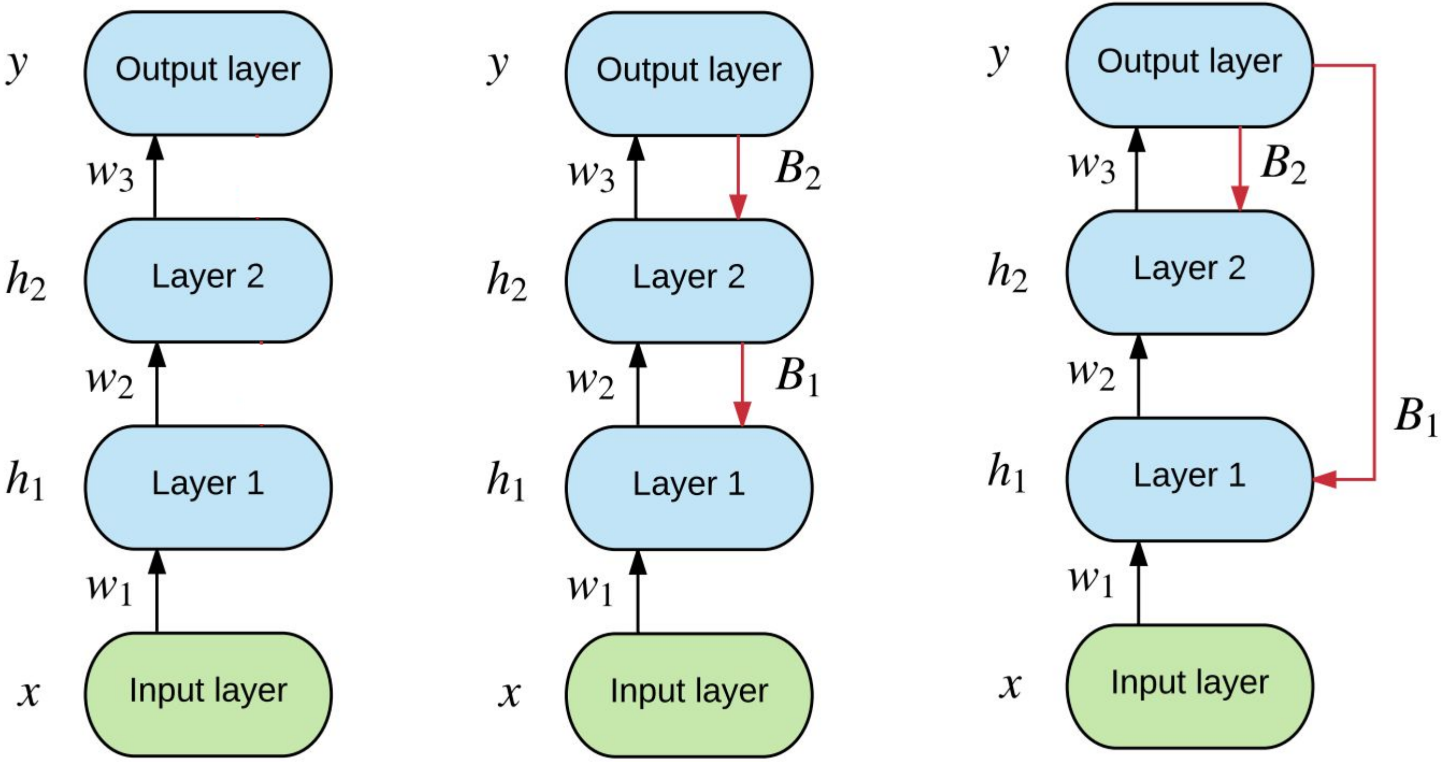
- focus on improving the architecture
- focus on transformers and NLP
- interest in adaptive architectures

“Making models more resource-efficient”

- during training or inference?
- what resource are we minimizing? time? memory?
- how? examples:
    - improve training (optimizer)
    - improve tensor operation efficiency (FP16, FlashAttention…)
    - reduce number of parameters (growing, pruning, distillation…)
    - improve architecture (linear attention alternatives, adaptive…)
    - improve hyperparameter search (NAS, warmup with Adam…)

Alternative optimization algorithms

Direct Feedback Alignment



(A) BP

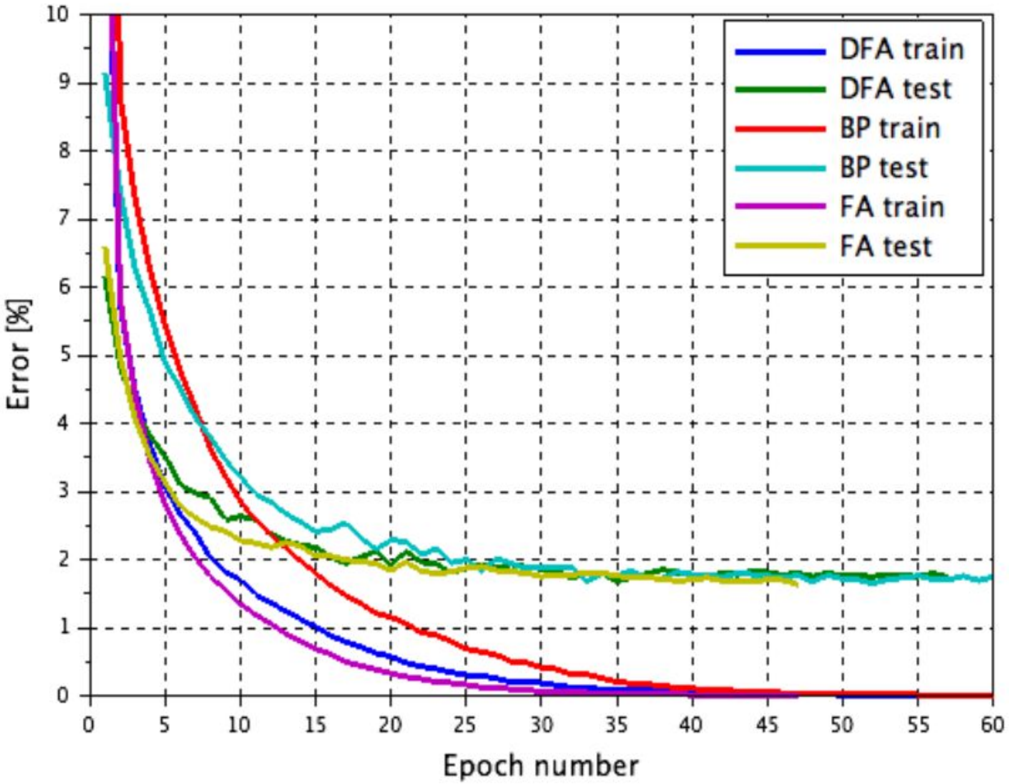
(B) FA

(C) DFA

TP Lillicrap · 2014

A Nøkland · 2016

Direct Feedback Alignment



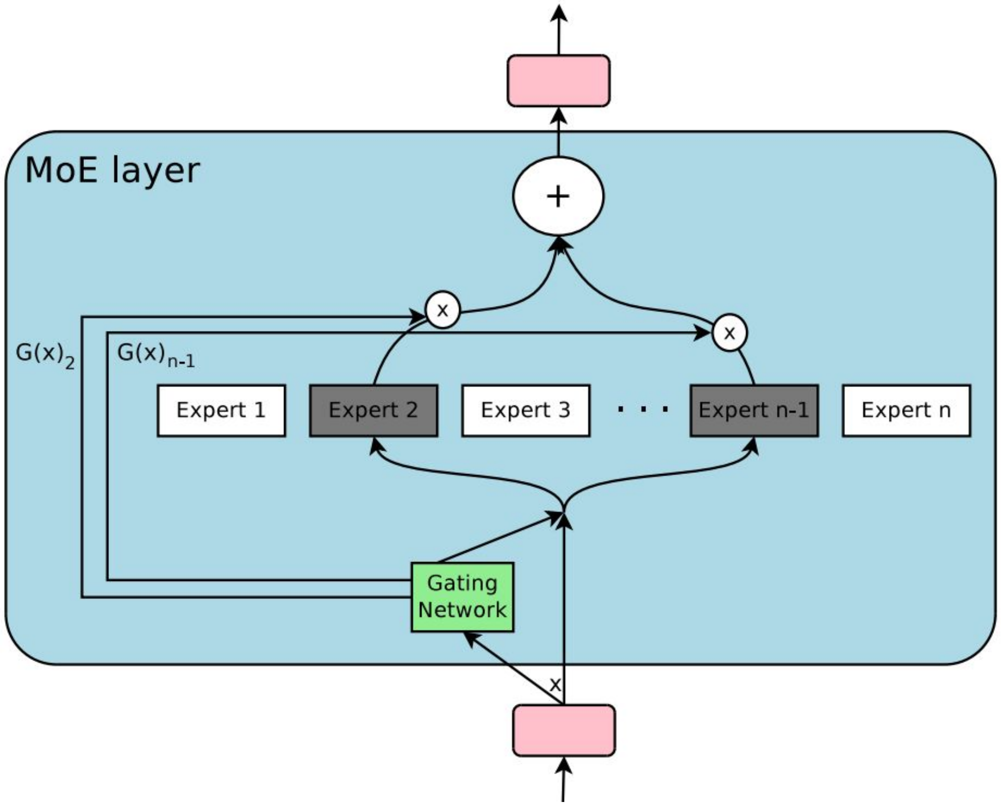
2x800 tanh network on MNIST

Alternative optimization algorithms

- Direct Feedback Alignment
- Target propagation
- Forward-forward
- …

Adaptive architectures

Mixture of experts



→ More weights for free!

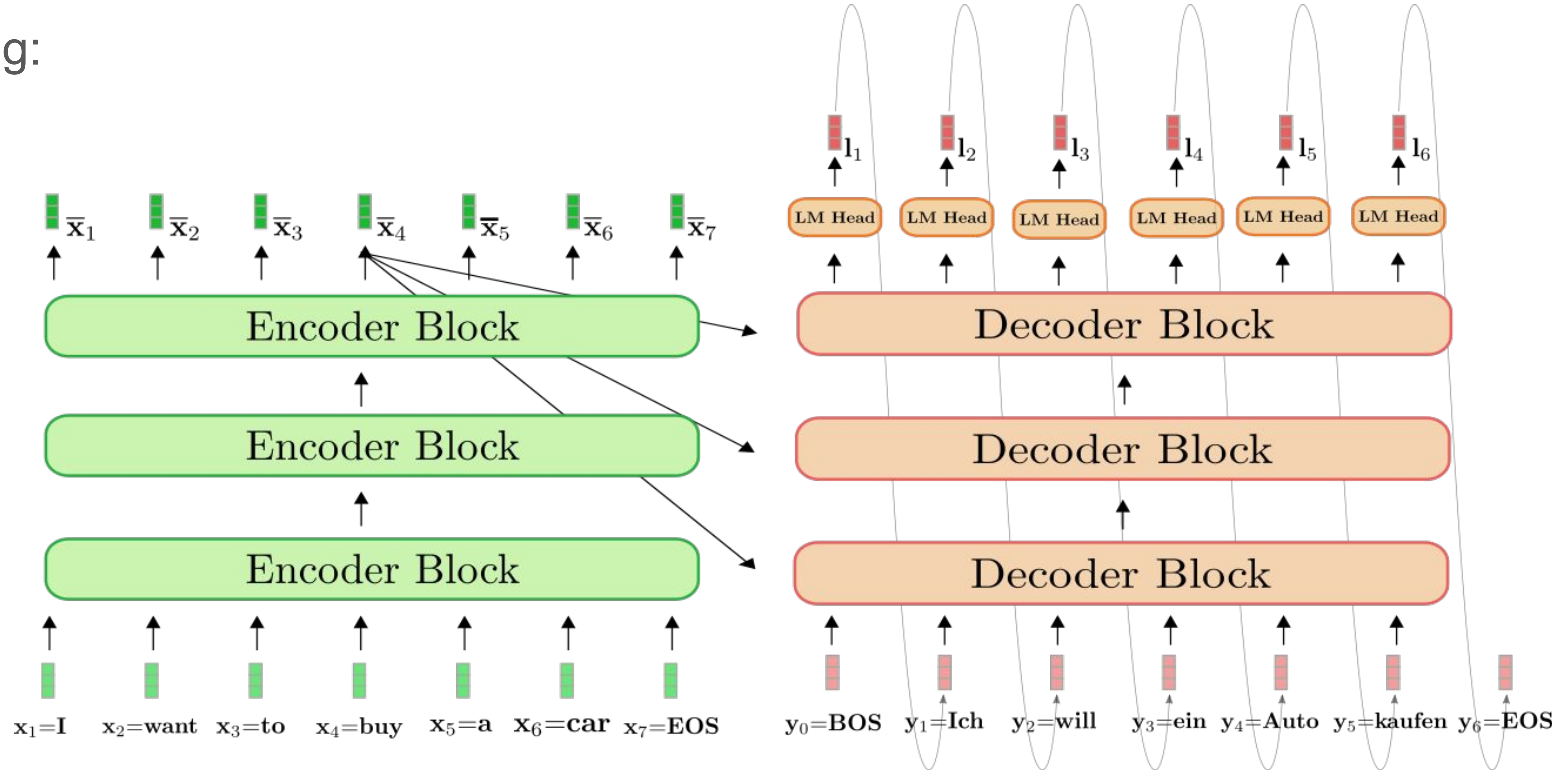
Adaptive architectures

Confident Adaptive Language Modeling (CALM)

Transformer overview

https://huggingface.co/blog/encoder-decoder

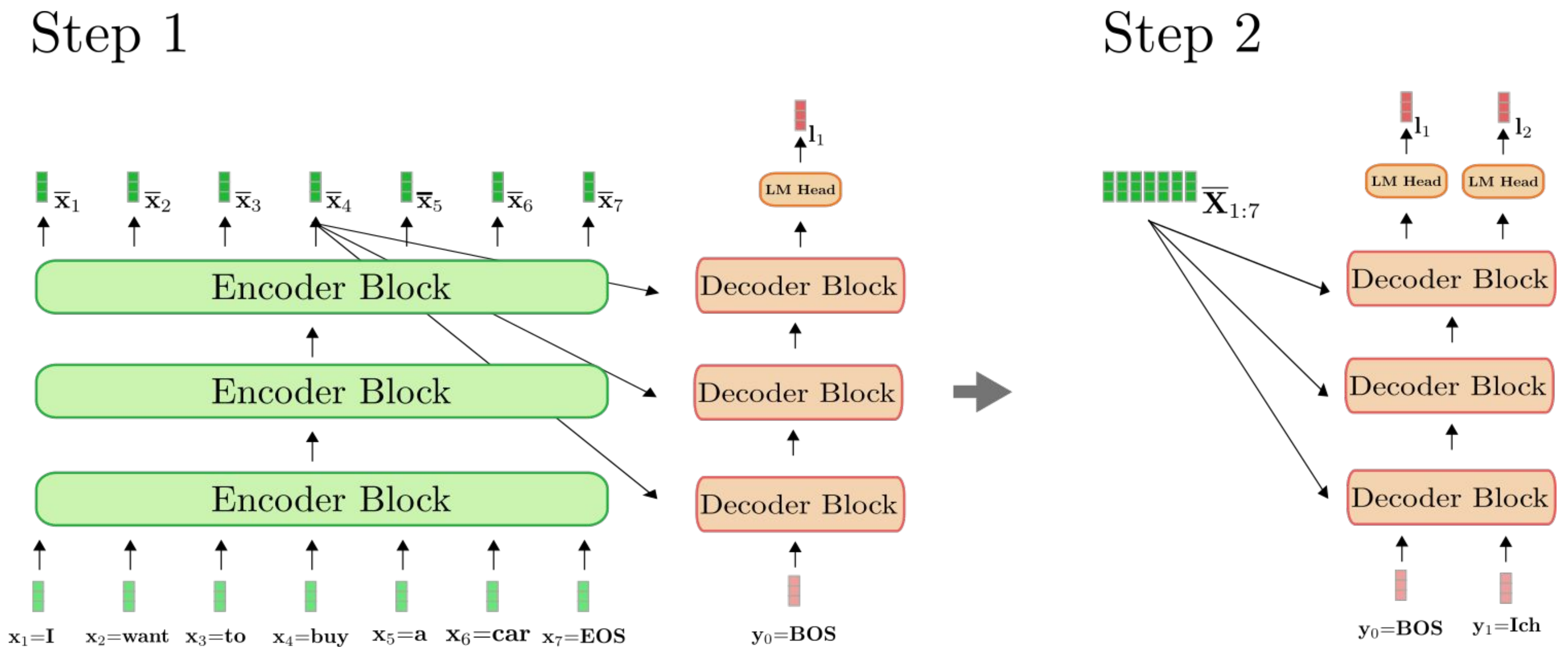
Training:



Transformer overview

https://huggingface.co/blog/encoder-decoder

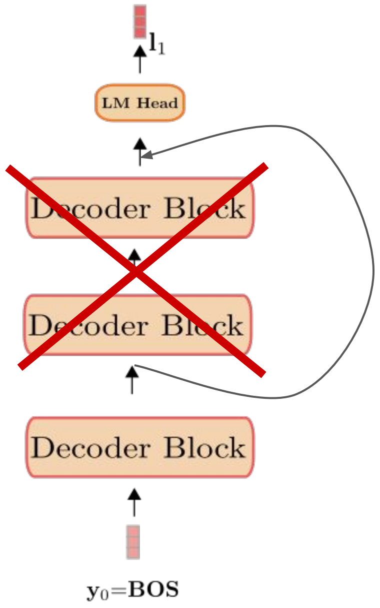
Generating text:



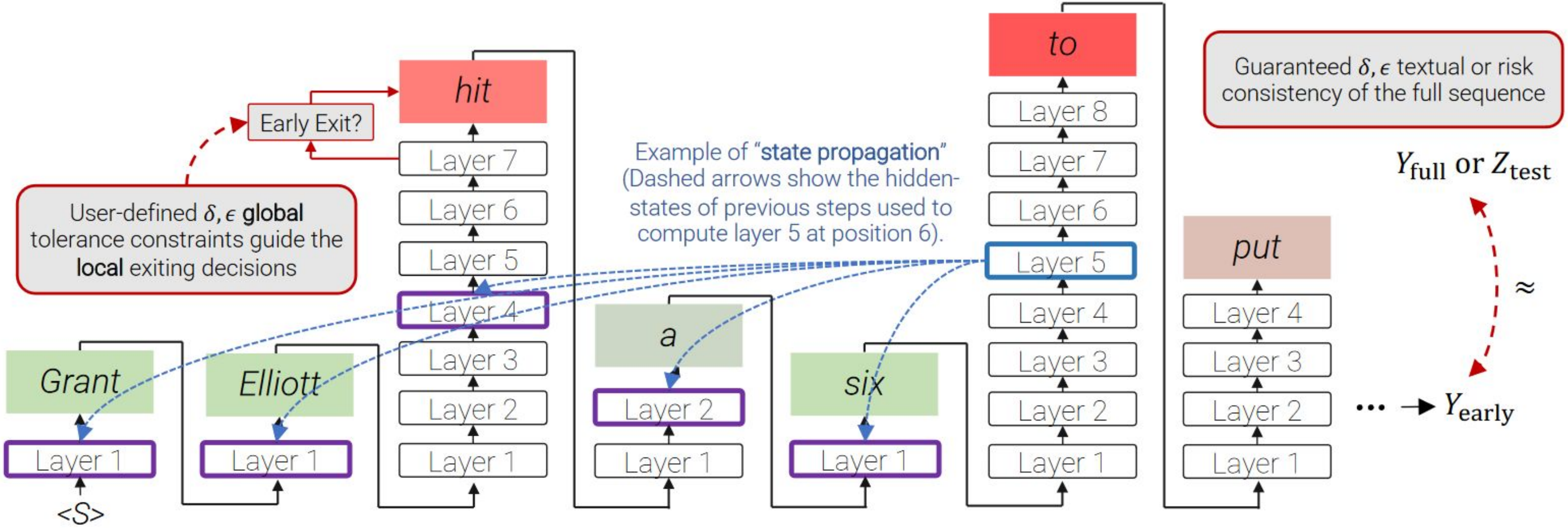
Early exit

To predict the next token :

- stop at layer *i* to make the prediction
- *i* depends on the input (adaptive)
- decision taken by another small neural network



CALM overview

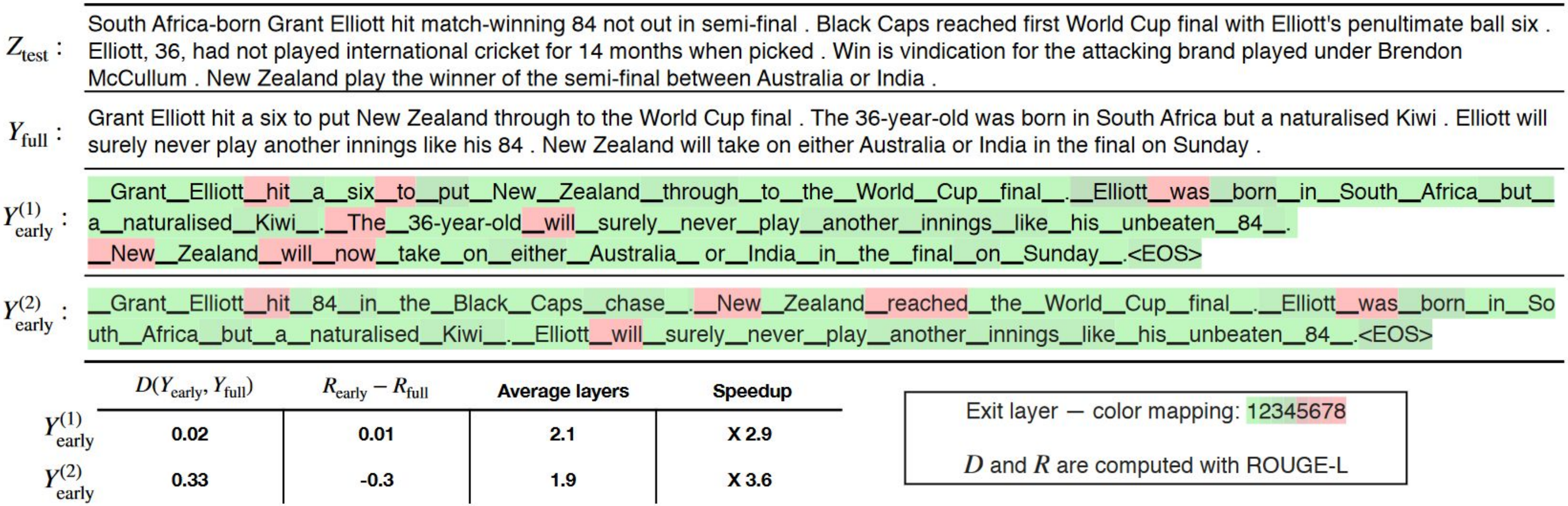


User-defined δ, ϵ **global** tolerance constraints guide the **local** exiting decisions

Example of "**state propagation**" (Dashed arrows show the hidden-states of previous steps used to compute layer 5 at position 6).

Guaranteed δ, ϵ textual or risk consistency of the full sequence

Results



Z_{test} : South Africa-born Grant Elliott hit match-winning 84 not out in semi-final . Black Caps reached first World Cup final with Elliott's penultimate ball six . Elliott, 36, had not played international cricket for 14 months when picked . Win is vindication for the attacking brand played under Brendon McCullum . New Zealand play the winner of the semi-final between Australia or India .

Y_{full} : Grant Elliott hit a six to put New Zealand through to the World Cup final . The 36-year-old was born in South Africa but a naturalised Kiwi . Elliott will surely never play another innings like his 84 . New Zealand will take on either Australia or India in the final on Sunday .

$Y^{(1)}_{\text{early}}$: __Grant__Elliott__hit__a__six__to__put__New__Zealand__through__to__the__World__Cup__final__.__Elliott__was__born__in__South__Africa__but__a__naturalised__Kiwi__.__The__36-year-old__will__surely__never__play__another__innings__like__his__unbeaten__84__.__New__Zealand__will__now__take__on__either__Australia__ or__India__in__the__final__on__Sunday__.<EOS>

$Y^{(2)}_{\text{early}}$: __Grant__Elliott__hit__84__in__the__Black__Caps__chase__.__New__Zealand__reached__the__World__Cup__final__.__Elliott__was__born__in__South__Africa__but__a__naturalised__Kiwi__.__Elliott__will__surely__never__play__another__innings__like__his__unbeaten__84__.<EOS>

	$D(Y_{\text{early}}, Y_{\text{full}})$	$R_{\text{early}} - R_{\text{full}}$	**Average layers**	**Speedup**
$Y^{(1)}_{\text{early}}$	**0.02**	**0.01**	**2.1**	**X 2.9**
$Y^{(2)}_{\text{early}}$	**0.33**	**-0.3**	**1.9**	**X 3.6**

Exit layer — color mapping: 12345678

D and R are computed with ROUGE-L

Challenges of dynamic architectures

- Trying to learn a non-differentiable decision function
- Not always possible to use during training
- Hard to parallelize in a batch
- Uneven weight updates

Thank you!

Any question?