An Incremental Encoder For Sequence-to-Sequence Modelling

Dennis Ulmer[1], Dieuwke Hupkes[1,2], Elia Bruni[1,2]

[1] Universiteit van Amsterdam

[2] Institute for Language, Logic and Computation

Contents

What is incrementality?

(and why do we care)

What is incrementality?

- "In incremental processing, representations are built up as rapidly as possible as the input is encountered" (Christiansen and Chater, 2016)

What is incrementality?

- “In incremental processing, representations are built up as rapidly as possible as the input is encountered” (Christiansen and Chater, 2016)

→ “Incrementally integrating information”

Why do we care?

- “Now-or-Never bottleneck” seems fundamental to *Human* Language Processing (Christiansen and Chater, 2016)

Why do we care?

- "Now-or-Never bottleneck" seems fundamental to *Human* Language Processing (Christiansen and Chater, 2016)
- Incrementality is closely related to *Compositionality*, a possible milestone to human-level intelligence (Lake et al., 2017)

Why do we care?

- “Now-or-Never bottleneck” seems fundamental to *Human* Language Processing (Christiansen and Chater, 2016)
- Incrementality is closely related to *Compositionality*, a possible milestone to human-level intelligence (Lake et al., 2017)
- Attention-based models are not biologically plausible; don’t give incentive to encode efficiently

Contents

Metrics

How to measure incrementality?

1. How well / long is past information stored?

Metrics

How to measure incrementality?

1. How well / long is past information stored?
2. How much new information is integrated?

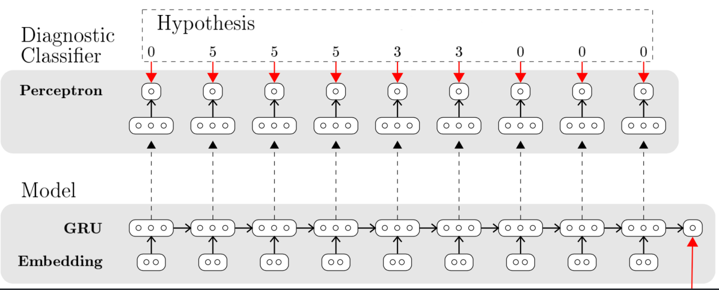
Metrics

How to measure incrementality?

1. How well / long is past information stored?
2. How much new information is integrated?
3. Do representations for the same type resemble each other?

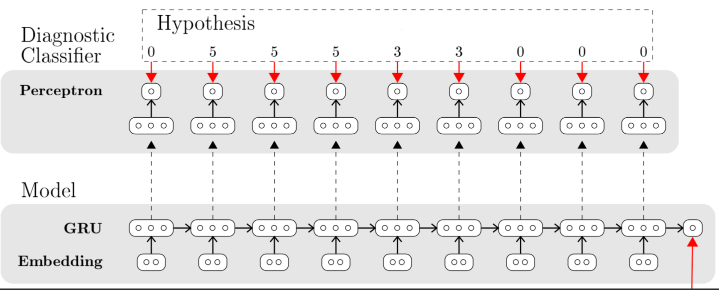
1. How well / long is past information stored?

- Use *Diagnostic Classifiers* (Hupkes et al., 2018)



1. How well / long is past information stored?

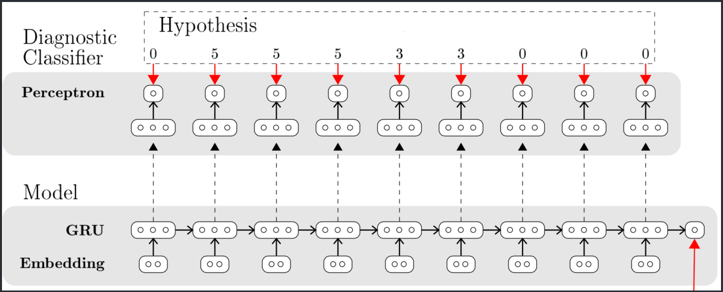
- Use *Diagnostic Classifiers* (Hupkes et al., 2018)
  - Use hidden activations as input for a simple Perceptron, try to predict some information



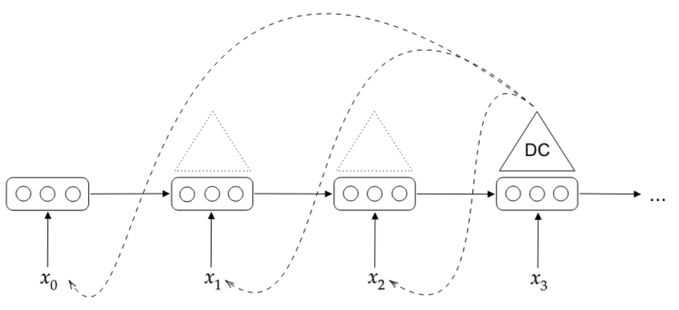
1. How well / long is past information stored?

- Use *Diagnostic Classifiers* (Hupkes et al., 2018)
  - Use hidden activations as input for a simple Perceptron, try to predict some information

  → High Accuracy = Information is likely being stored

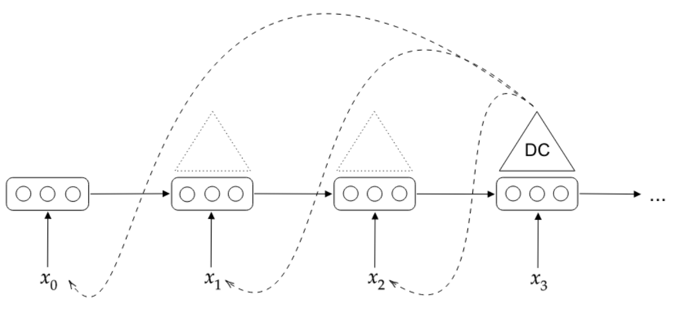


1. How well / long is past information stored?



- Average Diagnostic Classifier Accuracy
  - Use activations h_t to predict the occurrence of token $x_{t'}$

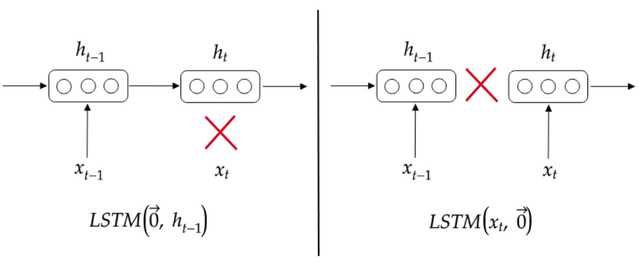
1. How well / long is past information stored?



- Average Diagnostic Classifier Accuracy
  - Use activations h_t to predict the occurrence of token $x_{t'}$
- Weighed Average Diagnostic Classifier Accuracy
  - Weigh by distance $t - t'$

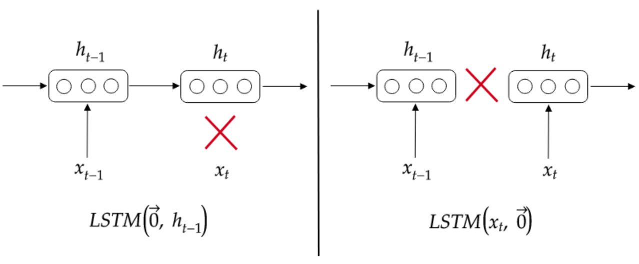
2. How much new information is integrated?

- If no new information is integrated, then $||h_t - \mathrm{LSTM}(\vec{0}, h_{t-1})|| = 0$



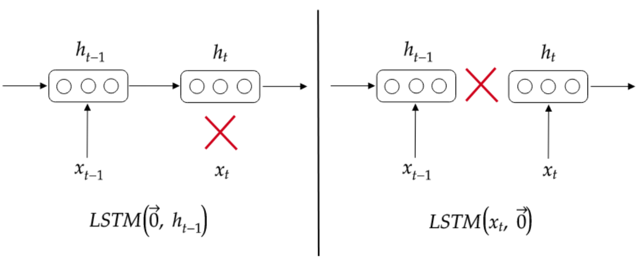
2. How much new information is integrated?

- If no new information is integrated, then $||h_t - \text{LSTM}(\vec{0}, h_{t-1})|| = 0$
- If the whole history is forgotten, then $||h_t - \text{LSTM}(x_t, \vec{0})|| = 0$



2. How much new information is integrated?

- If no new information is integrated, then $||h_t - \text{LSTM}(\vec{0}, h_{t-1})|| = 0$
- If the whole history is forgotten, then $||h_t - \text{LSTM}(x_t, \vec{0})|| = 0$
- Hypothesis: Incremental model finds trade-off between old and new information



2. How much new information is integrated?

→ Integration Ratio

$$\phi = \frac{||h_t - \mathrm{LSTM}(\vec{0}, h_{t-1})||}{||h_t - \mathrm{LSTM}(x_t, \vec{0})||} \tag{1}$$

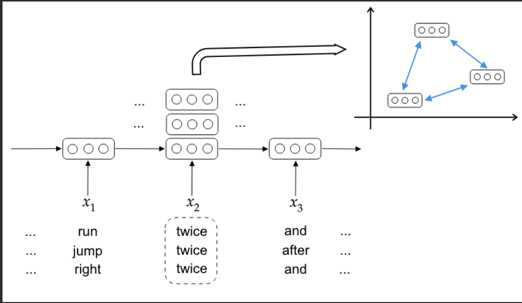
$$\mathrm{intratio}(\{x\}_0^T) = \frac{1}{T}\sum_{t=1}^{T} \min\left(\phi, \phi^{-1}\right) \tag{2}$$

3. Do representations for the same type resemble each other?

- We expect representations to be **close** in hidden space when encoding the **same type**

3. Do representations for the same type resemble each other?

- We expect representations to be **close** in hidden space when encoding the **same type**
- Representational similarity
  - Calculate average euclidean distance between h_t for same x_t



Contents

Standard models

- Vanilla LSTM
- LSTM with dot-product attention

An incremental (?) model

- Part of what enables fast human processing seems to be the ability to anticipate future utterances (Christiansen and Chater, 2016)

An incremental (?) model

- Part of what enables fast human processing seems to be the ability to anticipate future utterances (Christiansen and Chater, 2016)

→ Add a secondary Anticipation Loss to training
  - Project hidden representations into vocabulary space, use cross-entropy loss to compare with actual next token

Contents

Dataset

- Use the SCAN (Lake and Baroni, 2018) data set
  - Translate commands in Natural Language into sequence of commands

    jump thrice and look ⟶ `I_JUMP I_JUMP I_JUMP I_LOOK`

  - Designed to test compositionality

Setup

- Train 15 models per class to account for variance

Setup

- Train 15 models per class to account for variance
- Compute metric scores for every model

Setup

- Train 15 models per class to account for variance
- Compute metric scores for every model
  - Are they measuring the same?

Setup

- Train 15 models per class to account for variance
- Compute metric scores for every model
  - Are they measuring the same?
  - Do high scores imply better performance?

Setup

- Train 15 models per class to account for variance
- Compute metric scores for every model
  - Are they measuring the same?
  - Do high scores imply better performance?

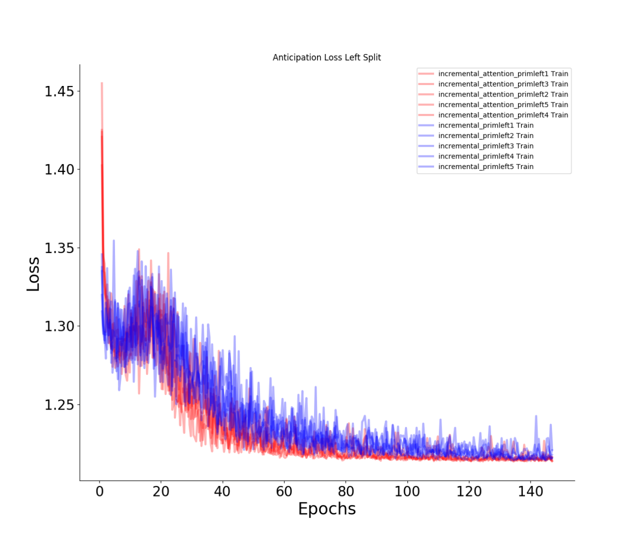
  $\longrightarrow$ Measure correlation!

Results

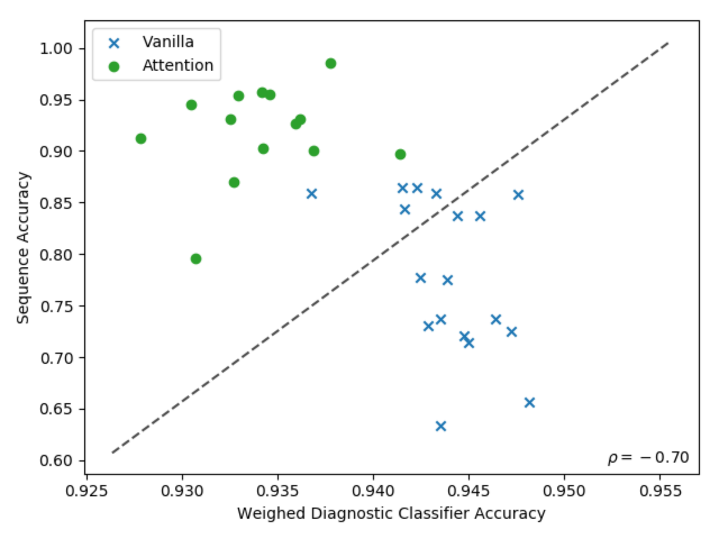
Model	seq_acc	dc_acc	wdc_acc	intratio	repsim
BL	$.765 \pm .07$	$.958 \pm 0$	$.945 \pm 0$	$\mathbf{.714 \pm .02}$	$4.399 \pm .08$
BL + Attn.	$\mathbf{.919 \pm .05}$	$.950 \pm 0$	$.935 \pm 0$	$.697 \pm .01$	$\mathbf{3.859 \pm .08}$
Antcp. Loss	$.661 \pm .23$	$.957 \pm 0$	$.943 \pm 0$	$.664 \pm .02$	$\mathbf{3.834 \pm .11}$

Figure: Results on SCAN `add_prim_left` with $n = 15$.

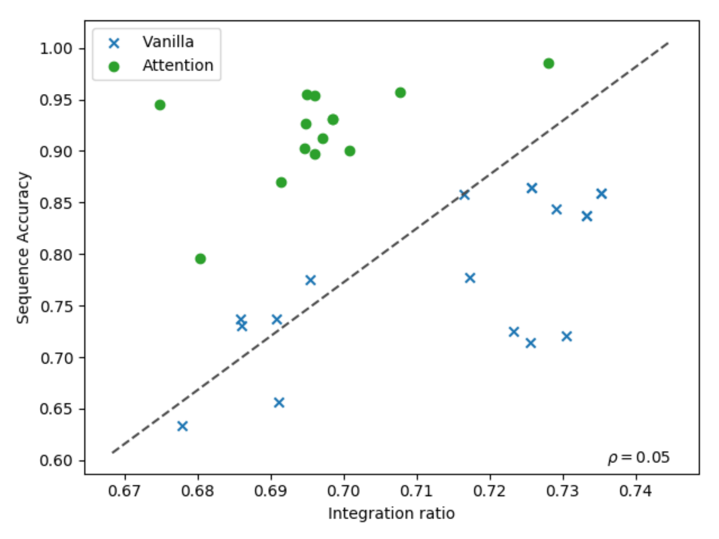
Anticipation Loss insights



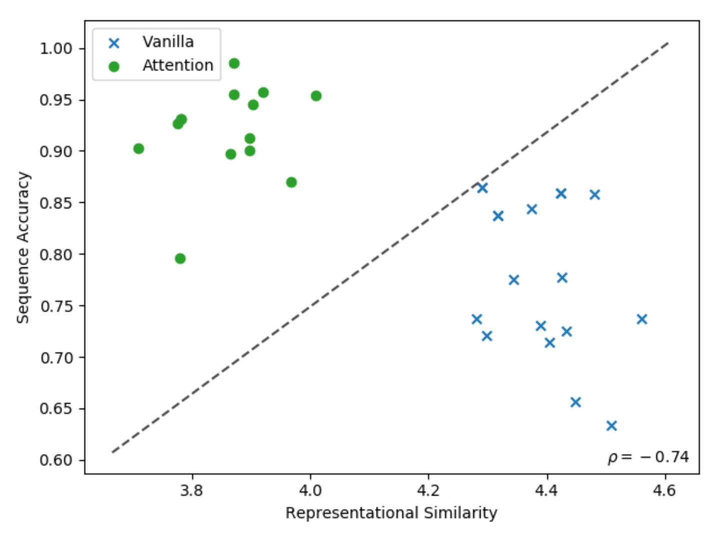
Metrics Insights I



Metrics Insights II



Metrics Insights III



Metrics Insights IV

	seq_acc	dc_acc	wdc_acc	intratio	repsim
seq_acc	1	−0.80	−0.79	0.05	−0.74
dc_acc		1	0.78	0.29	**0.81**
wdc_acc			1	0.40	**0.80**
intratio				1	0.39
repsim					1

Figure: Correlation between metrics measured with Pearson's ρ.

Contents

Open Questions

- Do the metrics actually measure incrementality? What would be a better metric?

Open Questions

- Do the metrics actually measure incrementality? What would be a better metric?
- What would an incremental model look like?

Open Questions

- Do the metrics actually measure incrementality? What would be a better metric?
- What would an incremental model look like?
- How to realize insights about human cognition in model architectures?

Future Work

- Improving metrics

Future Work

- Improving metrics
- Models based on *Chunk-and-Pass processing*

Future Work

- Improving metrics
- Models based on *Chunk-and-Pass processing*
- Your ideas?

Future Work

- Improving metrics
- Models based on *Chunk-and-Pass processing*
- Your ideas?
- Master thesis: Activation interventions

Contents

Bibliography

Morten H Christiansen and Nick Chater. The now-or-never bottleneck: A fundamental constraint on language. Behavioral and Brain Sciences, 39, 2016.

Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. Visualisation and'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. Journal of Artificial Intelligence Research, 61:907–926, 2018.

Brenden Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In International Conference on Machine Learning, pages 2879–2888, 2018.

Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. Behavioral and Brain Sciences, 40, 2017.