Paper Reading: The Importance of Being Recurrent for Modeling Hierarchical Structure

Ke Tran, Arianna Bisazza & Christof Monz

University of Amsterdam & Leiden University

Motivation

- Compare two architectures-recurrent versus non-recurrent.
- Ability of modeling hierarchical structure.

LSTM v.s. Fully Attentional Network (FAN)

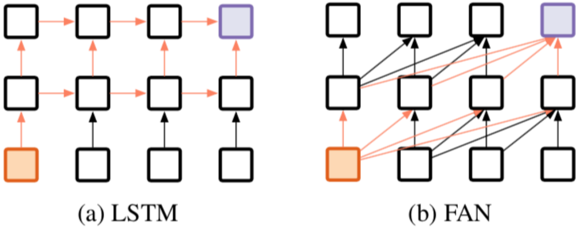


Figure 1: Diagram showing the main difference between a LSTM and a FAN. Purple boxes indicate the summarized vector at current time step t which is used to make prediction. Orange arrows indicate the information flow from a previous input to that vector.

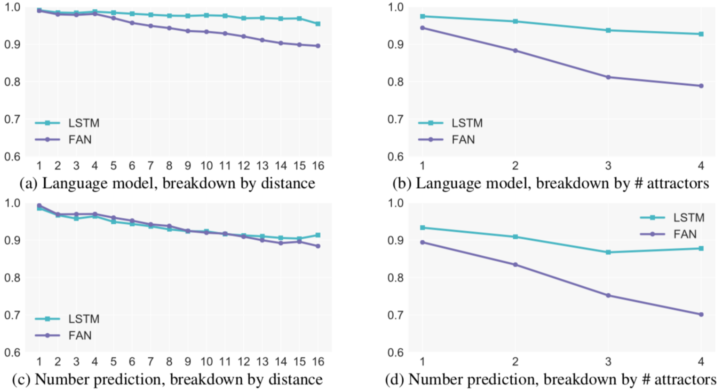
Task1: Subject-Verb Agreement

- (a) A general language model. i.e. Next word prediction objective.
- (b) An explicit supervision objective. i.e. Predicting the number of the verb given its sentence history.

Table 1: Examples of training and test conditions for the two subject-verb agreement subtasks. The full input sentence is "The **keys** to the cabinet **are** on the table" where verb and subject are bold and intervening nouns are underlined.

	Input	Train	Test
(a)	the keys to the cabinet	are	$p(\text{are}) > p(\text{is})$?
(b)	the keys to the cabinet	plural	plural/singular?

Task1 results



(a) Language model, breakdown by distance

(b) Language model, breakdown by # attractors

(c) Number prediction, breakdown by distance

(d) Number prediction, breakdown by # attractors

Figure 2: Results of subject-verb agreement with different training objectives.

Task2: Logical inference

- The vocabulary of this language includes six word typesa,b,c,d,e,f, and three logical operatorsor, and, not. There are also 7 logical relations.
- Reason for use the task: to correctly classify logical relations, the model must learn nested structures as well as the scope of logical operators.

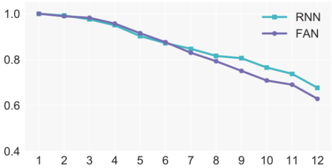
$$(\text{d} (\text{or f})) \sqsupset (\text{f} (\text{and a}))$$
$$(\text{d} (\text{and} (\text{c} (\text{or d})))) \ \# \ (\text{not f})$$
$$(\text{not} (\text{d} (\text{or} (\text{f} (\text{or c}))))) \sqsubset (\text{not} (\text{c} (\text{and} (\text{not d}))))$$

Model

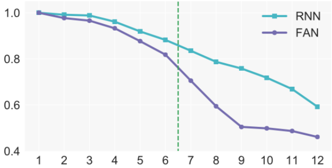
The LSTM architecture used in this experiment is similar to that of Bowman et al. (2015b). We simply take the last hidden state of the top LSTM layer as a fixed-size vector representation of the sentence. Here, we use a 2-layer LSTM with skip connections. The FAN maps a sentence x of length n to $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_n] \in \mathbb{R}^{d \times n}$. To obtain a fixed-size representation $\mathbf{z}$, we use a self-attention layer with two trainable queries $\mathbf{q}_1, \mathbf{q}_2 \in \mathbb{R}^{1 \times d}$:

$$
\begin{aligned}
\mathbf{z}_i &= \operatorname{softmax}\left(\frac{\mathbf{q}_i \mathbf{H}}{\sqrt{d}}\right) \mathbf{H}^\top \quad i \in \{1, 2\} \\
\mathbf{z} &= [\mathbf{z}_1, \mathbf{z}_2]
\end{aligned}
$$

Task2 results



(a) $n \leq 12$



(b) $n \leq 6$

Figure 4: Results of logical inference when training on all data (a) or only on samples with at most n logical operators (b).

Conclusions

- Recurrence is important for modeling hierarchical structure.
- The paper provides a way to do model analysis.

Paper Reading: How Much Attention Do You Need? A Granular Analysis of Neural Machine Translation Architectures

Tobias Domhan

Amazon

Motivation

- A granular analysis of neural machine translation architectures.
- In Transformer, how much each of components matters.

Architecture Definition Language

- Layers, Layer definitions, Layer chaining, Encoder/Decoder structure, Layer repetition.

Layer Definitions

Dropout, Fixed positional embeddings, Linear, Feed-Forward, Convolution, Identity, Concatenation, Recurrent Neural Network, Attention, Layer normalization, Residual layer.

What to attend to?

Encoder block	IWSLT	WMT'17
upper	25.4 ± 0.2	27.6 ± 0.0
increasing	25.4 ± 0.1	27.3 ± 0.1
decreasing	25.3 ± 0.2	27.1 ± 0.1

Table 2: BLEU scores when varying the encoder block used in the source attention mechanism of a Transformer on the EN$\rightarrow$DE IWSLT and WMT'17 datasets.

Network Structure

First,
Encoder: dropout → res_d(birnn) → repeat(5, res_d(rnn))
Decoder: dropout → repeat(6, res_d(rnn)) → res_d(dot_src_att) → res_d(ffl)

Then,
Encoder: pos → res_nd(birnn) → res_nd(ffl) → repeat(5, res_nd(rnn) → res_nd(ffl)) → norm
Decoder: pos → repeat(6, res_nd(rnn) → res_nd(mh_dot_src_att) → res_nd(ffl)) → norm

Transforming an RNN into a Transformer style architecture

Model	IWSLT EN→DE	WMT'17 EN→DE		WMT'17 LV→EN	
	BLEU	BLEU	METEOR	BLEU	METEOR
Transformer	25.4 ± 0.1	27.6 ± 0.0	47.2 ± 0.1	18.5 ± 0.0	51.3 ± 0.1
RNMT	23.2 ± 0.2	25.5 ± 0.2	45.1 ± 0.1	-	-
- input feeding	23.1 ± 0.2	24.6 ± 0.1	43.8 ± 0.2	-	-
RNN	22.8 ± 0.2	23.8 ± 0.1	43.3 ± 0.1	15.2 ± 0.1	45.9 ± 0.1
+ mh	23.7 ± 0.4	24.4 ± 0.1	43.9 ± 0.1	16.0 ± 0.1	47.1 ± 0.1
+ pos	23.9 ± 0.2	24.1 ± 0.1	43.5 ± 0.2	-	-
+ norm	23.7 ± 0.1	24.0 ± 0.2	43.2 ± 0.1	15.2 ± 0.1	46.3 ± 0.2
+ multi-att-1h	24.5 ± 0.0	25.2 ± 0.1	44.9 ± 0.1	16.6 ± 0.2	49.1 ± 0.2
/ multi-att	24.4 ± 0.3	25.5 ± 0.0	45.3 ± 0.0	17.0 ± 0.2	49.4 ± 0.1
+ ff	25.1 ± 0.1	26.7 ± 0.1	46.4 ± 0.2	17.8 ± 0.1	50.5 ± 0.1

Table 3: Transforming an RNN into a Transformer style architecture. + shows the incrementally added variation. / denotes an alternative variation to which the subsequent + is relative to.

Transforming a CNN based model into a Transformer style architecture

	IWSLT EN-DE	WMT'17 EN→DE		WMT'17 LV→EN	
Model	BLEU	BLEU	METEOR	BLEU	METEOR
Transformer	25.4 ± 0.1	27.6 ± 0.0	47.2 ± 0.1	18.5 ± 0.0	51.3 ± 0.1
CNN GLU	24.3 ± 0.4	25.0 ± 0.3	44.4 ± 0.2	16.0 ± 0.5	47.4 ± 0.4
+ norm	24.1 ± 0.1	-	-	-	-
+ mh	24.2 ± 0.2	25.4 ± 0.1	44.8 ± 0.1	16.1 ± 0.1	47.6 ± 0.2
+ ff	25.3 ± 0.1	26.8 ± 0.1	46.0 ± 0.1	16.4 ± 0.2	47.9 ± 0.2
CNN ReLU	23.6 ± 0.3	23.9 ± 0.1	43.4 ± 0.1	15.4 ± 0.1	46.4 ± 0.3
+ norm	24.3 ± 0.1	24.3 ± 0.2	43.6 ± 0.1	16.0 ± 0.2	47.1 ± 0.5
+ mh	24.2 ± 0.2	24.9 ± 0.1	44.4 ± 0.1	16.1 ± 0.1	47.5 ± 0.2
+ ff	25.3 ± 0.3	26.9 ± 0.1	46.1 ± 0.0	16.4 ± 0.2	47.9 ± 0.1

Table 4: Transforming a CNN based model into a Transformer style architecture.

Self-attention variations

		IWSLT EN→DE	WMT'17 EN→DE		WMT'17 LV→EN	
Encoder	Decoder	BLEU	BLEU	METEOR	BLEU	METEOR
self-att	self-att	25.4 ± 0.2	27.6 ± 0.0	47.2 ± 0.1	18.3 ± 0.0	51.1 ± 0.1
self-att	RNN	25.1 ± 0.1	27.4 ± 0.1	47.0 ± 0.1	18.4 ± 0.2	51.1 ± 0.1
self-att	CNN	25.4 ± 0.4	27.6 ± 0.2	46.7 ± 0.1	18.0 ± 0.3	50.3 ± 0.3
RNN	self-att	25.8 ± 0.1	27.2 ± 0.1	46.7 ± 0.1	17.8 ± 0.1	50.6 ± 0.1
CNN	self-att	25.7 ± 0.1	26.6 ± 0.3	46.3 ± 0.1	16.8 ± 0.4	49.4 ± 0.4
RNN	RNN	25.1 ± 0.1	26.7 ± 0.1	46.4 ± 0.2	17.8 ± 0.1	50.5 ± 0.1
CNN	CNN	25.3 ± 0.3	26.9 ± 0.1	46.1 ± 0.0	16.4 ± 0.2	47.9 ± 0.2
self-att	*combined*	25.1 ± 0.2	27.6 ± 0.2	47.2 ± 0.2	18.3 ± 0.1	51.1 ± 0.1
self-att	*none*	23.7 ± 0.2	25.3 ± 0.2	43.1 ± 0.1	15.9 ± 0.1	45.1 ± 0.2

Table 5: Different variations of the encoder and decoder self-attention layer.

Conclusion

- Source attention on lower encoder layers brings no additional benefit.
- Multiple source attention layers and residual feed-forward layers are key.
- Self-attention is more important for the source than for the target side.