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

Tobias Domhan domhant@amazon.de

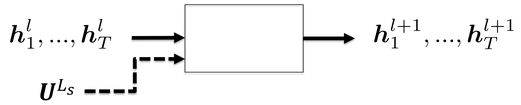
Amazon Machine Learning
Development Center Germany

Introduction

- **Architecture Definition Language**
- Allows easy experimentation of **Neural Machine Translation (NMT) architecture variations**
- Detailed analysis of NMT architecture combinations
- **Multiple source attention layers and residual feed-forward layer are key**
- **Self-attention more important on the source side**

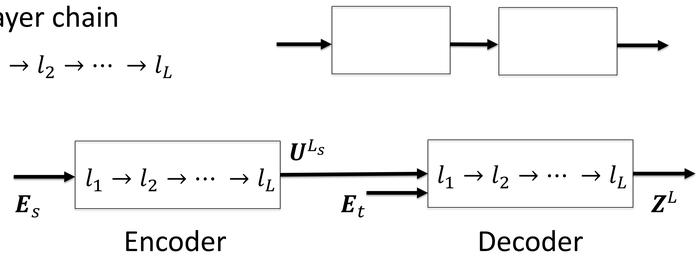
Architecture Definition Language (ADL)

Layer l



Layer chain

$l_1 \rightarrow l_2 \rightarrow \cdots \rightarrow l_L$



Layer	Description
pos	Positional embeddings.
$norm,\ bnorm$	Layer and batch normalization.
$repeat(n, l)$	Layer repetition.
$res(l)$	Residual connection.
$birnn, rnn$	Recurrent neural network.
cnn	Convolutional neural network.
$ff, linear, fff$	Feed-forward and linear layer.
$[mh]\_dot\_src\_att(h)$	[Multi-head] dot source attention.
$[mh]\_dot\_self\_att(h)$	[Multi-head] dot self-attention.
$res\_nd(l)$	$res(norm \rightarrow l \rightarrow dropout)$
$res\_d(l)$	$res(l \rightarrow dropout)$
$act, dropout$ $parallel, highway$	Other layers.

NMT Architectures defined by ADL

e.g. Transformer [Vaswani et al. 2017]

$$t_{\text{enc}} = res\_nd(mh\_dot\_self\_att) \rightarrow res\_nd(fff)$$

$$t_{\text{dec}} = res\_nd(mh\_dot\_self\_att) \rightarrow res\_nd(mh\_dot\_src\_att) \rightarrow res\_nd(fff)$$

$$\mathbf{U}^{L_s} = pos \rightarrow repeat(n, t_{\text{enc}}) \rightarrow norm$$

$$\mathbf{Z}^{L} = pos \rightarrow repeat(n, t_{\text{dec}}) \rightarrow norm$$

Usage in Sockeye, our NMT toolkit:

```
> sockeye-train -s train.de -t train.en
  --custom-seq-encoder repeat(6,birnn)
  --custom-seq-decoder pos->res(norm->mh_dot_self_att)->res(mh_dot_src_att)
```

https://github.com/awslabs/sockeye/tree/acl18

Encoder/Decoder Combinations

- Baseline: 6 layer Transformer, 512 hidden units
- Data sets: IWSLT'16, WMT'17
- Metrics: BLEU (and METEOR in the paper)
- 3 runs, reporting mean and standard deviation

$t_{\text{enc}} = res\_nd(\quad) \rightarrow res\_nd(fff)$

$t_{\text{dec}} = res\_nd(\quad) \rightarrow res\_nd(mh\_dot\_src\_att) \rightarrow res\_nd(fff)$

Encoder	Decoder	IWSLT EN→DE	WMT'17 EN→DE	WMT'17 LV→EN
self-att	self-att	25.4 ± 0.2	27.6 ± 0.0	18.3 ± 0.0
self-att	RNN	25.1 ± 0.1	27.4 ± 0.1	18.4 ± 0.2
self-att	CNN	25.4 ± 0.4	27.6 ± 0.2	18.0 ± 0.3
RNN	self-att	25.8 ± 0.1	27.2 ± 0.1	17.8 ± 0.1
CNN	self-att	25.7 ± 0.1	26.6 ± 0.3	16.8 ± 0.4
RNN	RNN	25.1 ± 0.1	26.7 ± 0.1	17.8 ± 0.1
CNN	CNN	25.3 ± 0.3	26.9 ± 0.1	16.4 ± 0.2
self-att	*combined*	25.1 ± 0.2	27.6 ± 0.2	18.3 ± 0.1
self-att	*none*	23.7 ± 0.2	25.3 ± 0.2	15.9 ± 0.1

Network Structure Experiments

RNN to Transformer

$$\mathbf{U}^{L_s} = dropout \rightarrow res\_d(birnn) \rightarrow repeat(5, res\_d(rnn))$$

$$\mathbf{Z}^{L} = dropout \rightarrow repeat(6, res\_d(rnn)) \rightarrow res\_d(dot\_src\_att) \rightarrow res\_d(fff)$$

Model	IWSLT EN→DE	WMT'17 EN→DE	WMT'17 LV→EN
Transformer	25.4 ± 0.1	27.6 ± 0.0	18.5 ± 0.0
RNMT	23.2 ± 0.2	25.5 ± 0.2	-
- input feeding	23.1 ± 0.2	24.6 ± 0.1	-
RNN	22.8 ± 0.2	23.8 ± 0.1	15.2 ± 0.1
+ mh	23.7 ± 0.4	24.4 ± 0.1	16.0 ± 0.1
+ pos	23.9 ± 0.2	24.1 ± 0.1	15.6 ± 0.1
+ norm	23.7 ± 0.1	24.0 ± 0.2	15.2 ± 0.1
+ multi-att-1h	24.5 ± 0.0	25.2 ± 0.1	16.6 ± 0.2
/ multi-att	24.4 ± 0.3	25.5 ± 0.0	17.0 ± 0.2
+ ff	25.1 ± 0.1	26.7 ± 0.1	17.8 ± 0.1

$$\mathbf{U}^{L_s} = pos \rightarrow res\_nd(birnn) \rightarrow res\_nd(fff) \rightarrow repeat(5, res\_nd(rnn)) \rightarrow res\_nd(fff) \rightarrow norm$$

$$\mathbf{Z}^{L} = pos \rightarrow repeat(6, res\_nd(rnn) \rightarrow res\_nd(mh\_dot\_src\_att) \rightarrow res\_nd(fff)) \rightarrow norm$$

CNN to Transformer

$$\mathbf{U}^{L_s} = pos \rightarrow repeat(6, res\_d(cnn))$$

$$\mathbf{Z}^{L} = pos \rightarrow repeat(6, res\_d(cnn) \rightarrow res\_d(dot\_src\_att))$$

Model	IWSLT EN→DE	WMT'17 EN→DE	WMT'17 LV→EN
Transformer	25.4 ± 0.1	27.6 ± 0.0	18.5 ± 0.0
CNN GLU	24.3 ± 0.4	25.0 ± 0.3	16.0 ± 0.5
+ norm	24.1 ± 0.1	-	16.1 ± 0.2
+ mh	24.2 ± 0.2	25.4 ± 0.1	16.1 ± 0.1
+ ff	25.3 ± 0.1	26.8 ± 0.1	16.4 ± 0.2
CNN ReLU	23.6 ± 0.3	23.9 ± 0.1	15.4 ± 0.1
+ norm	24.3 ± 0.1	24.3 ± 0.2	16.0 ± 0.2
+ mh	24.2 ± 0.2	24.9 ± 0.1	16.1 ± 0.1
+ ff	25.3 ± 0.3	26.9 ± 0.1	16.4 ± 0.2

$$\mathbf{U}^{L_s} = pos \rightarrow repeat(6, res\_nd(cnn) \rightarrow res\_nd(fff)) \rightarrow norm$$

$$\mathbf{Z}^{L} = pos \rightarrow repeat(6, res\_nd(cnn) \rightarrow res\_nd(mh\_dot\_src\_att) \rightarrow res\_nd(fff)) \rightarrow norm$$