Neural Machine Translation In Linear Time

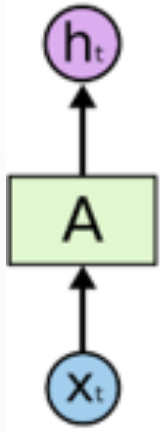
Authors:
Nal Kalchbrenner, Lasse Espeholt,
Karen Simonyan, Aaron van den Oord,
Alex Graves, Koray Kavukcuoglu

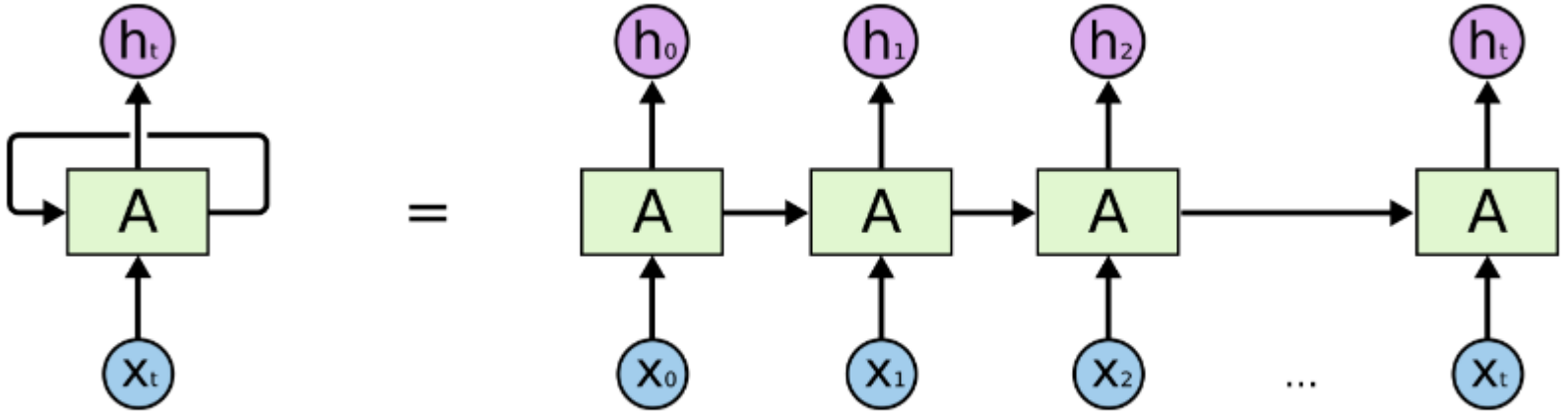
Presenter:
SunMao sm4206
YuZheng yz2978

OVERVIEW

- Brief review of RNN and one-dimension CNN
- ByteNet architecture
- Merits of this model compared with other models
- Techniques used in this model
- Performance comparison

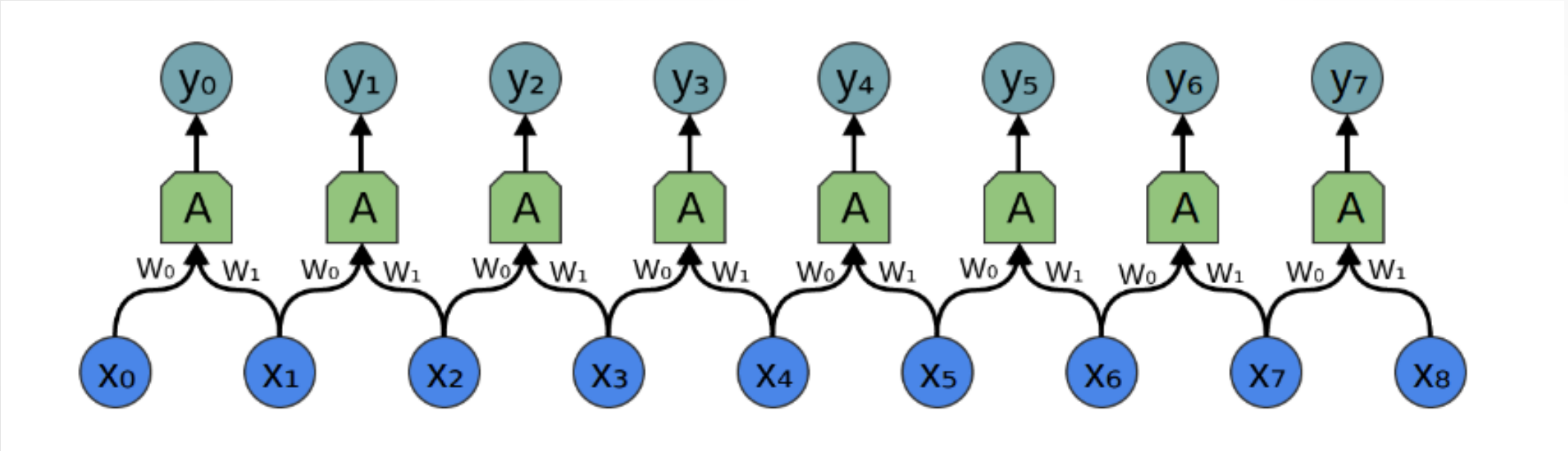
RNN



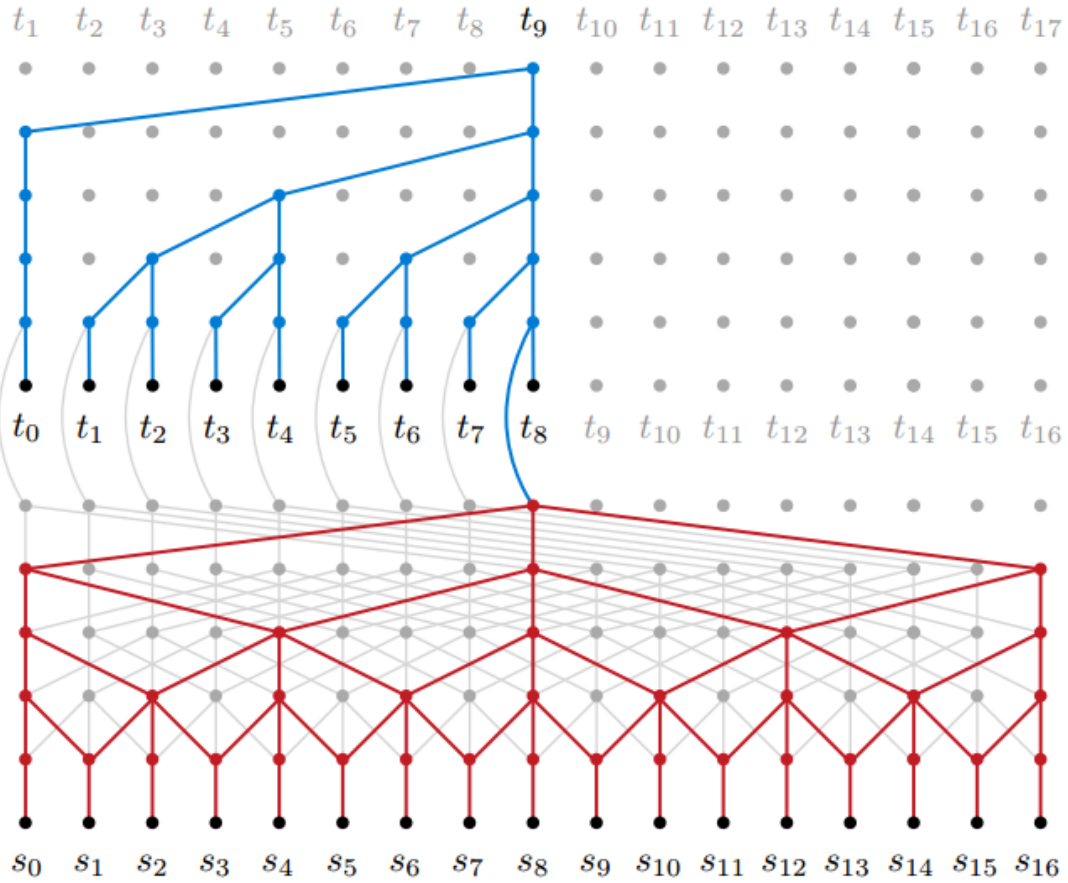


An unrolled recurrent neural network.

ONE – DIMENSION CNN



BYTENET ARCHITECTURE

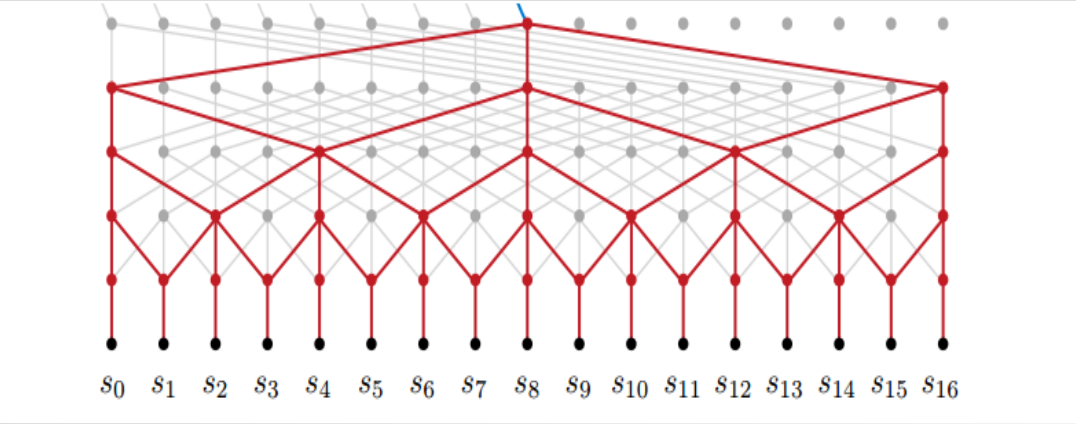


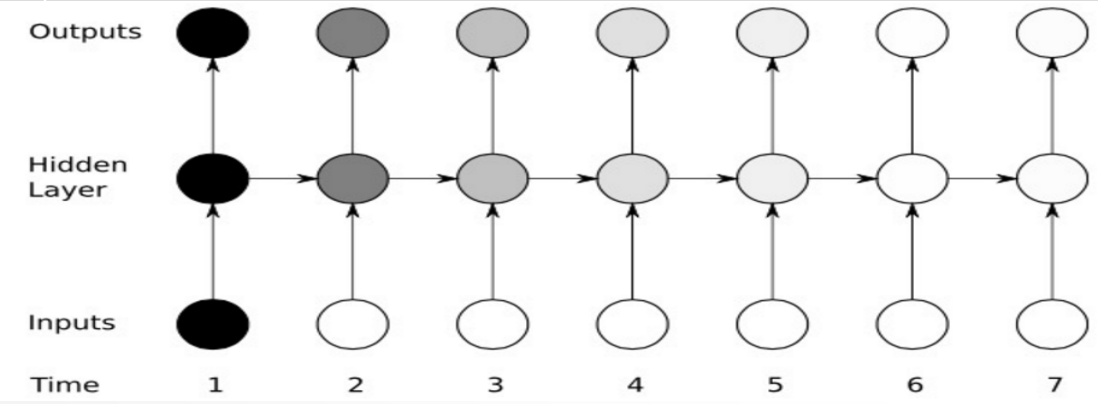
2 sub-networks:

- Source Network: Process the source sequence into a representation (CNN)
- Target Network: Use the representat to generate the target sequence (One novel structure designed to capture a very long range of past inputs)

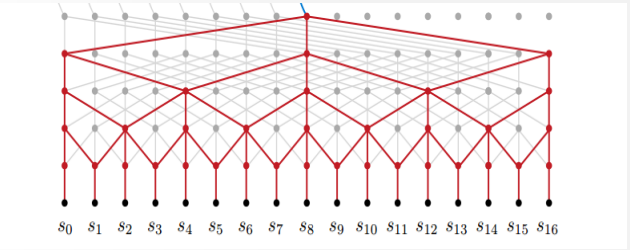
MERITS OF BYTENET USING CNN

ByteNet	RNN
Using CNN, can do parallel computation Running in linear time in the length of the source and target sequence	RNNs have an inherently serial structure that prevents them from being run in parallel along the sequence length.
the distance traversed by forward and backward signals between any input and output tokens in the networks corresponds to the fixed depth of the networks and is largely independent of the distance between the tokens.	Forward and backward signals in a RNN also need to traverse the full distance of the serial path to reach from one point to another in the sequence.
Dependencies over large distances are connected by short paths and can be learnt more easily	Dependencies over long distance is hard to be learnt



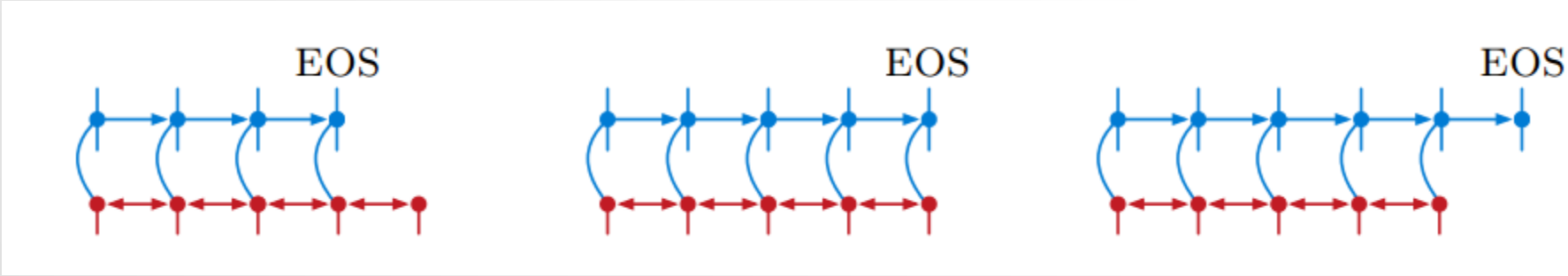


TECHNIQUES USED IN THIS MODEL: DYNAMIC UNFOLDING



- The source network builds a representation that has **the same width as the source sequence**.
- At each step the target network takes as input the corresponding column of The source representation until the target network produces the **end-of-sequence symbol**. (As shown below)
- The source representation is **zero-padded** on the fly:

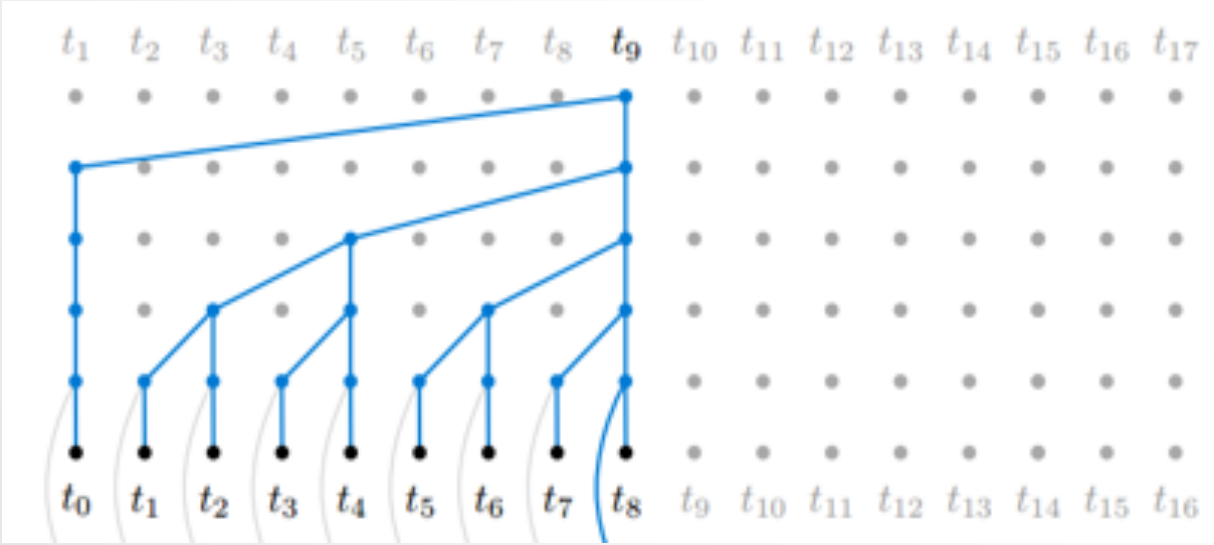
if the target network produces symbols beyond the length of the source sequence, the corresponding conditioning column is set to zero.



TECHNIQUES USED IN THIS MODEL: MASKED ONE-DIMENSION CONVOLUTION

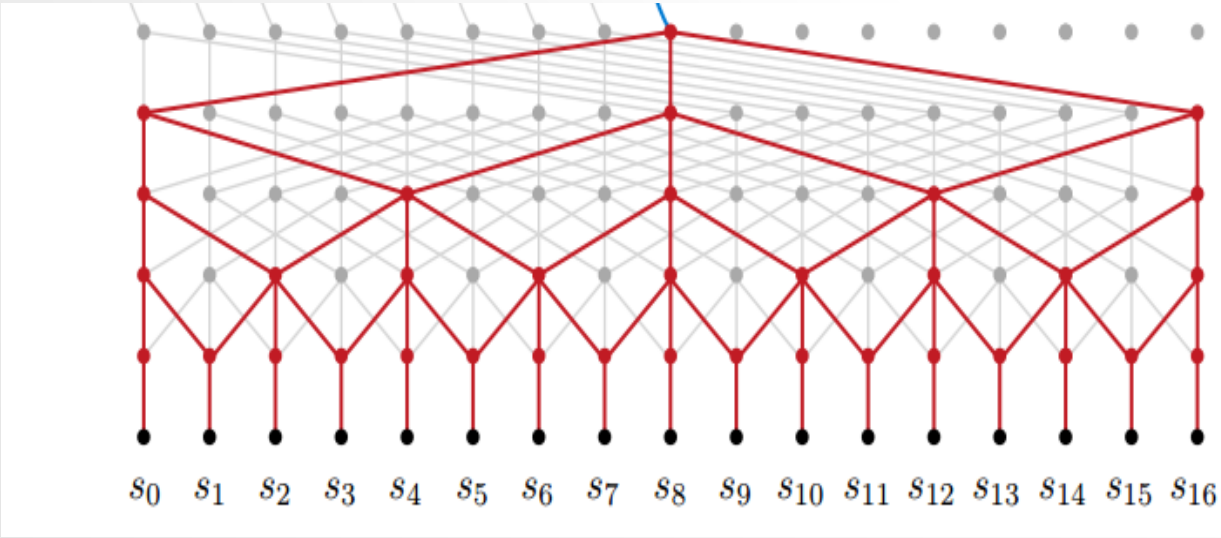
- Given a target string **t** = ***t0***,……,***tn*** the target network embeds each of the first *n* tokens ***t0***,……,***tn-1*** via a look-up table (the n tokens ***t1***,……,***tn*** serve as targets for the predictions)

- The target network applies masked one-dimensional convolutions to the embedding tensor that have a masked kernel of size *k*. The masking ensures that information from future tokens does not affect the prediction of the current token.



TECHNIQUES USED IN THIS MODEL: DILATION

- The masked convolutions use dilation to **increase the receptive field** of the target network
- Dilation makes the receptive field **grow exponentially** in terms of the depth of the networks, as opposed to linearly.
- We use a dilation scheme whereby the dilation rates are doubled every layer up to a maximum rate *r* (for our experiments *r* = 16). The scheme is repeated multiple times in the network always starting from a dilation rate of 1

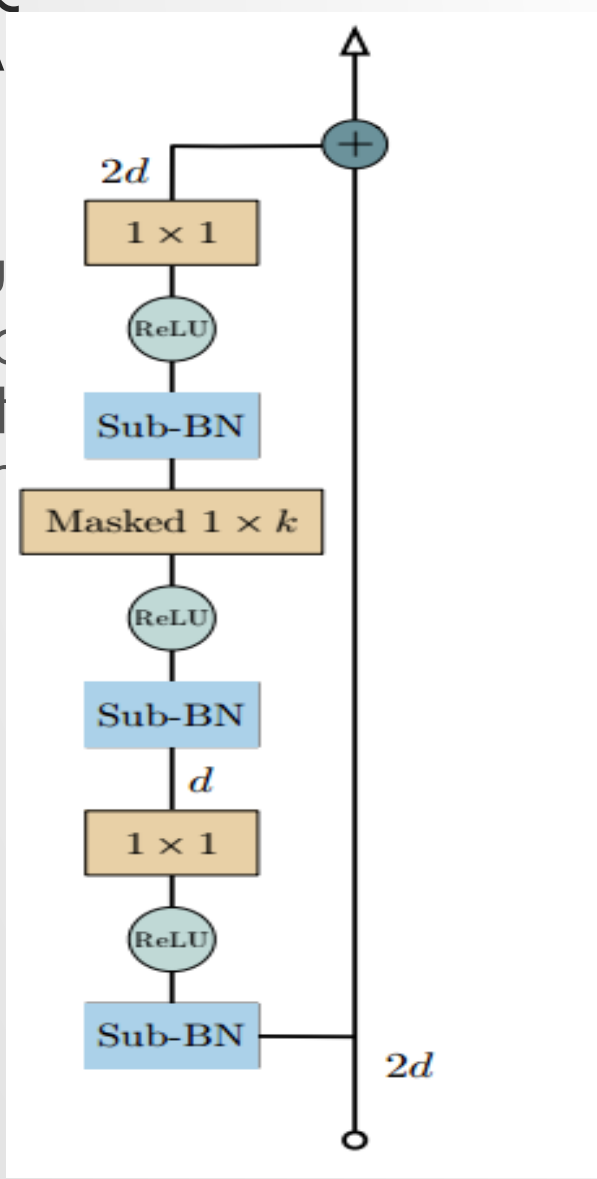


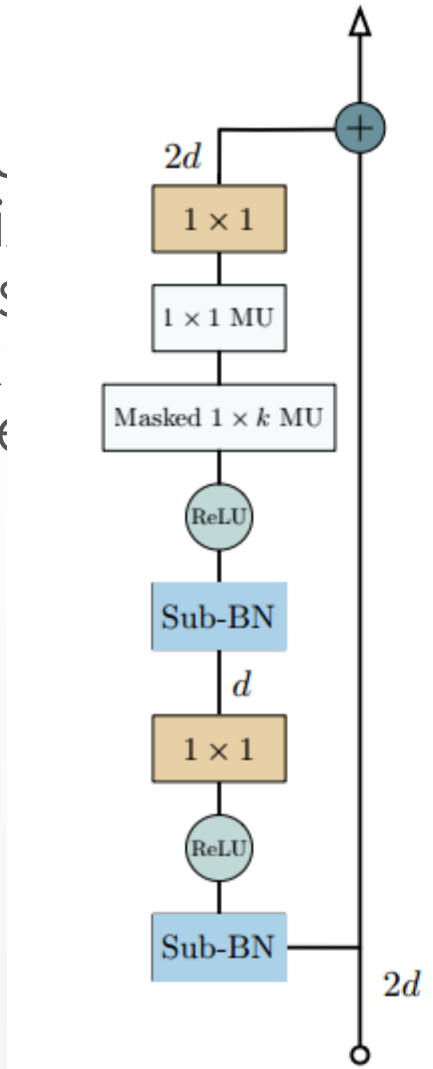
TECHNIQUES USED IN THIS MODEL: RESIDUA

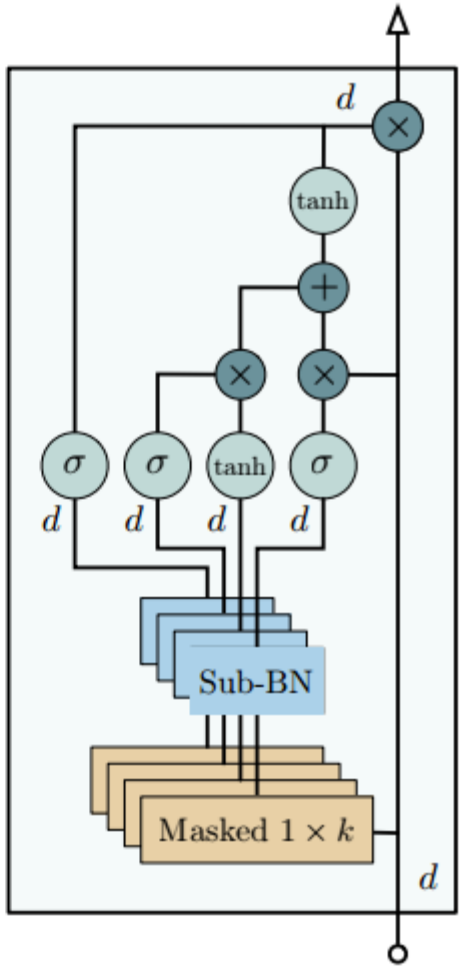
- Each
convolu
- We ac
used in t
Units, wh

a residu
ers of si
f the res
ition ex
nguage

h is
tive







TECHNIQUES USED IN THIS MODEL:
SUB-BATCH NORMALIZATION

- **Standard BN**: computes the mean and variance of the activations of a given convolutional layer along the batch, height, and width dimensions

- **Problem**: BN output for each target token would incorporate the information about the tokens that follow it. This breaks the conditioning structure since the succeeding tokens are yet to be predicted.

- **Solution**: For each layer, the mean and variance of its activations are computed over the auxiliary batch, but are used for the batch normalization of the main batch. At the same time, the loss is computed only on the predictions of the main batch, ignoring the predictions from the auxiliary batch

TECHNIQUES USED IN THIS MODEL: BAG OF CHARACTER N-GRAMS

- The tokens that we adopt correspond to characters in the input sequences.
- An efficient way to increase the capacity of the models is to use input embeddings not just for single tokens, but also for *n*-grams of adjacent tokens. At each position we **sum the embeddings** of the respective *n*-grams for $1 \leq n \leq 5$ component-wise into a single vector.
- The **length** of the sequences corresponds to the number of characters and does **not change** when using bags of *n*-grams.

MODEL COMPARISON

The way of combining source and target networks is not tied to the networks being strictly convolutional. We may consider two variants of the ByteNet that use recurrent

networks for one or both of the sub-networks

MODEL COMPARISON

Model	Net_S	Net_T	Time	RP	Path_S	Path_T
RCTM 1	CNN	RNN	$\|S\|\|S\| + \|T\|$	no	$\|S\|$	$\|T\|$
RCTM 2	CNN	RNN	$\|S\|\|S\| + \|T\|$	yes	$\|S\|$	$\|T\|$
RNN Enc-Dec	RNN	RNN	$\|S\| + \|T\|$	no	$\|S\| + \|T\|$	$\|T\|$
RNN Enc-Dec Att	RNN	RNN	$\|S\|\|T\|$	yes	1	$\|T\|$
Grid LSTM	RNN	RNN	$\|S\|\|T\|$	yes	$\|S\| + \|T\|$	$\|S\| + \|T\|$
Extended Neural GPU	cRNN	cRNN	$\|S\|\|S\| + \|S\|\|T\|$	yes	$\|S\|$	$\|T\|$
Recurrent ByteNet	RNN	RNN	$\|S\| + \|T\|$	yes	$\max(\|S\|, \|T\|)$	$\|T\|$
Recurrent ByteNet	CNN	RNN	$c\|S\| + \|T\|$	yes	c	$\|T\|$
ByteNet	CNN	CNN	$c\|S\| + c\|T\|$	yes	c	c

Properties of various previously and presently introduced neural translation models. The ByteNet models have both linear running time and are resolution preserving.

MODEL COMPARISON

Model	Test
Stacked LSTM (Graves, 2013)	1.67
GF-LSTM (Chung et al., 2015)	1.58
Grid-LSTM (Kalchbrenner et al., 2016a)	1.47
Layer-normalized LSTM (Chung et al., 2016a)	1.46
MI-LSTM (Wu et al., 2016b)	1.44
Recurrent Highway Networks (Srivastava et al., 2015)	1.42
Recurrent Memory Array Structures (Rocki, 2016)	1.40
HM-LSTM (Chung et al., 2016a)	1.40
Layer Norm HyperLSTM (Ha et al., 2016)	1.38
Large Layer Norm HyperLSTM (Ha et al., 2016)	1.34
ByteNet Decoder	**1.33**

Negative log-likelihood results in bits/byte on the Hutter Prize Wikipedia benchmark.

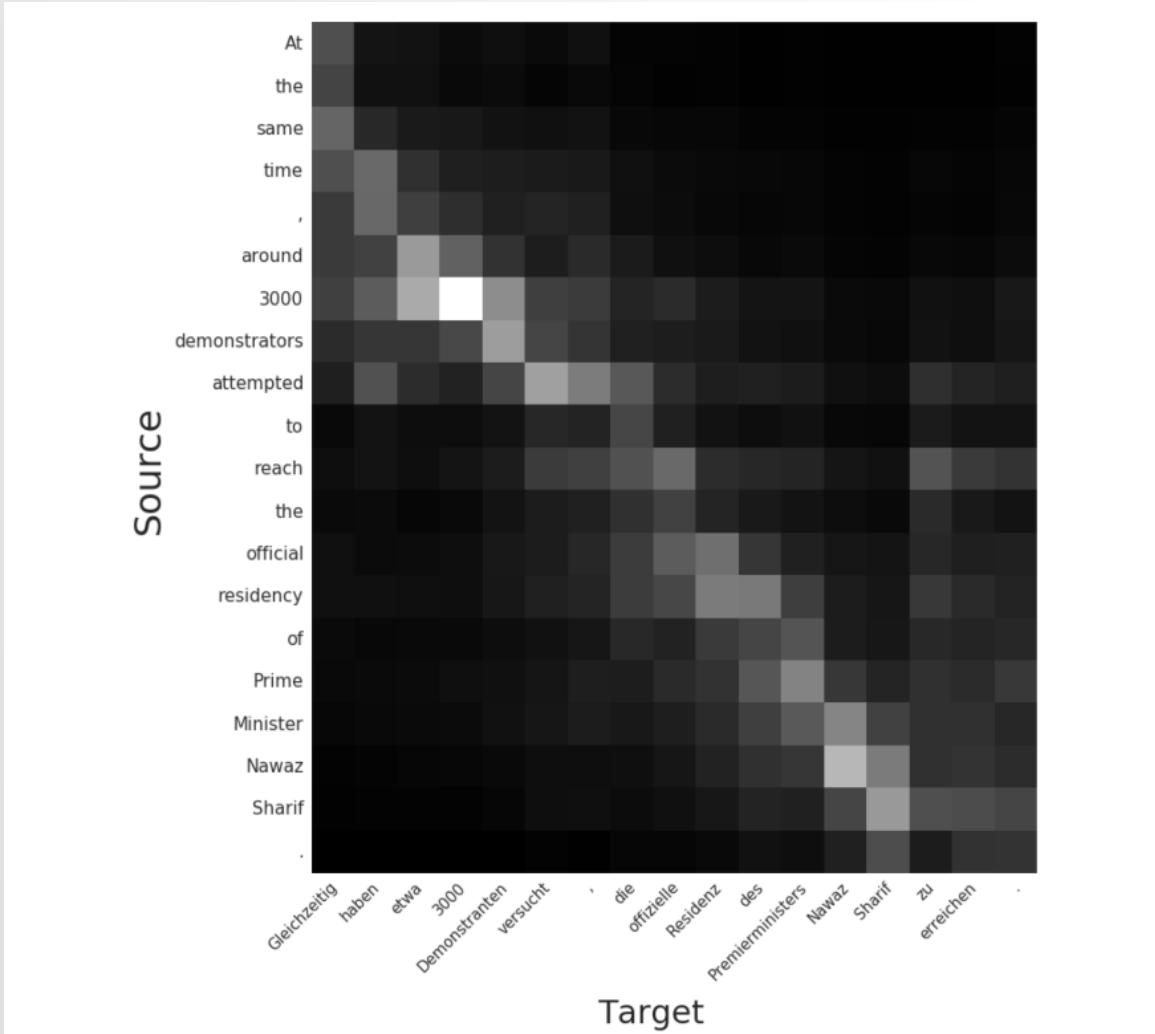
MODEL COMPARISON

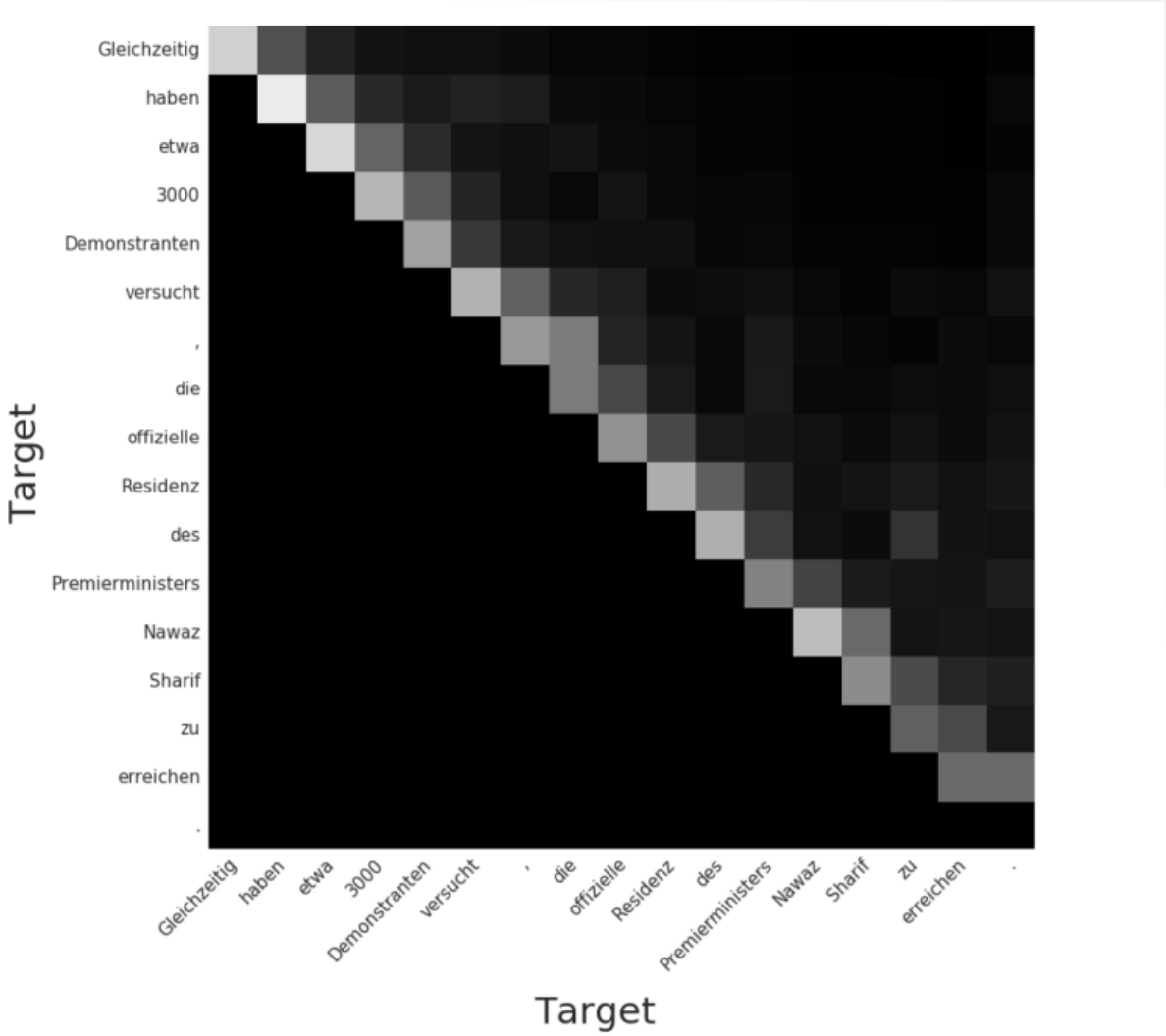
Model	WMT Test '14	WMT Test '15
Phrase Based MT	20.7[1]	24.0[2]
RNN Enc-Dec	11.3[3]	
RNN Enc-Dec + reverse	14.0[3]	
RNN Enc-Dec Att	16.9[3]	
RNN Enc-Dec Att + deep (Zhou et al., 2016)	20.6	
RNN Enc-Dec Att + local p + unk replace	20.9[3]	
RNN Enc-Dec Att + BPE in + BPE out	19.98[4]	21.72[4]
RNN Enc-Dec Att + BPE in + char out	21.33[4]	23.45[4]
GNMT + char in + char out (Wu et al., 2016a)	22.8	
ByteNet	*18.9*	*21.7*

BLEU scores on En-De WMT NewsTest 2014 and 2015 test sets.

MODEL COMPARISON

Lengths of sentences in characters and their correlation coefficient for the En-De and the En-Ru WMT NewsTest-2013 validation data.


Source
At
the
same
time
,
around
3000
demonstrators
attempted
to
reach
the
official
residency
of
Prime
Minister
Nawaz
Sharif
.
Target
Gleichzeitig
haben
etwa
3000
Demonstranten
versucht
,
die
offizielle
Residenz
des
Premierministers
Nawaz
Sharif
zu
erreichen
.


Target
Gleichzeitig
haben
etwa
3000
Demonstranten
versucht
,
die
offizielle
Residenz
des
Premierministers
Nawaz
Sharif
zu
erreichen
.
Target