UNIVERSITY OF AMSTERDAM

Annual Conference of the European Association for Machine Translation

2017

Convolutional over Recurrent Encoder for Neural Machine Translation

Praveen Dakwale and Christof Monz

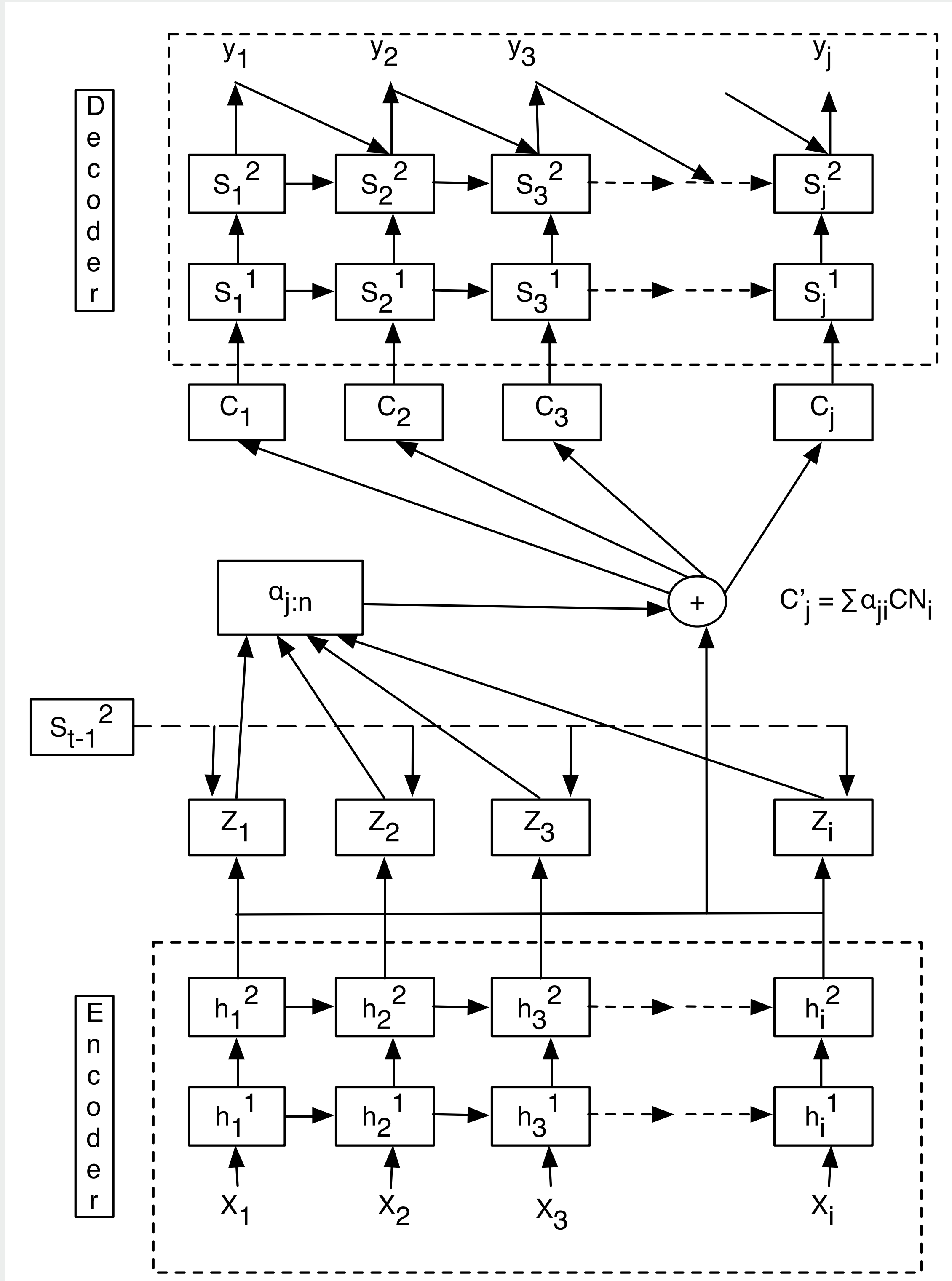
Neural Machine Translation

- End to end neural network with RNN architecture where the output of an RNN (decoder) is conditioned on another RNN (encoder).

$$p(y_i|y_1,\ldots,y_{i-1},\mathbf{x}) = g(y_{i-1}, s_i, c_i),$$

- *c* is a fixed length vector representation of source sentence encoded by RNN.
- Attention Mechanism :
  - (Bahdanau et al 2015) : compute conext vector as weighted average of annotations of source hidden states.

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j.$$


Decoder
Encoder
y_1 y_2 y_3 y_j
s_1^2 s_2^2 s_3^2 s_j^2
s_1^1 s_2^1 s_3^1 s_j^1
C_1 C_2 C_3 C_j
$\alpha_{j:n}$
+
$C'_j = \sum \alpha_{ji} CN_i$
S_{t-1}^2
Z_1 Z_2 Z_3 Z_i
h_1^2 h_2^2 h_3^2 h_i^2
h_1^1 h_2^1 h_3^1 h_i^1
X_1 X_2 X_3 X_i

Why RNN works for NMT ?

- Recurrently encode history for variable length large input sequences
- Capture the long distance dependency which is an important occurrence in natural language text

RNN for NMT:

- Disadvantages :
  - Slow : Doesn't allow parallel computation within sequence
  - Non-uniform composition : For each state, first word is over-processed and the last one only once
  - Dense representation : each h_i is a compact summary of the source sentence up to word 'i'
  - Focus on global representation not on local features

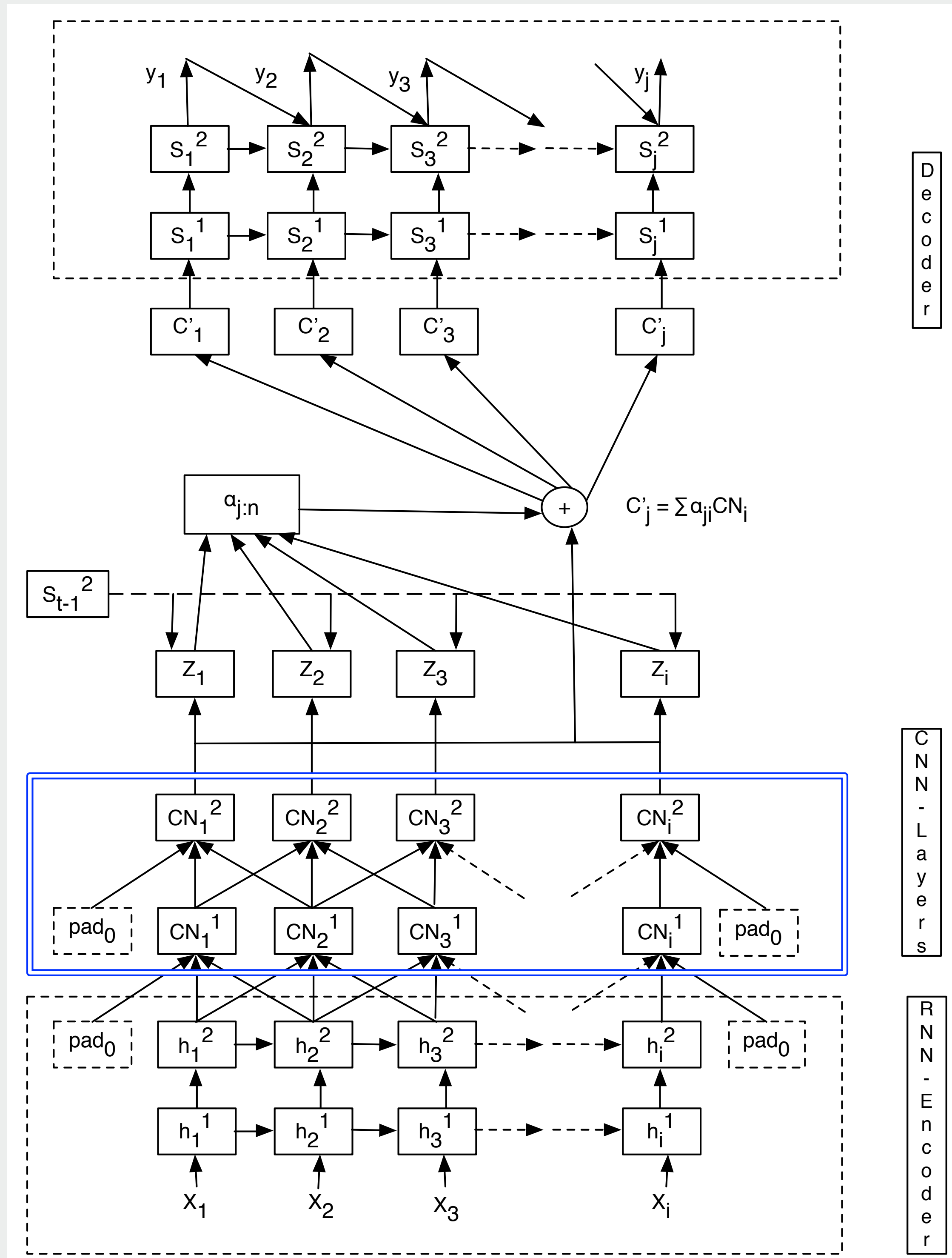
CNN in NLP :

- Unlike RNN, CNN apply over a fixed size window of input
  - This allows for parallel computation
- Represent sentence in terms of features:
  - a weighted combination of multiple words or n-grams
- Very successful in learning sentence representations for various tasks
  - Sentiment analysis, question classification (*Kim 2014, Kalchbrenner et al 2014*)

Convolution over Recurrent encoder (CoveR):

- Can CNN help for NMT ?
  - Instead of single recurrent outputs, we can use a composition of multiple hidden state outputs of the encoder
- Convolution over recurrent :
  - We apply multiple layers of fixed size convolution filters over the output of the RNN encoder at each time step
  - Can provide wider context about the relevant features of the source sentence

CoveR model


y1
y2
y3
yj
S1 2
S2 2
S3 2
Sj 2
S1 1
S2 1
S3 1
Sj 1
C'1
C'2
C'3
C'j
Decoder
αj:n
+
C'j = Σ αji CNi
St-1 2
Z1
Z2
Z3
Zi
CN1 2
CN2 2
CN3 2
CNi 2
pad0
CN1 1
CN2 1
CN3 1
CNi 1
pad0
CNN-Layers
pad0
h1 2
h2 2
h3 2
hi 2
pad0
h1 1
h2 1
h3 1
hi 1
X1
X2
X3
Xi
RNN-Encoder

Convolution over Recurrent encoder:

- Each of the vectors CN_i now represents a feature produced by multiple kernels over h_i

$$CN_i^1 = \sigma(\theta \cdot h_{i-[(w-1)/2]:i+[(w-1)/2]} + b)$$

- Relatively uniform composition of multiple previous states and current state.
- Simultaneous hence faster processing at the convolutional layers

Related work:

- Gehring et al 2017:
  - Completely replace RNN encoder with CNN
  - Simple replacement doesn’t work, position embeddings required to model dependencies
  - Require 6-15 convolutional layers to compete 2 layer RNN
- Meng et al 2015 :
  - For Phrase-based MT, use CNN language model as additional feature

Experimental setting:

- Data :
  - WMT-2015 En-De training data : 4.2M sentence pairs
  - Dev : WMT2013 test set
  - Test : WMT2014,WMT2015 test sets
- Baseline :
  - Two layer unidirectional LSTM encoder
  - Embedding size, hidden size = 1000
  - Vocab : Source : 60k, Target : 40k

Experimental setting:

- CoveR :
  - Encoder : 3 convolutional layers over RNN output
  - Decoder : same as baseline
  - Convolutional filters of size : 3
  - Output dimension : 1000
  - Zero padding on both sides at each layer, no pooling
  - Residual connection (He et, al 2015) between each intermediate layer

Experimental setting:

- Deep RNN encoder :
  - Comparing 2 layer RNN encoder baseline to CoveR is unfair
    - Improvement maybe just due to increased number of parameters
  - We compare with a deep RNN encoder with 5 layers
  - 2 layers of decoder initialized through a non-linear transformation of encoder final states

Result

BLEU scores (* = significant at $p < 0.05$)

BLEU	Dev	wmt14	wmt15
Baseline	17.9	15.8	18.5
Deep RNN encoder	18.3	16.2	18.7
CoveR	**18.5**	**16.9***	**19.0***

- Compared to baseline:
  - +1.1 for WMT-14 and 0.5 for WMT-15
- Compared to deep RNN encoder :
  - +0.7 for WMT-14 and 0.3 for WMT-15

Result

#parameters and decoding speed

BLEU	#parameters (millions)	avg sec/ sent
Baseline	174	0.11
Deep RNN encoder	283	0.28
CoveR	**183**	**0.14**

- CoveR model:
  - Slightly slower than baseline but faster than deep RNN
  - Slightly more parameter than baseline but less than deep RNN
- Improvements not just due to increased number of parameters

Qualitative analysis :

- Increased output length

Source :	as the reverend martin luther king jr. said fifty years ago
Reference :	wie pastor martin luther king jr. vor fünfzig jahren sagte :
Baseline :	wie der martin luther king jr. sagte
Cover :	wie der martin luther king jr. sagte **vor fünfzig jahren** :

BLEU	Avg sent length
Baseline	18.7
Deep RNN	19.0
CoveR	**19.9**
Reference	20.9

- With additional context, CoveR model generates complete translation

Qualitative analysis :

- More uniform attention distribution

Source :	he said the itinerary is still being worked out .
Reference :	er sagte , das genaue reiseroute werde noch ausgearbeitet .
Baseline :	er sagte , dass die strecke noch <unk> ist .
Cover :	er sagte , die **reiseroute** wird noch **ausgearbeitet** .

- Generation of correct composite word

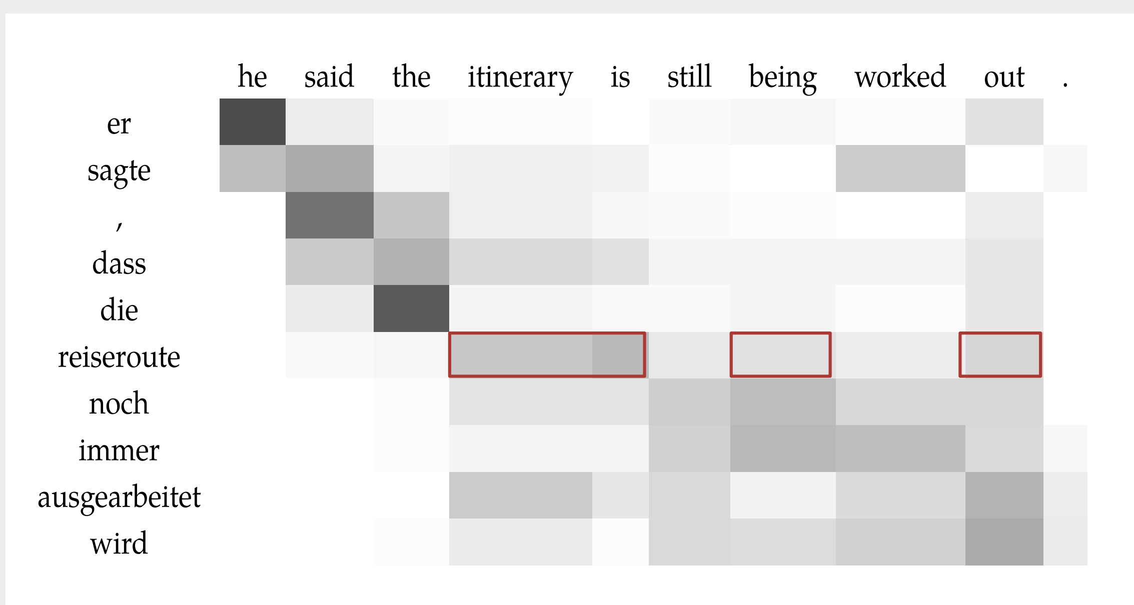
Qualitative analysis :

- More uniform attention distribution

Baseline

- Baseline translates : ‘*itinerary*’ to ‘*strecke*’ (road, distance)
- Pays attention only to ‘*itinerary*’ for this position

CoveR



- Cover translates : ‘*itinerary*’ to ‘*reiseroute*’
- Also pays attention to final verb

Conclusion :

- CoveR : multiple convolutional layers over RNN encoder
- Significant improvements over standard LSTM baseline
- Increasing LSTM layers improves slightly, but convolutional layers perform better
- Faster and less parameters than fully RNN encoder of same size
- CoveR model can improve coverage and provide more uniform attention distribution

Thanks

Questions ?