Quasi-Recurrent Neural Networks

Published as a conference paper at ICLR 2017

QUASI-RECURRENT NEURAL NETWORKS

James Bradbury,* Stephen Merity,* Caiming Xiong & Richard Socher
Salesforce Research
Palo Alto, California
{james.bradbury,smerity,cxiong,rsocher}@salesforce.com

Problem

Jak przyspieszyć RNNy, nie tracąc na jakości?

Rozwiązanie

- Intuicja: Części sekwencji, które nie zależą od kontekstu mogą być przetwarzane równolegle, reszta musi być przetwarzana sekwencyjnie
- Połączenie CNN i RNN
- Splot z maską (masked convolution)
- Zredukowanie sekwencyjnych obliczeń
- Do 16x przyspieszenie uczenia

QRNN vs LSTM

masked convolution

$$\mathbf{f} = \sigma(\mathbf{W}_f * \mathbf{x})$$
$$\mathbf{i} = \sigma(\mathbf{W}_i * \mathbf{x})$$
$$\mathbf{o} = \sigma(\mathbf{W}_o * \mathbf{x})$$
$$\mathbf{z} = \tanh(\mathbf{W}_z * \mathbf{x})$$
$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \mathbf{z}_t$$
$$\mathbf{h}_t = \mathbf{o}_t \odot \mathbf{c}_t$$

ifo-pooling

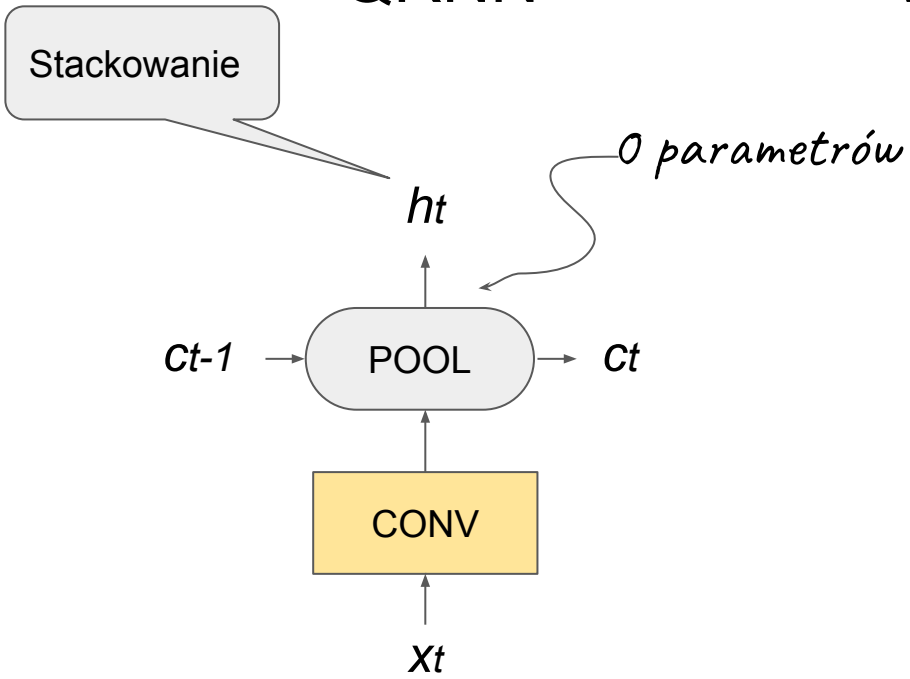
(równoległe w kanałach)

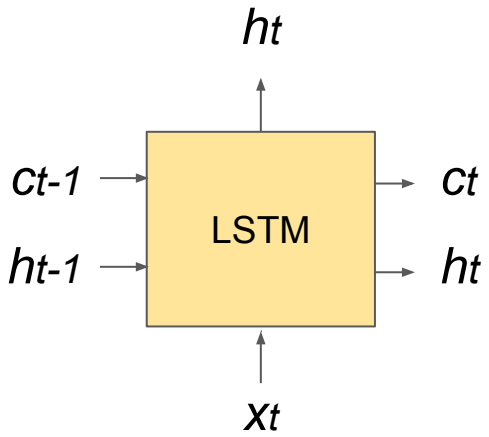
$$\mathbf{f}_t = \sigma(\mathbf{W}_f [\mathbf{h}_{t-1}, \mathbf{x}_t])$$
$$\mathbf{i}_t = \sigma(\mathbf{W}_i [\mathbf{h}_{t-1}, \mathbf{x}_t])$$
$$\mathbf{o}_t = \sigma(\mathbf{W}_o [\mathbf{h}_{t-1}, \mathbf{x}_t])$$
$$\mathbf{z}_t = \tanh(\mathbf{W}_z [\mathbf{h}_{t-1}, \mathbf{x}_t])$$

(a_t)

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \mathbf{z}_t$$
$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t)$$

QRNN vs LSTM


Stackowanie
0 parametrów
ht
ct-1
POOL
ct
CONV
xt


ht
ct-1
LSTM
ct
ht-1
ht
xt

Eksperymenty

1. Klasyfikacja sentymentu
2. Modelowanie języka
3. Tłumaczenie maszynowe

Klasyfikacja sentymentu

- Klasyfikacja sentymentu recenzji z IMDB (50000 dokumentów, średnio 231 słów na dokument).
- 256 units / layer, dropout, skip-connections.

Model	**Time / Epoch (s)**	**Test Acc (%)**
NBSVM-bi (Wang & Manning, 2012)	–	91.2
2 layer sequential BoW CNN (Johnson & Zhang, 2014)	–	92.3
Ensemble of RNNs and NB-SVM (Mesnil et al., 2014)	–	92.6
2-layer LSTM (Longpre et al., 2016)	–	87.6
Residual 2-layer bi-LSTM (Longpre et al., 2016)	–	90.1
Our models		
Densely-connected 4-layer LSTM (cuDNN optimized)	480	90.9
Densely-connected 4-layer QRNN	150	91.4
Densely-connected 4-layer QRNN with $k = 4$	160	91.1

Klasyfikacja sentymentu cd.

- 3.2x szybciej od porównywalnego LSTM na epokę
- do 16.9x szybciej na batch

Model	Time / Epoch (s)
NBSVM-bi (Wang & Manning, 2012)	–
2 layer sequential BoW CNN (Johnson & Zhang, 2014)	–
Ensemble of RNNs and NB-SVM (Mesnil et al., 2014)	–
2-layer LSTM (Longpre et al., 2016)	–
Residual 2-layer bi-LSTM (Longpre et al., 2016)	–
Our models	
Densely-connected 4-layer LSTM (cuDNN optimized)	480
Densely-connected 4-layer QRNN	150
Densely-connected 4-layer QRNN with $k = 4$	160

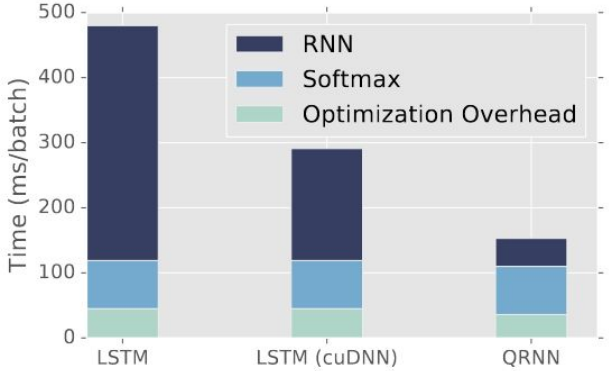
		Sequence length				
		32	64	128	256	512
Batch size	8	**5.5x**	**8.8x**	**11.0x**	**12.4x**	**16.9x**
	16	**5.5x**	**6.7x**	**7.8x**	**8.3x**	**10.8x**
	32	**4.2x**	**4.5x**	**4.9x**	**4.9x**	**6.4x**
	64	**3.0x**	**3.0x**	**3.0x**	**3.0x**	**3.7x**
	128	**2.1x**	**1.9x**	**2.0x**	**2.0x**	**2.4x**
	256	**1.4x**	**1.4x**	**1.3x**	**1.3x**	**1.3x**

Modelowanie języka (word-level)

- Predykcja sekwencji (POS); Zbiór Penn Treebank
- 2 warstwy po 640 jednostek, dropout + early stopping, bez dropout

Model	Parameters	Validation	Test
LSTM (medium) (Zaremba et al., 2014)	20M	86.2	82.7
Variational LSTM (medium, MC) (Gal & Ghahramani, 2016)	20M	81.9	79.7
LSTM with CharCNN embeddings (Kim et al., 2016)	19M	–	78.9
Zoneout + Variational LSTM (medium) (Merity et al., 2016)	20M	84.4	80.6
Our models			
LSTM (medium)	20M	85.7	82.0
QRNN (medium)	18M	82.9	79.9
QRNN + zoneout ($p = 0.1$) (medium)	18M	82.1	78.3

Table 2: Single model perplexity on validation and test sets for the Penn Treebank language modeling task. Lower is better. "Medium" refers to a two-layer network with 640 or 650 hidden units per layer. All QRNN models include dropout of 0.5 on embeddings and between layers. MC refers to Monte Carlo dropout averaging at test time.



Tłumaczenie maszynowe (char-level)

- IWSLT German-English: 200k par zdań z języka mówionego (średnio 103 znaków na zdanie)
- 320 units / layer, bez dropoutu, w enkoderze splot jest bez maski

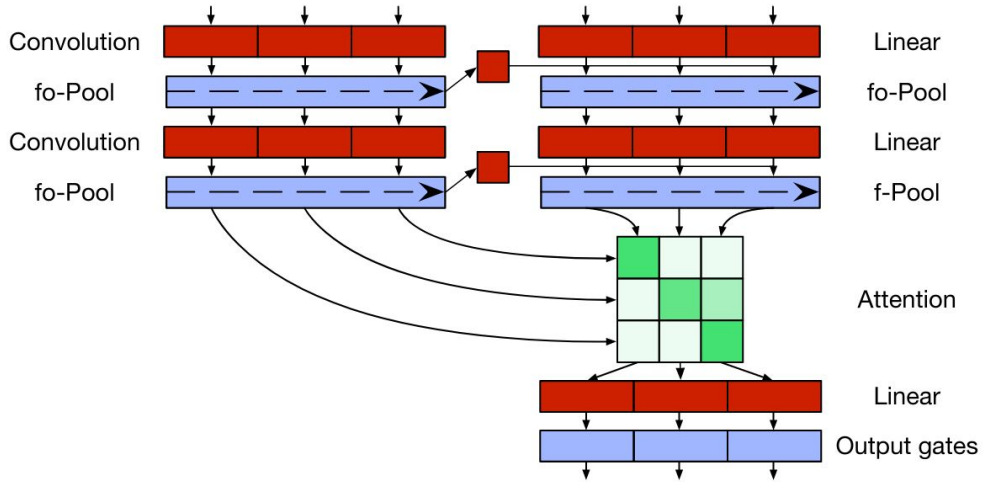


Figure 2: The QRNN encoder–decoder architecture used for machine translation experiments.

Tłumaczenie maszynowe (char-level) cd.

- IWSLT German-English: 200k par zdań z języka mówionego (średnio 103 znaków na zdanie)
- 320 units / layer, bez droupoutu, splot encodera bez maski

Model	**Train Time**	**BLEU (TED.tst2014)**
Word-level LSTM w/attn (Ranzato et al., 2016)	–	20.2
Word-level CNN w/attn, input feeding (Wiseman & Rush, 2016)	–	24.0
Our models		
Char-level 4-layer LSTM	4.2 hrs/epoch	16.53
Char-level 4-layer QRNN with $k = 6$	1.0 hrs/epoch	19.41

Podsumowanie

Dziękuję za uwagę