Efficient Purely Convolutional Text Encoding

Szymon Malik*, Adrian Lancucki*, Jan Chorowski

University of Wroclaw, Poland

szymon.w.malik@gmail.com, {adrian.lancucki, jan.chorowski}@cs.uni.wroc.pl

Introduction

We investigate Byte-Level Recursive Convolutional Auto-Encoder (BRCA) by Xiang and LeCun [1], which **auto-encodes paragraphs of text on byte-level**.

Research Questions

Q1: How does BRCA achieve its byte-level auto-encoding accuracy?
Q2: Is it useful for embedding sentences?

By answering those questions we are able to:

- increasing accuracy, speed up convergence, lower # of parameters,
- develop a light-weight **sentence embedding** method.

The Model and Our Modifications

The model was trained on raw paragraphs of English Wikipedia [1]. It truned out slow and hard to train. To mitigate that, we have introduced modifications:

- adding batch normalizaion (recursive issues),
- removing linear layers,
- different input padding.

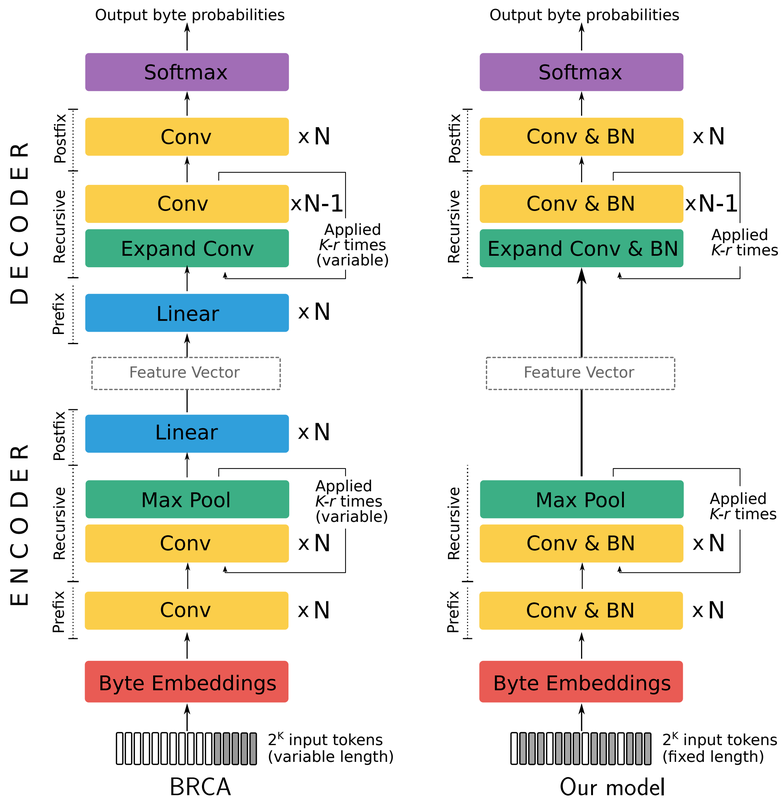


Figure: Structural comparison of Byte-Level Recursive Convolutional Auto-Encoder (BRCA) and our model. Dark boxes indicate input padding. BRCA pads the input form the right to the nearest power of two. We pad the input evenly to a fixed-size vector. Our model does not have postfix/prefix groups with linear layers and uses Batch Normalization (BN) after every layer.

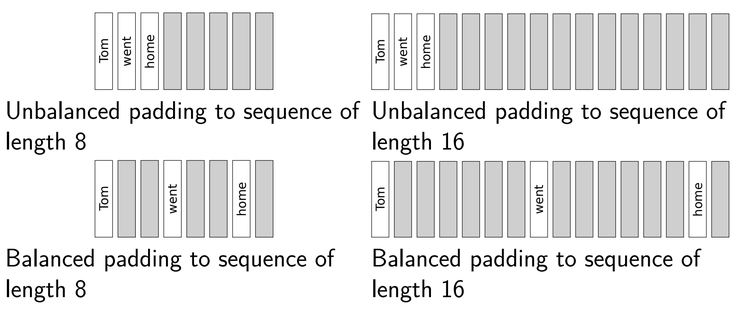


Figure: Unbalanced and balanced padding of an input sequence to a fixed-length sequence. Grey boxes are (zero) padding, white boxes are input embedding vectors.

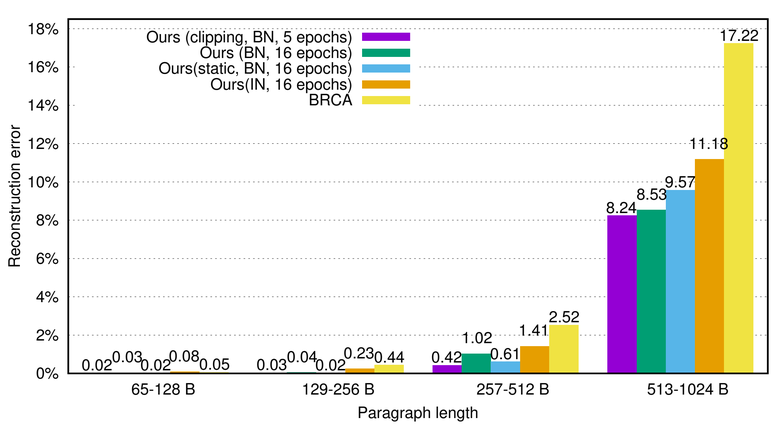


Figure: Decoding errors on unseen data for our best models ($N = 8$, no linear layers) with balanced input padding to a sequence of size 1024 compared with Byte-Level Recursive Convolutional Auto-Encoder (BRCA).

Model Analysis: Learning Identity

Table: Learning identity by training on random sequences of alphanumerical ASCII characters of different length. Accuracy is presented for BRCA (N=8) model.

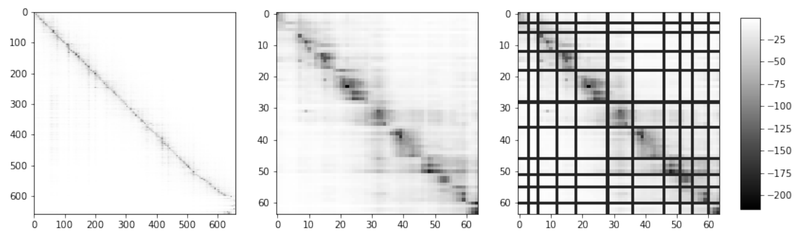
Training Lengths	Test Length	Accuracy
4 – 128	128	99.81%
4 – 512	128	60.79%
	256	22.99%
	512	9.81%

Model Analysis: Generalization to Longer Texts

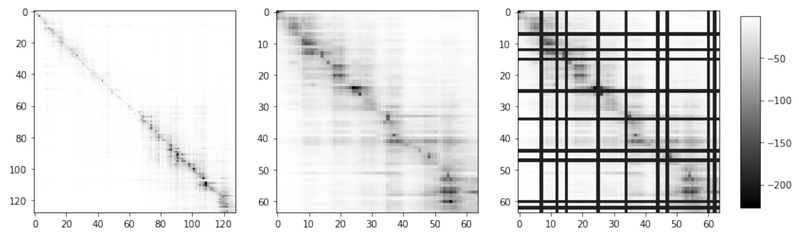
Table: Comparison of the ability of BRCA and LSTM encoder-decoder to learn an identity function and generalize to unseen data. Values represent byte-level decoding accuracy. Note that the LSTM decoder has the advantage of always being primed with the correct prefix sequence.

Lengths (bytes)	BRCA (N=2)	LSTM-LSTM
9-16	97.06%	91.17%
17-32	97.96%	90.20%
33-64	97.45%	91.72%
65-128	83.56%	86.34%
129-256	11.66%	72.88%
257-512	8.05%	58.80%

Model Analysis: Input-Output Relations



Input sentence: *One of Lem's major recurring themes, beginning from his very first novel, "The Man from Mars" (. . .)*



Input sentence: *Typical fuel is denatured alcohol, methanol, or isopropanol (. . .)*

Figure: Input-output byte relations (X axis vs. Y axis) as indicated by the method of Integrated Gradients [2] with 50 integration points. The plots correspond to (a) 659-byte, and (b) 128-byte Wikipedia paragraphs. The leftmost plots show relations between all input-output bytes, the middle plots for the first 64 bytes. The rightmost plots also plot spaces. Dark shades indicate strong relations - those lay along diagonals and do not cross word and phrases boundaries.

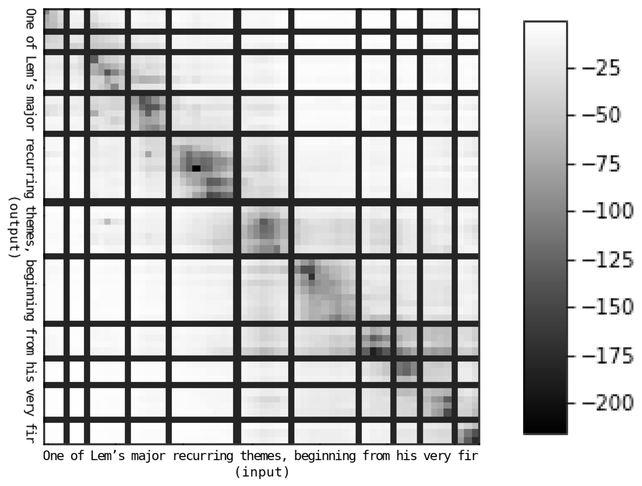


Figure: Input-output byte relations aligned with characters.

```
In :  9H3cxn4RIRUnOuPymw28dxUoA060LQ3heq1diKcbiUoinkzDjxucnE3Hk7FEFwHjzcTlOrhPUp3kgt9y8VAaw1sYpjPO9 (...)
Out:  9H3cxn4RIRUnOuPymw28dxUoA060LQ3heqddiKcbiUoinkzDjxucnE3Hk7FEFwHjzcTlOrhPUp3kgt9y8VAaw1sYpjPO9 (...)
```

Random string of characters (128 bytes)

```
In :  Lorsque ce m\xc3\xa9lange de cultures mondiales doit donner une signification au fait d'\xc3 (...)
Out:  Lorsque ce m\xc3\xa9lange de cultures mondiales doit donner une signification au fait d'\xc3 (...)

In :  When we first sequenced this genome in 1995, the standard of accuracy was one error per 10,00 (...)
Out:  When we first sequenced this genome in 1995, the standard of accuracy was one error per 10,00 (...)
```

French and English sentences (256 and 128 bytes respectively)

```
In :  Lorsque ce m\xc3\xa9lange de cultures mondiales doit donner une signification au fait d'\xc3 (...)
Out:  Larkere de Bu\xa9   lande by mortures mondiales leid dunner one mignification or Zaan ''\xc3 (...)

In :  When we first sequenced this genome in 1995, the standard of accuracy was one error per 10,00 (...)
Out:  When we first sequenced this genome in 1995, the standard of accuracy was one error per 10,00 (...)
```

French and English sentences concatenated 4 times (1024 and 512 bytes)

Figure: Auto-encoding capabilities of the model with errors marked in **bold red**. The model was trained only on English Wikipedia paragraphs. On short sequences, our model performs close to an identity function. On longer ones, it seems to correctly auto-encode only English paragraphs. Note that the model tries to map French words into English ones (`avec` → `open`, `une` → `one`).

Word-Level Sentence Encoder

Following the methods and work of [3], we apply our architecture to a practical task. Namely, we train models consisting of the recursive convolutional word-level encoder and a simple three-layer fully-connected classifier on Stanford Natural Language Inference (SNLI) corpus [4]. This dataset contains 570k sentence pairs, each one described by one of three relation labels: entailment, contradiction, and neutral. Then we test encoders on various transfer tasks measuring semantic similarities between sentences.

Table: Results for word-level sentence encoders. We compare bag-of-words (BoW), i.e. averaged word embeddings, WRCE - the encoder from Zhang and LeCun's model on word-level, our word-level model with balanced padding to 64 elements (Ours), and an ensemble of our model and BoW (Ours + BoW) for various supervised (classification accuracy) and unsupervised (Pearson/Spearman correlation coefficients) tasks.

	Model			
Task (dev/test acc%)	BoW	WRCE	Ours	Ours + BoW
SNLI	67.7 / 67.5	82.0 / 81.3	**83.8** / **83.1**	83.2 / 82.6
CR	**79.7** / 78.0	78.0 / 77.3	78.6 / 77.0	79.1 / **78.2**
MR	**77.7** / **77.0**	72.9 / 72.4	73.7 / 73.1	75.3 / 74.8
MPQA	**87.4** / 87.5	85.9 / 85.6	86.0 / 85.9	**87.4** / **87.6**
SUBJ	**91.8** / **91.4**	86.1 / 85.4	87.2 / 86.9	89.0 / 88.9
SST Bin. Class.	**80.4** / **81.4**	78.1 / 77.5	77.2 / 76.7	78.1 / 78.8
SST Fine-Grained Class.	**45.1** / **44.4**	38.3 / 40.5	40.5 / 39.3	41.9 / 41.4
TREC	**74.5** / **82.2**	67.0 / 72.4	69.2 / 71.4	71.0 / 77.4
MRPC	**74.4** / 73.2	72.4 / 71.1	73.5 / 72.5	74.1 / **73.3**
SICK-E	79.8 / 78.2	82.6 / 82.8	**83.6** / 81.9	83.2 / **83.0**
Task (correlation P/S)	BoW	WRCE	Ours	Ours + BoW
SICK-R	0.80 / 0.72	0.85 / 0.78	**0.87** / **0.80**	0.86 / **0.80**
STS12	0.53 / 0.54	0.56 / 0.57	0.60 / 0.60	**0.62** / **0.61**
STS13	0.45 / 0.47	0.55 / 0.54	0.53 / 0.54	**0.57** / **0.58**
STS14	0.53 / 0.54	0.65 / 0.63	0.68 / **0.70**	**0.69** / 0.66
STS15	0.56 / 0.59	0.68 / 0.69	0.70 / 0.70	**0.71** / **0.72**
STS16	0.52 / 0.57	0.69 / 0.70	0.70 / 0.72	**0.71** / **0.73**

Sample Usage in a Retrieval-Based Agent

User:	Hello, have you studied algebra?
A1 (BoW):	I never failed in mathematics. Before I was fifteen I had mastered differential and integral calculus.
A2 (Ours+BoW):	I never failed in mathematics. Before I was fifteen I had mastered differential and integral calculus.
User:	From what I understand it's a video game.
A1 (BoW):	Life is like a piano. What you get out of it depends on how you play it.
A2 (Ours+BoW):	From now on, I control all aspects of the game.
User:	God
A1 (BoW):	Knowledge of God is obedience to God.
A2 (Ours+BoW):	God and all attributes of God are eternal.
User:	Have you been to South Asia?
A1 (BoW):	We do not want chaos in South Africa.
A2 (Ours+BoW):	There is a wide knowledge gap between us and the developed world in the West and in Asia.
User:	How do you like Hackathon?
A1 (BoW):	Any thing you can do I can do better.
A2 (Ours+BoW):	Discover the things that you do that make you proud as hell.
User:	How do you think ?
A1 (BoW):	Any thing you can do I can do better.
A2 (Ours+BoW):	How you think matters more than what you think.

Figure: Sample answers of retrieval-based agents which embed sentences as either BoWs, or BoWs combined with embeddings from our method

Conclusions

- We observe popularity and effectiveness of retrieval-based systems.
- Analyzed model scales well and is highly parallelizable.
- Model achieves good results on semantic similarity tasks.

Room for improvement of the word-level encoder:

- better training (regularization and hyperparameters)
- learning on conversational data

References

[1] X. Zhang and Y. LeCun, "Byte-Level Recursive Convolutional Auto-Encoder for Text," *ArXiv e-prints*, Feb. 2018.

[2] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," *arXiv preprint arXiv:1703.01365*, 2017.

[3] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, "Supervised learning of universal sentence representations from natural language inference data," *CoRR*, vol. abs/1705.02364, 2017.

[4] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, 2015.