Reproducibility Challenge:

Imputing Out-of-Vocabulary Embeddings with LOVE Makes Language Models Robust with Little Cost

11/29/2022

Andrea Carotti: acarot2@uic.edu

Emilio Ingenito: eingen2@uic.edu

Introduction - Why Love

- In general, model performance deteriorates with unseen words (e.g. typos, slang, rare words …)

  Solution:

  → word embeddings on sub-word tokens
    - FastText
    - BERT

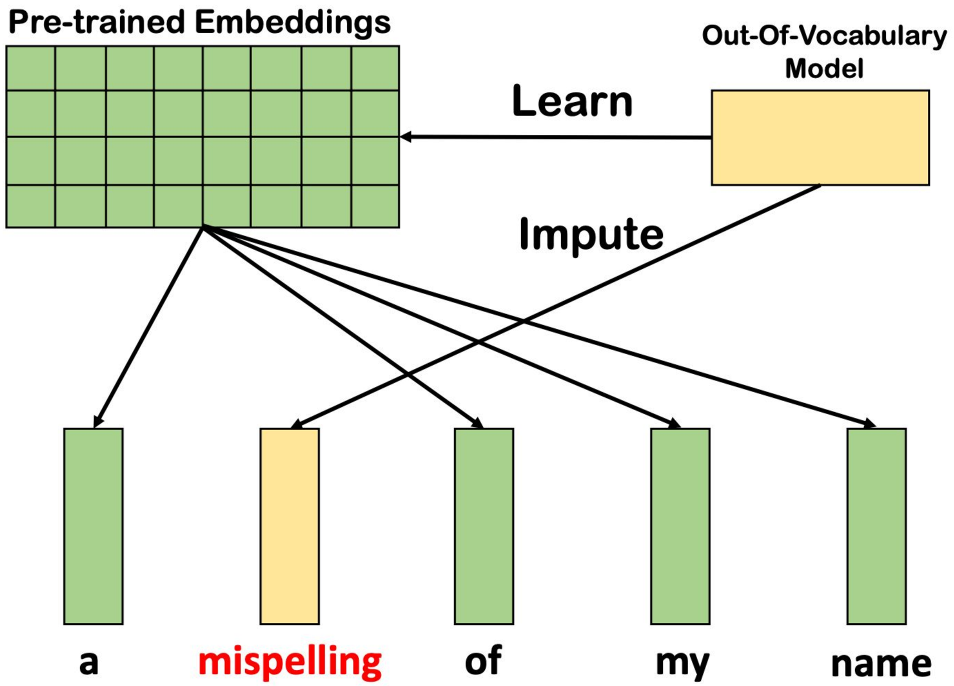
  } High Cost → Pre-training from scratch; → High memory footprint

  → MIMICK-like language models
    - MIMICK
    - BoS
    - KVQ-FH

  } Simpler → NO High-quality word vectors consistently; → Cannot used with pre-trained language models

Introduction - What is Love

- LOVE uses a novel type of data augmentation and hard negative generation
- Produces high-quality word representations robust to character perturbations
- LOVE is lightweight compared to FastText and BERT
- LOVE can be used in a plug-and-play fashion with FastText and BERT
  $\longrightarrow$ Increase Robustness



Introduction - Love Performance on intrinsic task

- Intrinsic evaluations measure syntactic or semantic relationships between words directly

	parameters		Word Similarity						Word Cluster		Avg
	embedding	others	RareWord	SimLex	MTurk	MEN	WordSim	SimVerb	AP	BLESS	
FastText (2017)	969M	-	48.1	30.4	66.9	78.1	68.2	25.7	58.0	71.5	55.9
MIMICK (2017)	9M	517K	27.1	15.9	32.5	36.5	15.0	7.5	**59.3**	**72.0**	33.2
BoS (2018)	500M	-	**44.2**	27.4	55.8	65.5	53.8	22.1	41.8	39.0	43.7
KVQ-FH (2019)	12M	-	42.4	20.4	55.2	63.4	53.1	16.4	39.1	42.5	41.6
LOVE	6.3M	200K	42.2	**35.0**	**62.0**	**68.8**	**55.1**	**29.4**	53.2	51.5	**49.7**

Introduction - Love Performance on extrinsic task

- Extrinsic evaluations measure the performance of word embeddings as input features to a downstream task
  - Named Entity Recognition (NER)
  - Text Classification

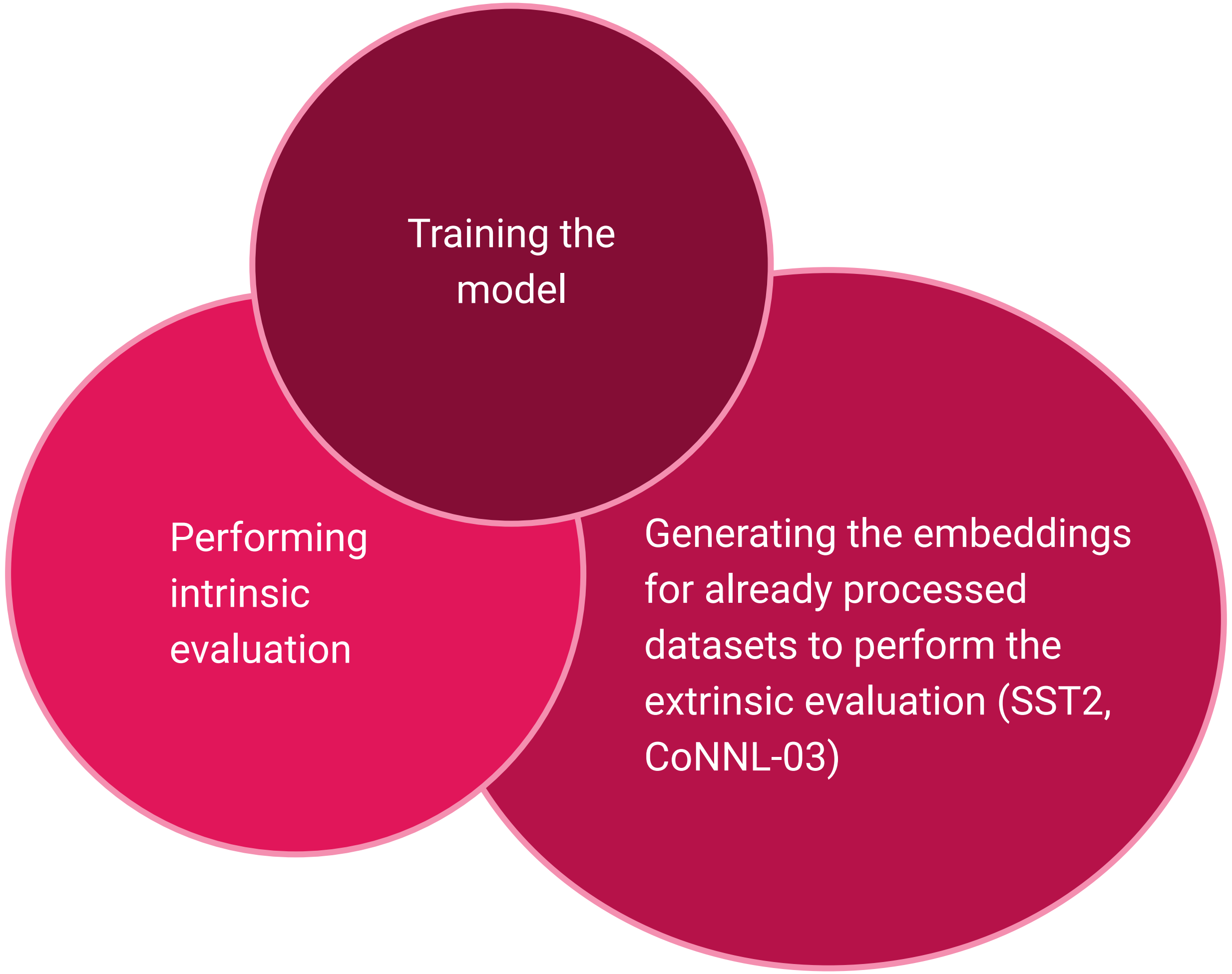
	parameters		SST2		MR		CoNLL-03		BC2GM		Avg
	embedding	others	original	+typo	original	+typo	original	+typo	original	+typo	
FastText (2017)	969M	-	82.3	60.5	73.3	62.2	86.4	66.3	71.8	53.4	69.5
Edit Distance	969M	-	-	67.4	-	68.3	-	76.2	-	66.6	-
MIMICK (2018)	9M	517K	69.7	62.3	73.6	61.4	68.0	65.2	56.6	56.7	64.2
BoS (2018)	500M	-	79.7	72.6	73.6	**69.5**	**79.5**	68.6	**66.4**	61.5	71.5
KVQ-FH (2019)	12M	-	77.8	71.4	72.9	66.5	73.1	**70.4**	46.2	53.5	66.5
LOVE	6.3M	200K	**81.4**	**73.2**	**74.4**	66.7	78.6	69.7	64.7	**63.8**	**71.6**

Introduction - LOVE performance on Extrinsic Task

- Introducing OCR typos increase the robustness of the model
- LOVE degrades performance on original datasets only marginally

	SST2						CoNLL-03						
Typo Probability	original	10%	30%	50%	70%	90%	original	10%	30%	50%	70%	90%	Avg
Static Embeddings													
FastText	**82.3**	68.2	59.8	56.7	57.8	60.3	**86.4**	81.6	78.9	73.9	70.2	63.4	70.0
FastText + LOVE	82.1	**79.8**	**74.9**	**74.2**	**68.8**	**67.2**	86.3	**84.7**	**81.8**	**77.5**	**73.1**	**71.3**	**76.8**
Dynamical Embeddings													
BERT	**91.5**	88.2	78.9	74.7	69.0	60.1	**91.2**	**89.8**	**86.2**	83.4	79.9	76.5	80.7
BERT + LOVE	**91.5**	**88.3**	**83.7**	**77.4**	**72.7**	**63.3**	89.9	88.3	86.1	**84.3**	**80.8**	**78.3**	**82.1**

Ease of reproduction



Extent of reproduction

We were **able** to reproduce the following analysis:

- Intrinsic Evaluation
  - 6 out of 8 Datasets → Word similarity tasks
- Extrinsic Evaluation
  - 4 out of 4 Datasets → Both NER and Text Classification
- Extrinsic Evaluation in a plug-and-play fashion
  - Only for FastText+Love for both SST2 and CoNLL-03 datasets

Extent of reproduction

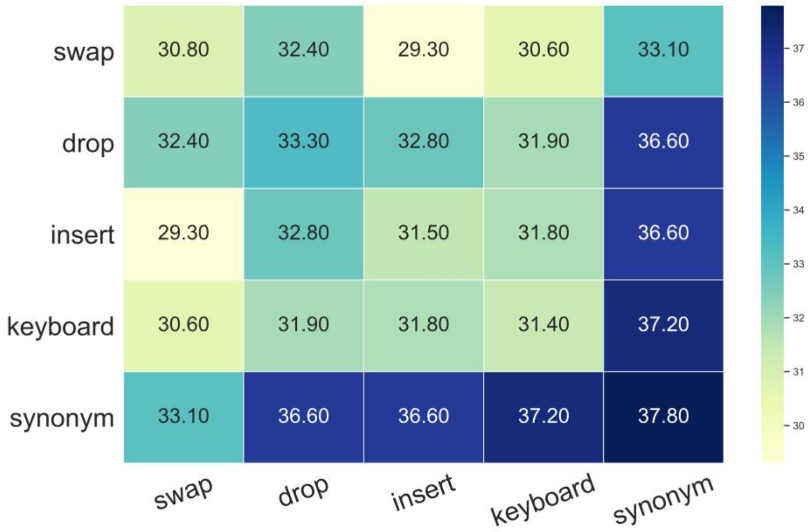
We were **unable** to reproduce the following tasks:

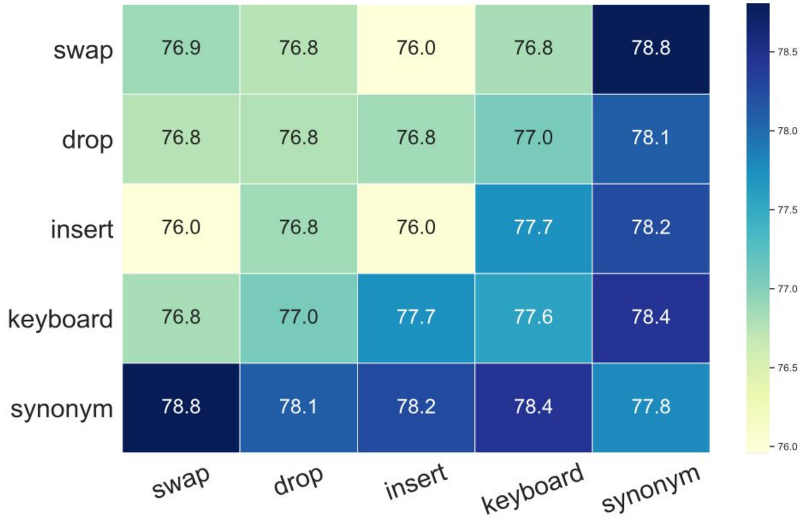
- Intrinsic Evaluation
  - 2 out of 8 Datasets (AP and BLESS) $\longrightarrow$ Word Cluster tasks
- Extrinsic Evaluation in a plug-and-play fashion
  - BERT+Love for both SST2 and CoNLL-03 datasets
- Demonstration of effectiveness of the architecture (Ablation study)
  - Varying input method, encoder and loss function
- The performance of mimicking BERT (Replacement strategy)

Extent of reproduction

What we were **unable** to reproduce:

- Performances of different augmentations on RareWord
- Performances of different augmentations on SST2





Results - intrinsic evaluation

Our Results:

RareWord	SimLex	MTurk	MEN	WordSim	SimVerb
42.65	35.02	63.77	68.4	55.89	28.72

Author's Results:

	parameters		Word Similarity						Word Cluster		Avg
	embedding	others	RareWord	SimLex	MTurk	MEN	WordSim	SimVerb	AP	BLESS	
FastText (2017)	969M	-	48.1	30.4	66.9	78.1	68.2	25.7	58.0	71.5	55.9
MIMICK (2017)	9M	517K	27.1	15.9	32.5	36.5	15.0	7.5	**59.3**	**72.0**	33.2
BoS (2018)	500M	-	**44.2**	27.4	55.8	65.5	53.8	22.1	41.8	39.0	43.7
KVQ-FH (2019)	12M	-	42.4	20.4	55.2	63.4	53.1	16.4	39.1	42.5	41.6
LOVE	6.3M	200K	42.2	**35.0**	**62.0**	**68.8**	**55.1**	**29.4**	53.2	51.5	**49.7**

Results - extrinsic evaluation

Our Results:

SST2	SST2+typo	MR	MR+typo	CoNLL-03	CoNLL-03+typo	BC2GM	BC2GM+typo
79.96	71.21	73.92	66.17	83.41	66.17	54.09	25.95

Author's Results:

	parameters		**SST2**		**MR**		**CoNLL-03**		**BC2GM**		Avg
	embedding	others	original	+typo	original	+typo	original	+typo	original	+typo	
FastText (2017)	969M	-	82.3	60.5	73.3	62.2	86.4	66.3	71.8	53.4	69.5
Edit Distance	969M	-	-	67.4	-	68.3	-	76.2	-	66.6	-
MIMICK (2018)	9M	517K	69.7	62.3	73.6	61.4	68.0	65.2	56.6	56.7	64.2
BoS (2018)	500M	-	79.7	72.6	73.6	**69.5**	**79.5**	68.6	**66.4**	61.5	71.5
KVQ-FH (2019)	12M	-	77.8	71.4	72.9	66.5	73.1	**70.4**	46.2	53.5	66.5
LOVE	6.3M	200K	**81.4**	**73.2**	**74.4**	66.7	78.6	69.7	64.7	**63.8**	**71.6**

Results - robustness evaluation

Our Results:

Extrinsic Evaluation Love+FastText on SST2

original	10%	30%	50%	70%	90%
79.96	78.69	78.09	73.75	71.21	69.67

Extrinsic Evaluation Love+FastText on CoNLL-03

original	10%	30%	50%	70%	90%
83.4	80.33	76.21	72.19	66.17	63.1

Author's Results:

	SST2						**CoNLL-03**						
Typo Probability	original	10%	30%	50%	70%	90%	original	10%	30%	50%	70%	90%	Avg
Static Embeddings													
FastText	**82.3**	68.2	59.8	56.7	57.8	60.3	**86.4**	81.6	78.9	73.9	70.2	63.4	70.0
FastText + LOVE	82.1	**79.8**	**74.9**	**74.2**	**68.8**	**67.2**	86.3	**84.7**	**81.8**	**77.5**	**73.1**	**71.3**	**76.8**

Key findings

- How to use LOVE+FastText in a plug-and-play fashion
- The model outperformed MIMICK-like models in intrinsic evaluation tasks and in extrinsic evaluation for SST2 and MR datasets
- Probabilities of word augmentation
  - Not specified the probability used for each specific word augmentation
  - Size of the synonym file from which *synonym augmentation* was extracted

Thank you!

Andrea Carotti: acarot2@uic.edu

Emilio Ingenito: eingen2@uic.edu