Encouraging Paragraph Embeddings to Remember Sentence Identity Improves Classification

Tu Vu, Mohit Iyyer

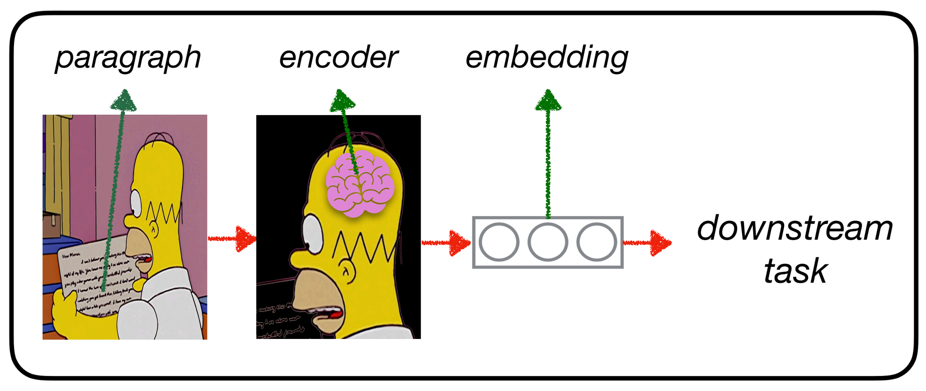
code at github.com/tuvuumass/scope

What are paragraph embeddings?

Encode a given paragraph into **a single fixed-length vector representation**

Applications

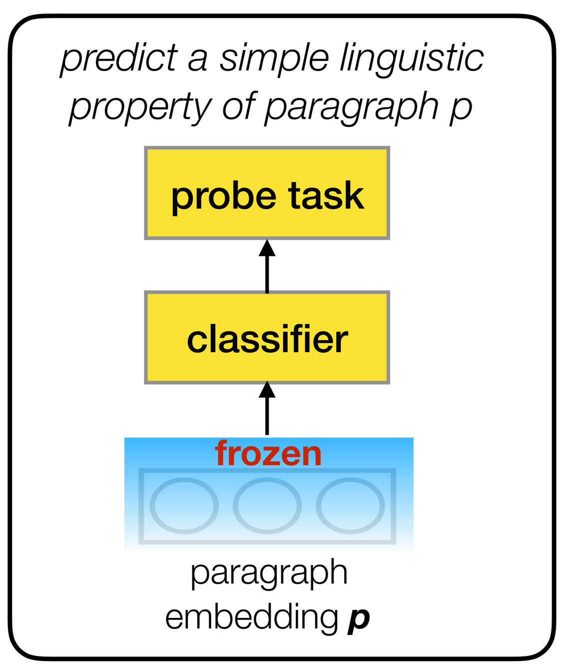
★ text classification
★ document retrieval
★ semantic similarity and relatedness



How can we examine what linguistic properties they encode?

Linguistic Probe Tasks

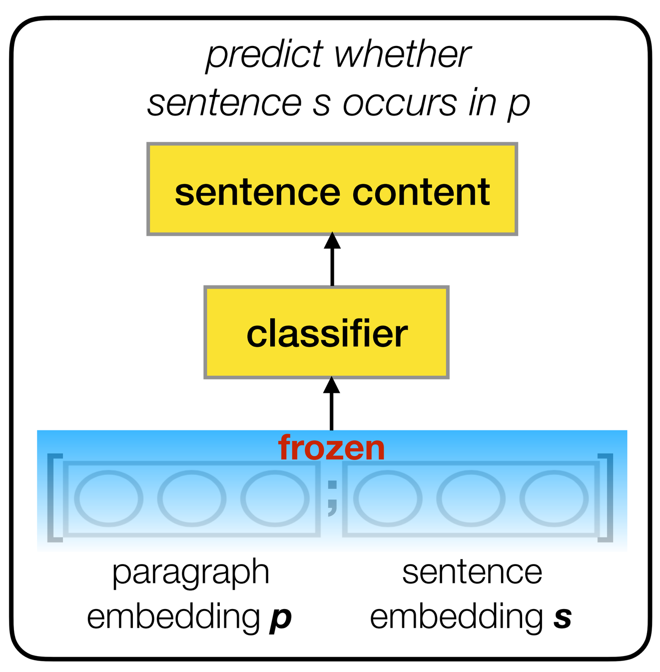
extended to the paragraph level



★ classification tasks
★ agnostic to the embedding method

Sentence Content

binary classification



★ positive instances: [**p**; **s**⁺], s^+ from p
★ negative instances: [**p**; **s**⁻], s^- from another paragraph p'

Motivation: word identity information is correlated with downstream sentence-level classification performance (Conneau et al., 2018)

How well do they encode the identity of the sentences within a paragraph?

Probe data

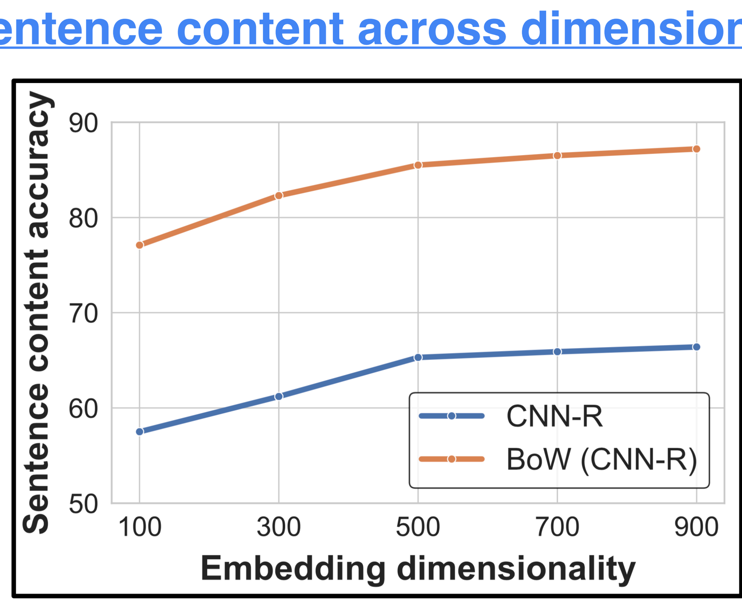
Hotel Reviews (Li et al., 2015; Zhang et al., 2017): 340K/20K/20K paragraphs for train/val/test

Paragraph Embedding Models

★ **CNN-R**, originally **CNN-DCNN** (Zhang et al., 2017):
- *convolutional-deconvolutional encoder-decoder* model + *reconstruction objective*
- powerful paragraph embeddings

★ **BOW (CNN-R)**
- *average of CNN-R's word vectors*

BoW (CNN-R) outperforms CNN-R on sentence content across dimensions



BoW models outperform more complex models on sentence content

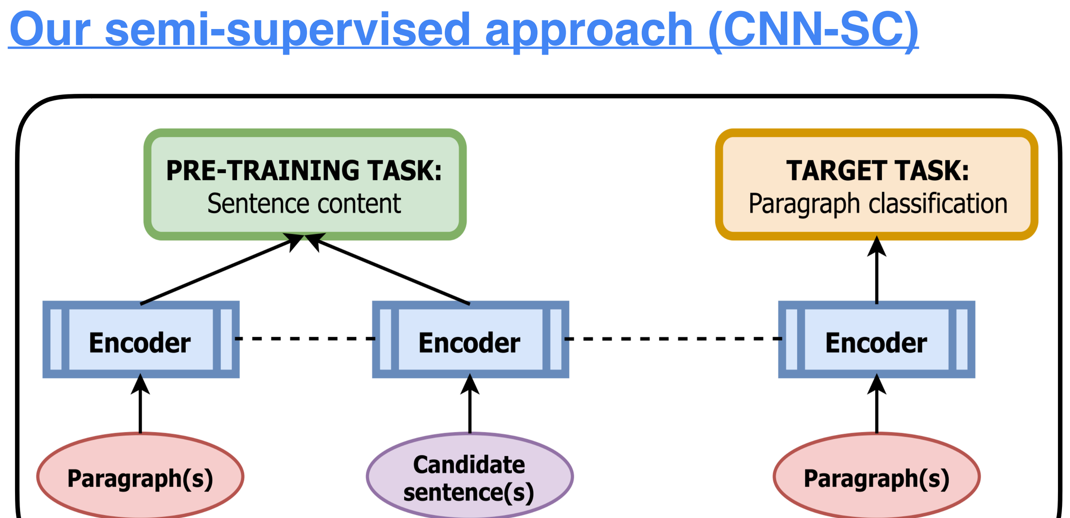
Model	Dimensionaltity	Accuracy
Random	—	50.0
trained on paragraphs from Hotel Reviews		
BoW (CNN-R)	900	**87.2**
Doc2VecC	900	**90.8**
CNN-R	900	66.4
LSTM-R	900	65.4
pre-trained on other datasets		
BOW (Glove)	300	**84.6**
BOW (ELMo)	1024	**88.1**
Skip-Thoughts	4800	78.9
InferSent	4096	68.7

BoW (CNN-R) relies more heavily on low-level matching than CNN-R

Setting	CNN-R	BOW (CNN-R)
Without s^+ excluded from p	61.2	**82.3**
With s^+ excluded from p	57.5	**61.7**

Sentence content as a pretraining task

Our semi-supervised approach (CNN-SC)



Classification tasks and datasets

Dataset	Type	# classes	# examples
Yelp	Sentiment	2	560K
DBPedia	Topic	14	560K
Yahoo	Topic	10	1.4M

Sentence content substantially boosts accuracy and generalization, outperforming reconstruction

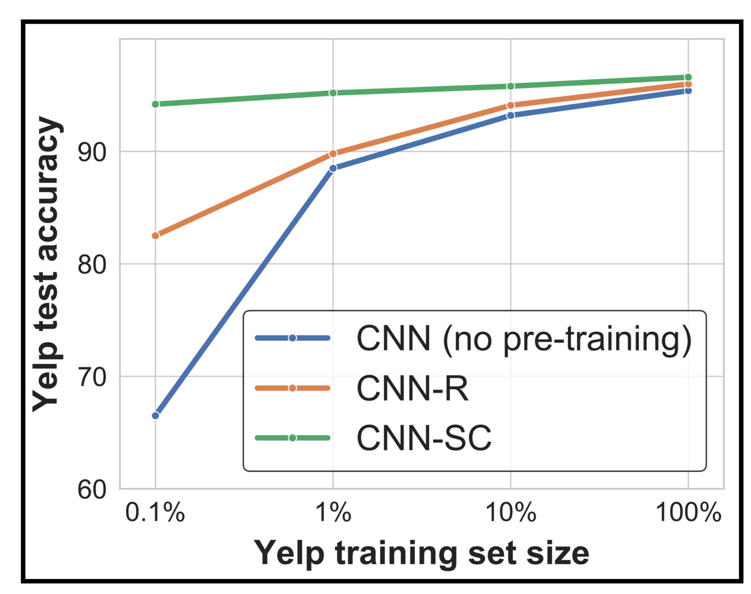
Without fine-tuning, CNN-SC outperforms CNN-R by a large margin on both in-domain and out-of-domain data

Pre-training	CNN-R	CNN-SC
On Yelp	67.4	**90.0**
On Wikipedia	61.4	**65.7**
Wall-clock speedup	1X	**4X**

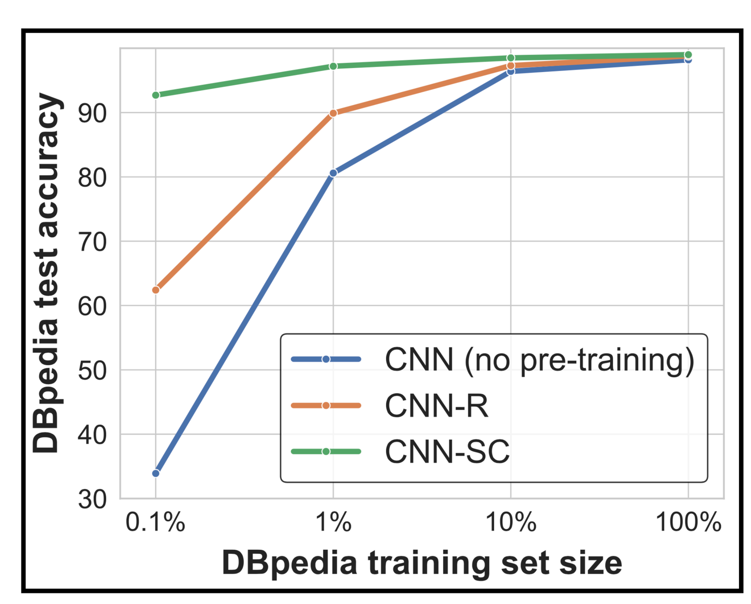
Yelp test accuracy

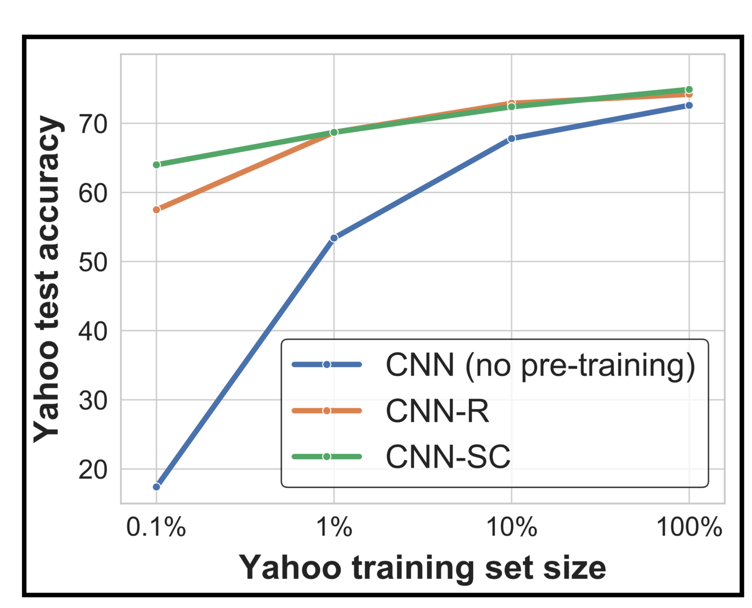
★ four times faster to train
★ better correlation to downstream accuracy

Fine-tuning CNN-SC substantially boosts accuracy and generalization



On Yelp, with only 500 labeled examples, it outperforms training from scratch on 200× more data

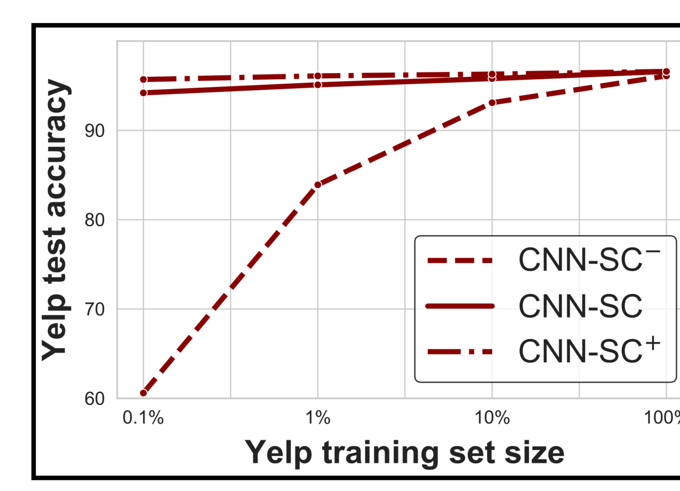




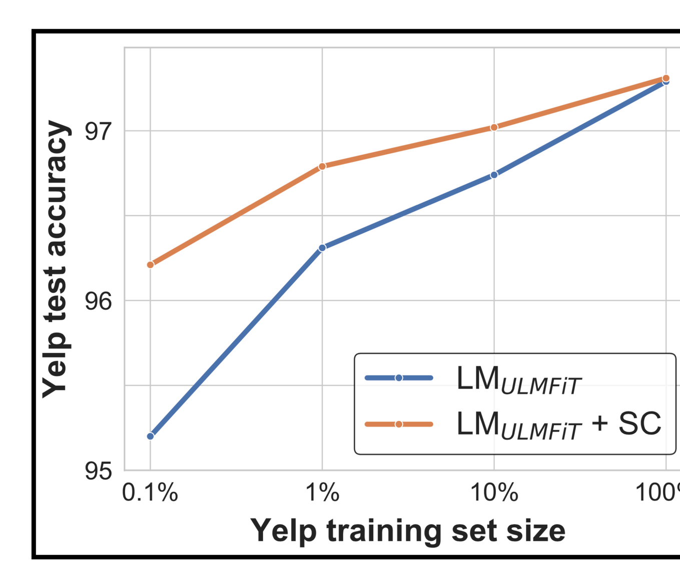
CNN-SC outperforms baseline models that do not use external data, including CNN-R

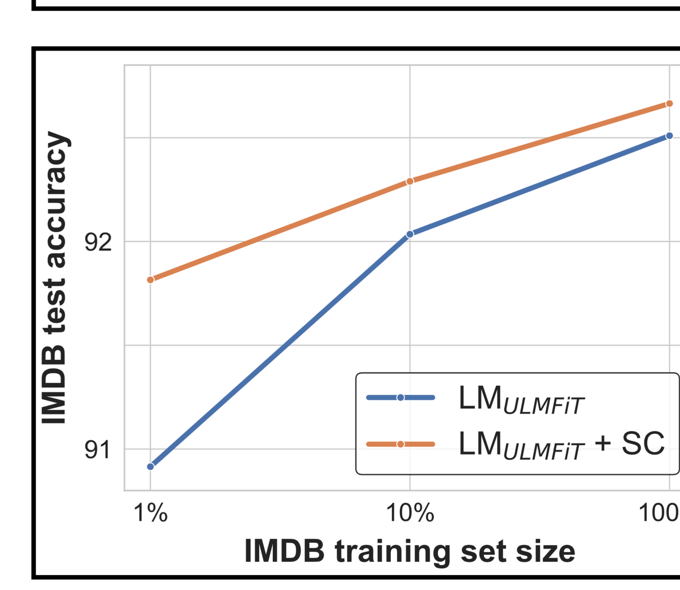
Model	Yelp	DBPedia	Yahoo
purely supervised w/o external data			
ngrams TFIDF	95.4	98.7	68.5
Large Word ConvNet	95.1	98.3	70.9
Small Word ConvNet	94.5	98.2	70.0
Large Char ConvNet	94.1	98.3	70.5
Small Char ConvNet	93.5	98.0	70.2
SA-LSTM (word level)	NA	98.6	NA
Deep ConvNet	95.7	98.7	73.4
CNN (Zhang et al., 2017)	95.4	98.2	72.6
pre-training + fine-tuning w/o external data			
CNN-R (Zhang et al., 2017)	96.0	98.8	74.2
CNN-SC (ours)	**96.6**	**99.0**	**74.9**
pre-training + fine-tuning w/ external data			
ULMFiT (Howard and Ruder, 2018)	97.8	99.2	NA

CNN-SC implicitly learns to distinguish between class labels



Sentence content learns complementary information to language modeling (LM)





Conclusions

★ BoW models outperform more complex models on our sentence content probe
★ Incorporating probe objectives into downstream models might help improve performance
★ Future work: more linguistically-informed research into embedding methods

References

★ Jiwei Li, Thang Luong, and Dan Jurafsky. 2015. A hierarchical neural autoencoder for paragraphs and documents. In ACL, pages 1106–1115.

★ Yizhe Zhang, Dinghan Shen, Guoyin Wang, Zhe Gan, Ricardo Henao, and Lawrence Carin. 2017. Deconvolutional paragraph representation learning. In NeurIPS, pages 4169–4179.

★ Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In ACL, pages 328–339.

★ Alexis Conneau, German Kruszewski, Guillaume Lample, Loıc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In ACL, pages 2126–2136.