Universal Sentence Encoder

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan,Chris Tar, Yun-Hsuan Sung, Brian Strope, Ray Kurzweil

February 28, 2019

Transfer Learning

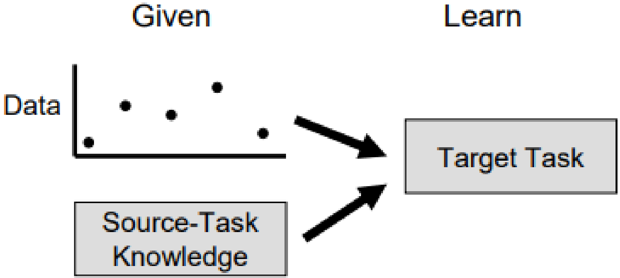


Figure: Transfer learning is machine learning with an additional source of information apart from the standard training data: knowledge from one or more related tasks.

Transfer Learning

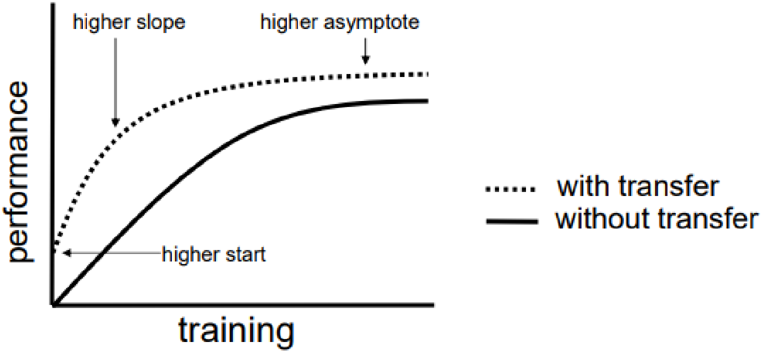
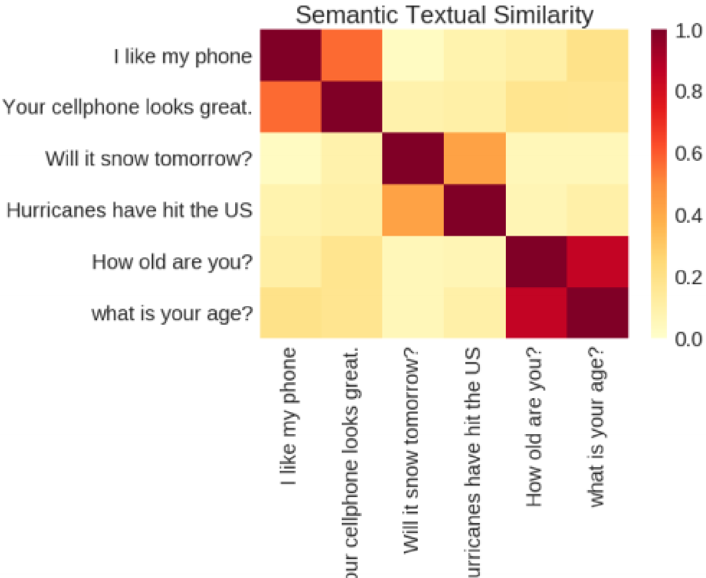


Figure: Three ways in which transfer might improve learning.

Transfer Learning in NLP tasks

- **transfer learning using sentence embeddings**
- transfer learning using word embeddings

Sentence Encoder Models



Sentence Encoder Models

- Transformer model
  - high accuracy
  - greater model complexity
  - greater resource consumption
- Deep Averaging Network (DAN) model
  - efficient inference
  - slightly reduced accuracy

Sentence Encoder Models

Transformer

Constructs sentence embeddings using the encoding sub-graph of the transformer architecture (Vaswani et al., 2017).

- input: PTB tokenized string.
- 1) compute context aware representations of words (ordering and identity)
- 2) representations (1) are converted to a fixed length vector (sentence encoding): element-wise sum of the representations at each word position
- output: 512 dimensional vector as the sentence embedding.

Sentence Encoder Models

Deep Averaging Network (DAN)

Makes use of a deep averaging network (DAN) (Iyyer et al., 2015).

- input: PTB tokenized string.
- 1) embeddings for words and bi-grams are averaged together
- 2) (1) passed through a feedforward deep neural network (DNN)
- output: 512 dimensional vector as the sentence embedding.

Transfer Tasks. Transfer Learning Models

- sentence classification tasks: DNN
- pairwise semantic similarity task: $sim(u, v) = (1 - arccos(\frac{u \cdot v}{\|u\| \|v\|})/\pi)$

Transfer Tasks. Transfer Learning Models

Baselines

- sentence + word level transfer
  - DNN
    - DAN model encoder
    - Transformer model encoder
  - CNN
    - DAN model encoder
    - Transformer model encoder
- sentence level transfer
  - DNN
    - DAN model encoder
    - Transformer model encoder
  - CNN
    - DAN model encoder
    - Transformer model encoder
- word level transfer
  - DNN
  - CNN
- no transfer
  - DNN
  - CNN

Model performance on transfer tasks

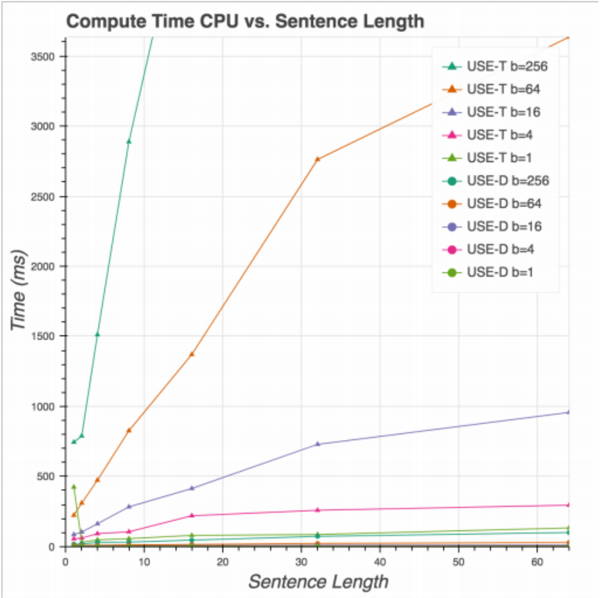
Model	MR	CR	SUBJ	MPQA	TREC	SST	STS Bench (dev / test)
Sentence & Word Embedding Transfer Learning							
USE_D+DAN (w2v w.e.)	77.11	81.71	93.12	87.01	94.72	82.14	–
USE_D+CNN (w2v w.e.)	78.20	82.04	93.24	85.87	97.67	85.29	–
USE_T+DAN (w2v w.e.)	81.32	86.66	93.90	88.14	95.51	86.62	–
USE_T+CNN (w2v w.e.)	81.18	87.45	93.58	87.32	98.07	86.69	–
Sentence Embedding Transfer Learning							
USE_D	74.45	80.97	92.65	85.38	91.19	77.62	0.763 / 0.719 (r)
USE_T	81.44	87.43	93.87	86.98	92.51	85.38	0.814 / 0.782 (r)
USE_D+DAN (lrn w.e.)	77.57	81.93	92.91	85.97	95.86	83.41	–
USE_D+CNN (lrn w.e.)	78.49	81.49	92.99	85.53	97.71	85.27	–
USE_T+DAN (lrn w.e.)	81.36	86.08	93.66	87.14	96.60	86.24	–
USE_T+CNN (lrn w.e.)	81.59	86.45	93.36	86.85	97.44	87.21	–
Word Embedding Transfer Learning							
DAN (w2v w.e.)	74.75	75.24	90.80	81.25	85.69	80.24	–
CNN (w2v w.e.)	75.10	80.18	90.84	81.38	97.32	83.74	–
Baselines with No Transfer Learning							
DAN (lrn w.e.)	75.97	76.91	89.49	80.93	93.88	81.52	–
CNN (lrn w.e.)	76.39	79.39	91.18	82.20	95.82	84.90	–

Task performance on SST for varying amounts of training data

Model	SST 1k	SST 2k	SST 4k	SST 8k	SST 16k	SST 32k	SST 67.3k
Sentence & Word Embedding Transfer Learning							
USE_D+DNN (w2v w.e.)	78.65	78.68	79.07	81.69	81.14	81.47	82.14
USE_D+CNN (w2v w.e.)	77.79	79.19	79.75	82.32	82.70	83.56	85.29
USE_T+DNN (w2v w.e.)	85.24	84.75	85.05	86.48	86.44	86.38	86.62
USE_T+CNN (w2v w.e.)	84.44	84.16	84.77	85.70	85.22	86.38	86.69
Sentence Embedding Transfer Learning							
USE_D	77.47	76.38	77.39	79.02	78.38	77.79	77.62
USE_T	84.85	84.25	85.18	85.63	85.83	85.59	85.38
USE_D+DNN (lrn w.e.)	75.90	78.68	79.01	82.31	82.31	82.14	83.41
USE_D+CNN (lrn w.e.)	77.28	77.74	79.84	81.83	82.64	84.24	85.27
USE_T+DNN (lrn w.e.)	84.51	84.87	84.55	85.96	85.62	85.86	86.24
USE_T+CNN (lrn w.e.)	82.66	83.73	84.23	85.74	86.06	86.97	87.21
Word Embedding Transfer Learning							
DNN (w2v w.e.)	66.34	69.67	73.03	77.42	78.29	79.81	80.24
CNN (w2v w.e.)	68.10	71.80	74.91	78.86	80.83	81.98	83.74
Baselines with No Transfer Learning							
DNN (lrn w.e.)	66.87	71.23	73.70	77.85	78.07	80.15	81.52
CNN (lrn w.e.)	67.98	71.81	74.90	79.14	81.04	82.72	84.90

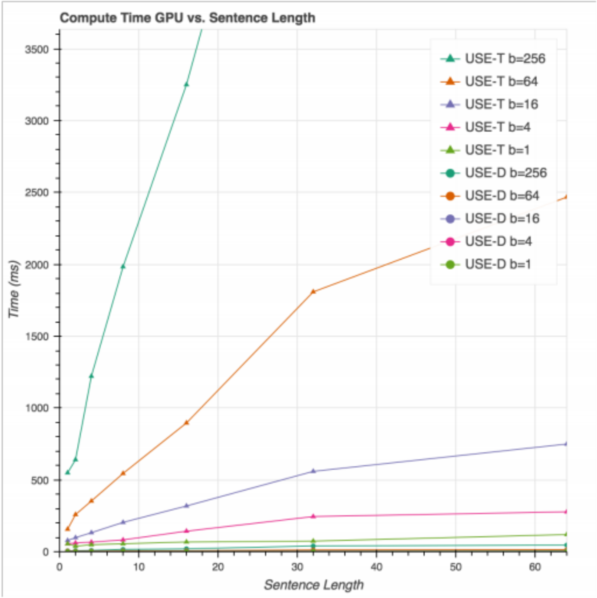
Resource Usage

Compute Usage



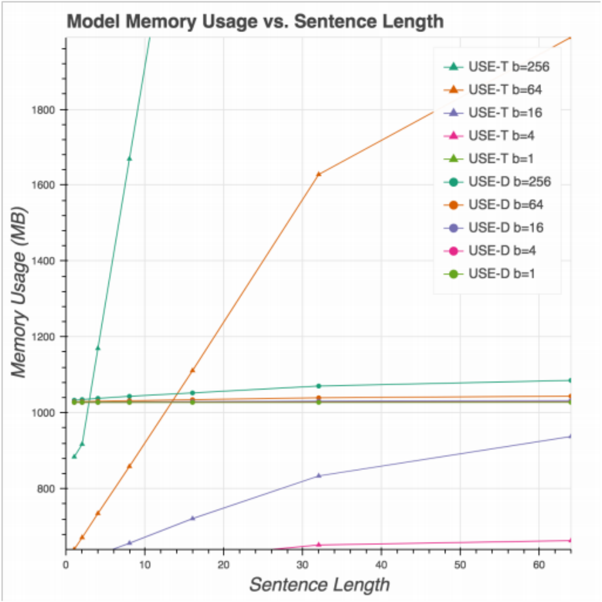
Resource Usage

Compute Usage



Resource Usage

Memory Usage



Conclusion

- 1) sentence level embeddings are better tnan only word level embeddings
- 2) sentence level + word level embeddings are even better than 1)
- 3) transfer learning is most helpful when limited training data is available
- 4) the encoding models make different trade-offs (accuracy, model complexity) that should be considered when choosing a model for a particular application