Universal Sentence Encoder

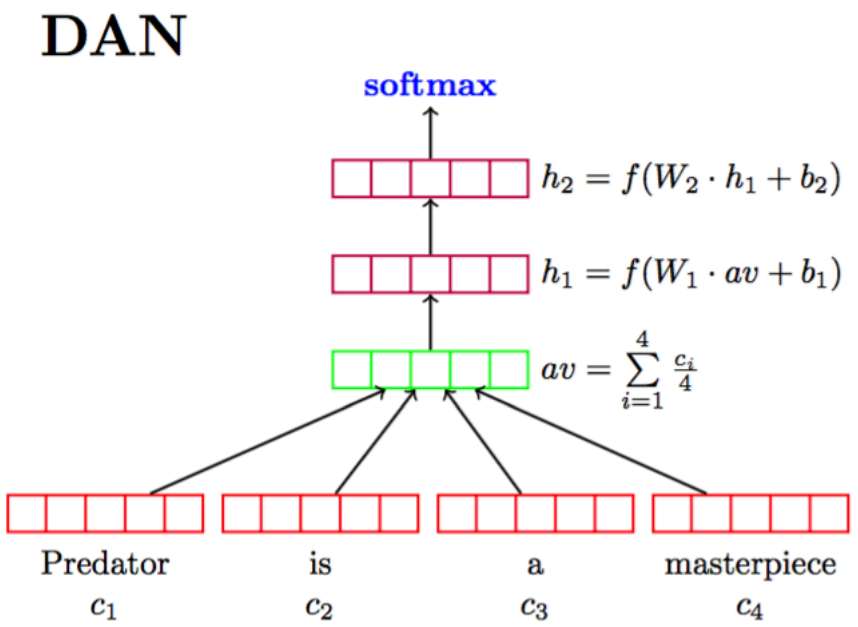
Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brain Strope, Ray Kurzeweil

Google

Method

Present two models for producing sentence embeddings that demonstrate good transfer to other NLP tasks

- transformer based
  - compute the element-wise-sum of the representations at each word position
  - divide by the square root of the length of the sentence so that the differences between short sentences are not dominated by sentence length effects
- deep averaging network (DAN) based
  - input embeddings for words and bi-grams are first averaged together and then passed through a feedforward deep neural network



Training procedure:

multi-task learning: skip-thought like task; conversational input-response task; classification task

Transfer Learning Models

- For sentence classification transfer tasks, the output of the transformer and DAN sentence encoders are provided to a task specific DNN
- For the pairwise semantic similarity task, we directly assess the similarity of the sentence embeddings produced by the two encoders

$$\text{sim}(\mathbf{u}, \mathbf{v}) = \left(1 - \arccos\left(\frac{\mathbf{u} \cdot \mathbf{v}}{||\mathbf{u}||\,||\mathbf{v}||}\right)/\pi\right) \quad (1)$$

- baseline: pretrained word embeddings (word2vec skip-gram) are fed into CNN or DAN.
- combined transfer models: combine the sentence and word level transfer models by concatenating their representations prior to feeding the combined representation to the transfer task classification layers

Experiments

- transformer based sentence encoder > DAN based encoder
- sentence + word level transfer > sentence level transfer > word level transfer

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	–

Table 2: Model performance on transfer tasks. *USE_T* is the universal sentence encoder (USE) usi
Transformer. *USE_D* is the universal encoder DAN model. Models tagged with *w2v w.e.* make use
pre-training word2vec skip-gram embeddings for the transfer task model, while models tagged with *l*
w.e. use randomly initialized word embeddings that are learned only on the transfer task data. Accura
is reported for all evaluations except STS Bench where we report the Pearson correlation of the simila
ity scores with human judgments. Pairwise similarity scores are computed directly using the sentenc
embeddings from the universal sentence encoder as in Eq. (1).

Experiments

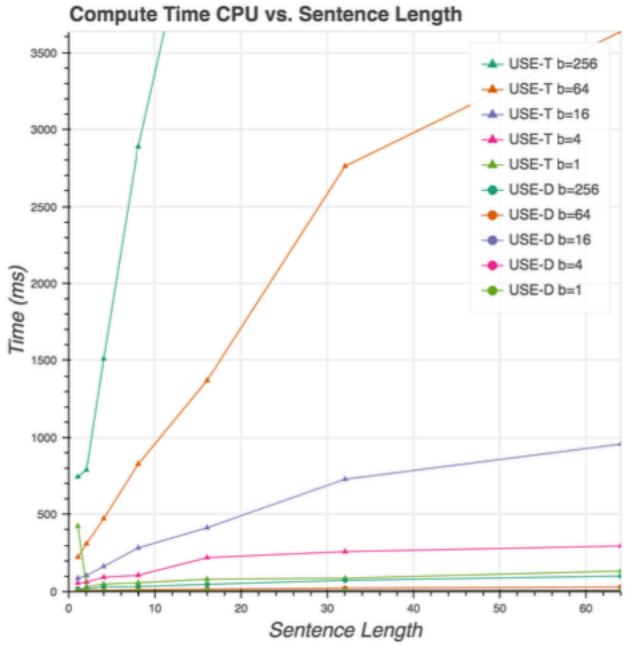
- for smaller quantities of data, sentence level transfer learning can achieve surprisingly good task performance
- as the training set size increases, models that do not make use of transfer learning approach the performance of the other models

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

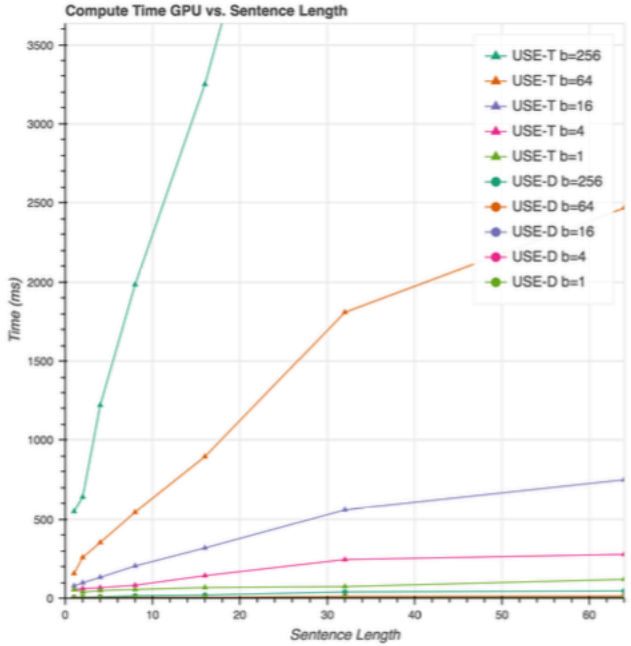
Table 3: Task performance on SST for varying amounts of training data. SST 67.3k represents the full training set. Using only 1,000 examples for training, transfer learning from USE_T is able to obtain performance that rivals many of the other models trained on the full 67.3 thousand example training set.

Experiments

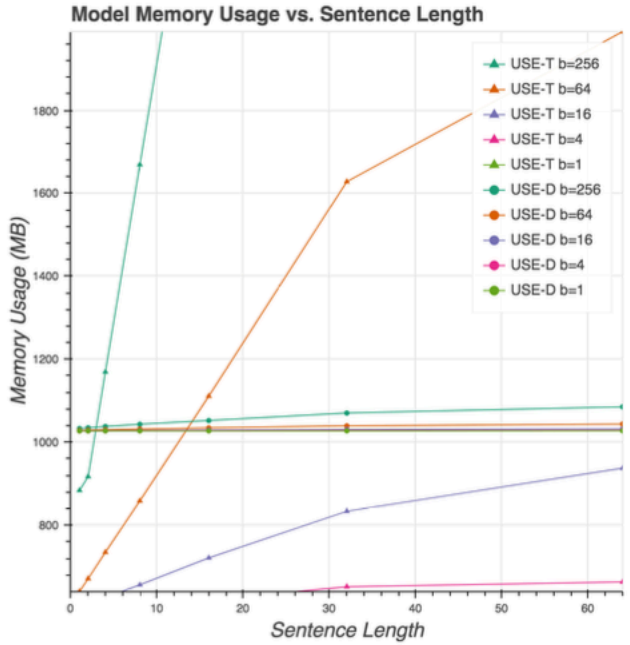
- compute time for transformer increases noticeably as sentence length increases
- the compute time for the DAN model stays nearly constant as sentence length is increased.



(a) CPU Time vs. Sentence Length



(b) GPU Time vs. Sentence Length



(c) Memory vs. Sentence Length