Understanding Neural Networks through Representation Learning

Jiwei Li, Will Monroe and Dan Jurafsky

Motivation

One of the hidden layers

Word embedding

- Black box
  - What does each word vector dimension stand for?
  - What do hidden units in intermediate levels stand for?
  - How does the model combine meaning from different parts of the sentence, filtering the informational wheat from the chaff?
  - How is the final decision made at the output layer?

- If we can answer these,
  - Error analysis may lead to correcting model mistakes.

Methodology

- Several techniques of erasure
  - interpret decisions from a neural model by observing the effects on the model of **erasing** various parts of the representation, such as
    - **input word-vector dimensions**
    - **intermediate hidden units,**
    - **or input words**
- Evaluated in two ways
  - computing its impact on evaluation metrics
    - Compute the log-likelihood difference when a particular dimension is erased
  - using reinforcement learning to erase the minimum set of input words in order to flip a neural model's decision.

Linking Word Vector Dimensions to Linguistics Features

- Train classifier models on benchmarks for
  - POS tagging
  - NER tagging
  - Chunking
  - Prefix and suffix (predicting a prefix/suffix given a word)
  - Sentiment
  - Shape (X, XX, XXX… very easy task that depends on the model to reason on the length of the word)
  - Frequency (regression for the frequency in Wikipedia)
- Then employ our method, explained in the next slide

“Importance” concept

$S(e, c) = -logP(L_e = c)$

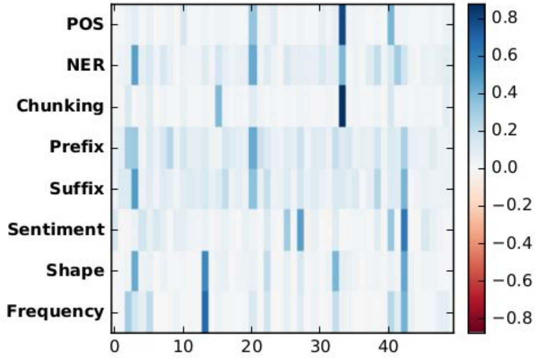
Everything same but
dimension d is erased

$$I(d) = \frac{1}{|E|} \sum_{e \in E} \frac{S(e,c) - S(e,c,\neg d)}{S(e,c)}$$

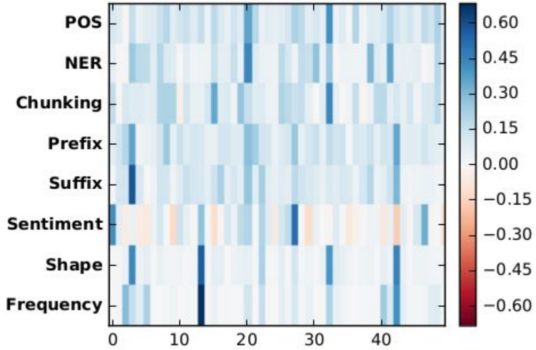
examples

Some details on training and the model

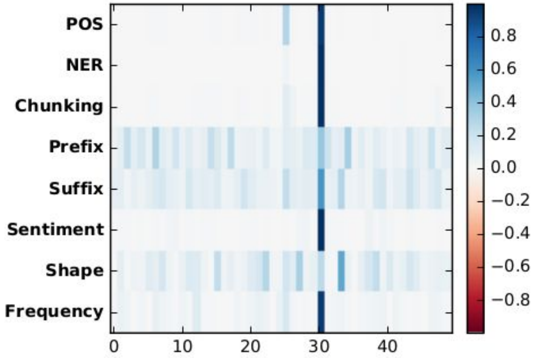
- Train a 50 dimensional embedding on Gigaword-Wiki corpus
  - Word2vec
  - GloVe
- 4-layer NN
  - Input, 2 intermediate and output layer
  - 50 neurons for each



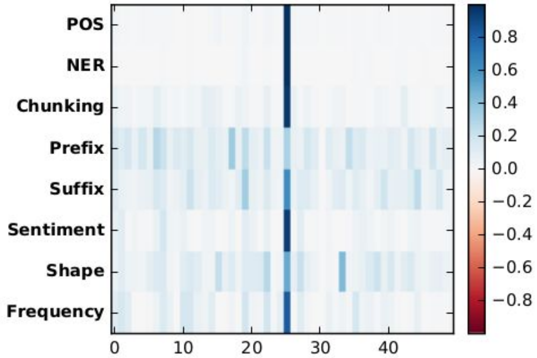
(a) Word2vec, no dropout.



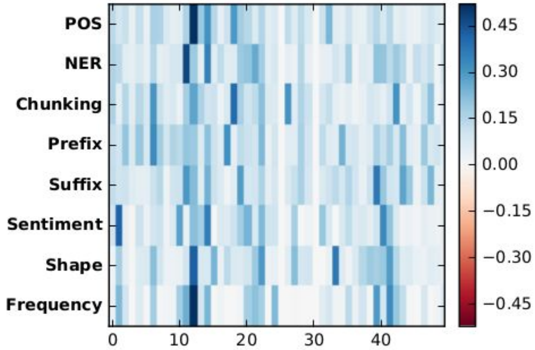
(b) Word2vec, with dropout.



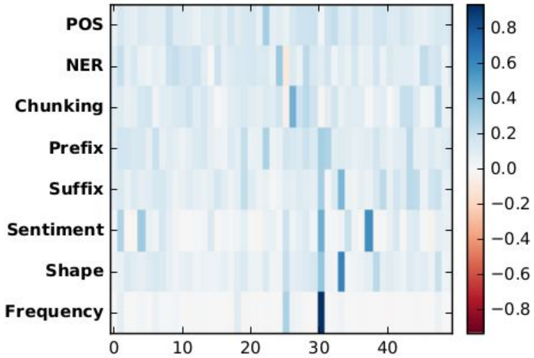
(c) GloVe, no dropout.



(d) GloVe, no dropout; 31rd dimension removed.



(e) GloVe, no dropout; 31rd, 26th dimensions removed.



(f) GloVe, with dropout.

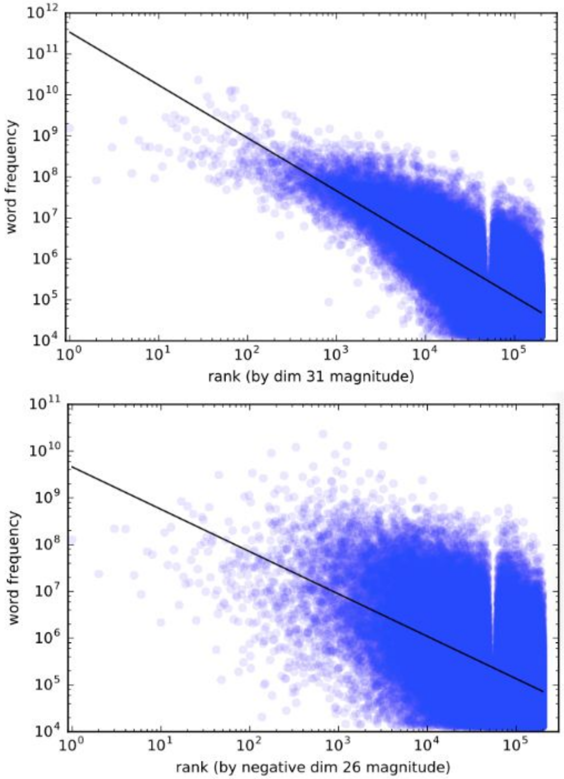
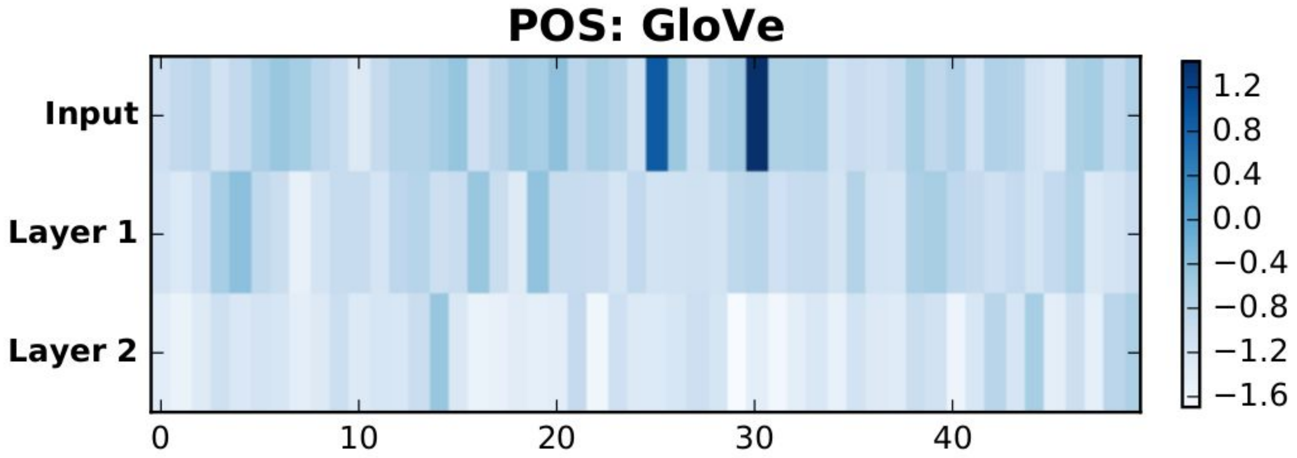


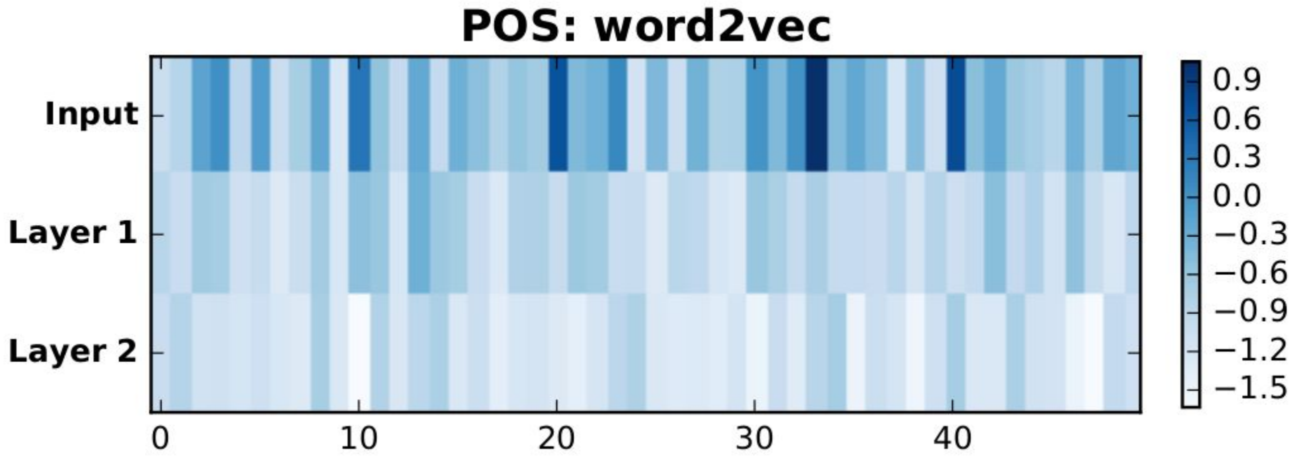
Figure 3: Correlation with word frequency of the magnitude of (a) the 31st dimension ($R^2 = 0.55$, $p < 1 \times 10^{-5}$) and (b) the 26th dimension ($R^2 = 0.27$, $p < 1 \times 10^{-5}$) of GloVe vectors.

Compare Word2vec and GloVe on this aspect

…

From the paper: “presumably because tokens are omitted in proportion to word-type frequency in word2vec models”





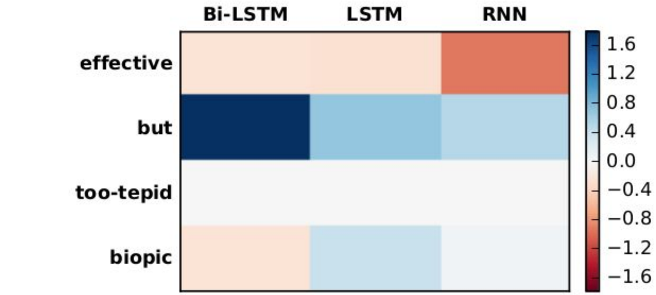
Paper says “this indicates robustness”

…

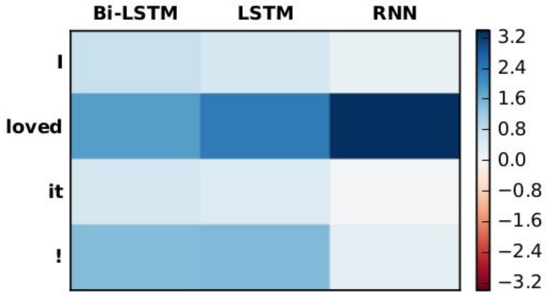
I ask: What if we choose more than one dimension at once? From different layers?

Word level

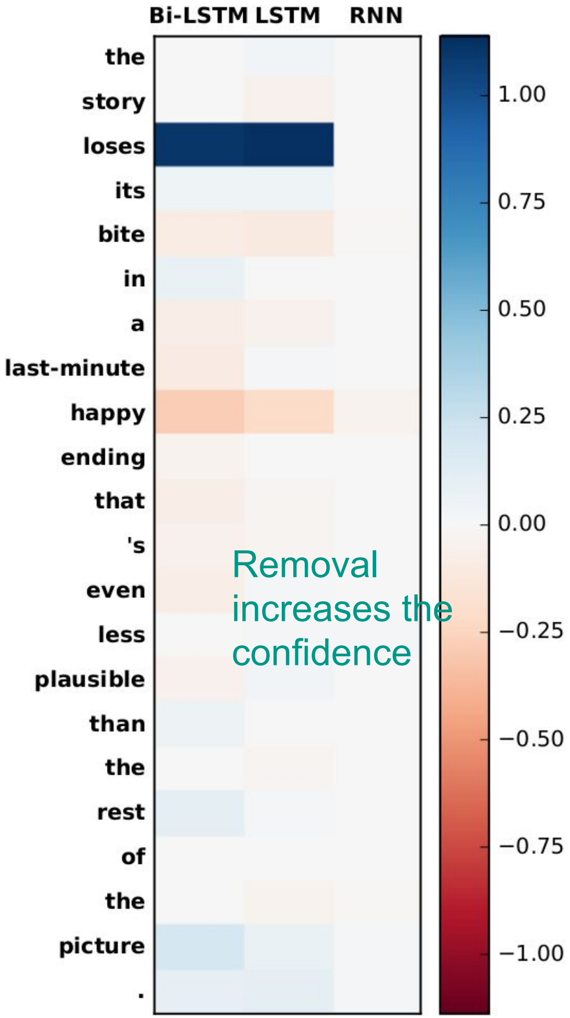
- Stanford Sentiment Treebank dataset
- Same as before but now we **erase word units**
- **Three models**
  - **RNN**
  - **LSTM**
  - **Bi-LSTM**

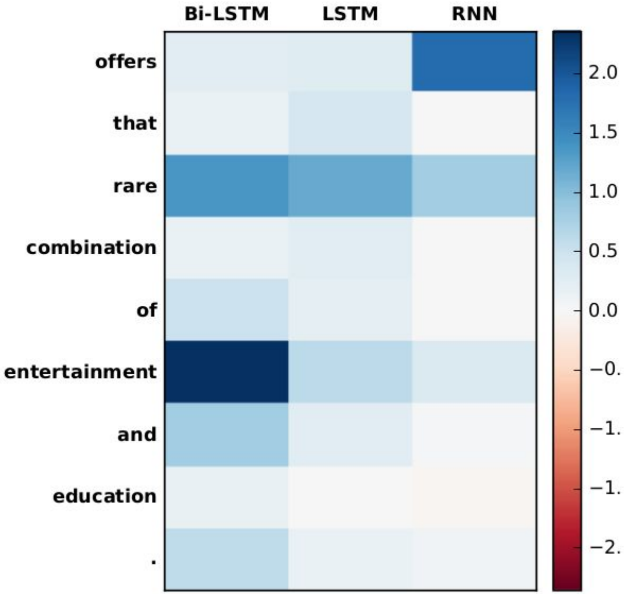


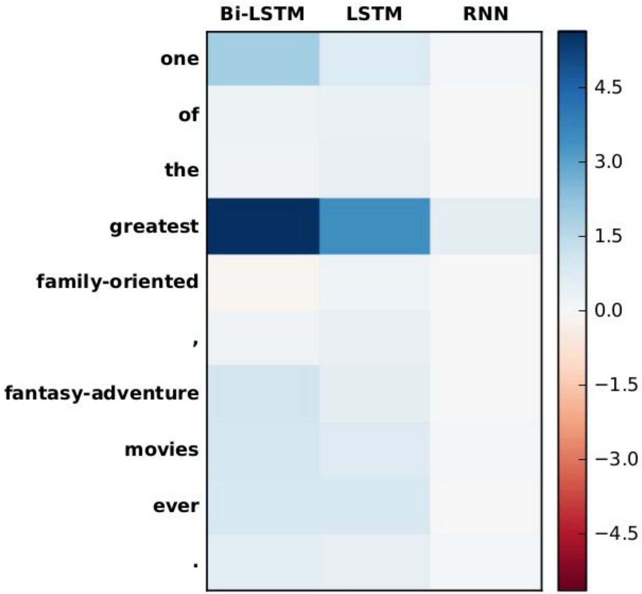
(a) Neutral



(b) Strong positive







Reinforcement Learning

Objective:

$$\min_{D} |D| \quad s.t.\ L_{e-D} \neq L_e$$

- Naive method
  - Enumerate all combinations of word units and try. Intractable.
- Proposal: Reinforcement learning based method
  - Reward function

$$J(\theta) = \mathbb{E}_{\pi}(R(e)|\theta)$$

$$L(e, D) = \frac{1}{|D|} \cdot \mathbf{1}(L_{e-D} \neq L_e)$$

$$\Omega(e, z) = \gamma \sum_{s \in S} \sum_{t \in s} |z_t - z_{t-1}|$$

$$R(e) = L(e, D) - \Omega(z_{1:N})$$

(1) clean updated room. friendly efficient staff . rate was too high 199 plus they charged 10 day for internet access in the room .
(2) the location is fantastic. the staff are helpful and service oriented . sleeping rooms meeting rooms and public lavatories not cleaned on a daily basis . the hotel seems a bit old and a bit tired overall . trolley noise outside can go into the wee hours . if you get a great price for a few nights this hotel may be a good choice . breakfast is very nice remember if you just stick to the cold buffet it is cheaper .
(3) location is nice . but goes from bad to worse once you walk through the door . staff very surly and unhelpful . room and hallway had a very strange smell . rooms very run down . so bad that i checked out immediately and went to another hotel . intercontinental chain should be ashamed .
(4) i took my daughter and her step sister to see a show at webster hall . it is so overpriced i ’m in awe . i felt safe . the rooms were tiny . lots of street noise all night from the partiers at the ale house below .

(a) Examples of minimal set of erased words based on *Bi-LSTM* model

(1) clean updated room. friendly efficient staff . rate was too high 199 plus they charged 10 day for internet access in the room .
the location is fantastic. the staff are helpful and service oriented . (2) sleeping rooms meeting rooms and public lavatories not cleaned on a daily basis . the hotel seems a bit old and a bit tired overall . trolley noise outside can go into the wee hours . if you get a great price for a few nights this hotel may be a good choice . breakfast is very nice remember if you just stick to the cold buffet it is cheaper .
(3) location is nice . but goes from bad to worse once you walk through the door . staff very surly and unhelpful . room and hallway had a very strange smell . rooms very run down . so bad that i checked out immediately and went to another hotel . intercontinental chain should be ashamed .
(4) i took my daughter and her step sister to see a show at webster hall . it is so overpriced i ’m in awe . i felt safe . the rooms were tiny . lots of street noise all night from the partiers at the ale house below .

References

- Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199.
- Anh Nguyen, Jason Yosinski, and Jeff Clune. 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 427–436. IEEE.