Beyond Word Importance: Contextual Decomposition to Extract Interactions from LSTMs

Presenter: Arshdeep Sekhon
https://qdata.github.io/deep2Read

W. James Murdoch, Peter J. Liu, Bin Yu

July 2019

Introduction

- LSTM interpretation model
- extracts information about not only which words contributed to an LSTM's prediction
- also how they were combined in order to yield the final prediction
- mathematically decomposing the LSTM's output, able to disambiguate the contributions made at each step by different parts of the sentence.

Method: LSTM Decomposition

LSTM equations:

$$o_t = \sigma(W_o x_t + V_o h_{t-1} + b_o) \tag{1}$$
$$f_t = \sigma(W_f x_t + V_f h_{t-1} + b_f) \tag{2}$$
$$i_t = \sigma(W_i x_t + V_i h_{t-1} + b_i) \tag{3}$$
$$g_t = \tanh(W_g x_t + V_g h_{t-1} + b_g) \tag{4}$$
$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \tag{5}$$
$$h_t = o_t \odot \tanh(c_t) \tag{6}$$

After processing the full sequence, the final state h_T used as input to a linear layer +SoftMax (multinomial logistic regression), to return a probability distribution p over C classes, with

$$p_j = \mathrm{SoftMax}(Wh_T)_j = \frac{\exp(W_j h_T)}{\sum_{k=1}^{C} \exp(W_k h_t)} \tag{7}$$

Contextual Decomposition

Given an arbitrary phrase $x_q, ..., x_r$, where $1 \leq q \leq r \leq T$, decompose each output h_t and cell state c_t

$$h_t = \beta_t + \gamma_t \tag{8}$$

$$c_t = \beta_t^c + \gamma_t^c \tag{9}$$

β_t corresponds to contributions made solely by the given phrase to h_t, and that γ_t corresponds to contributions involving, at least in part, elements outside of the phrase. Similarly, β_t^c and γ_t^c. final output state Wh_T :

$$p = \text{SoftMax}(W\beta_T + W\gamma_T) \tag{10}$$

Contextual Decomposition

$$
\begin{aligned}
i_t &= \sigma(W_i x_t + V_i h_{t-1} + b_i) && (11)\\
&= L_\sigma(W_i x_t) + L_\sigma(V_i h_{t-1}) + L_\sigma(b_i) && (12)
\end{aligned}
$$

$$
\begin{aligned}
f_t \odot c_{t-1} =& (L_\sigma(W_f x_t) + L_\sigma(V_f \beta_{t-1}) + L_\sigma(V_f \gamma_{t-1}) + L_\sigma(b_f)) \odot (\beta^c_{t-1} + && (13)\\
=& ([L_\sigma(W_f x_t) + L_\sigma(V_f \beta_{t-1}) + L_\sigma(b_f)] \odot \beta^c_{t-1}) && (14)\\
& + (L_\sigma(V_f \gamma_{t-1}) \odot \beta^c_{t-1} + f_t \odot \gamma^c_{t-1}) && \\
=& \beta^f_t + \gamma^f_t && (15)
\end{aligned}
$$

Contextual Decomposition

$$
\begin{aligned}
i_t \odot g_t =& [L_\sigma(W_i x_t) + L_\sigma(V_i \beta_{t-1}) + L_\sigma(V_i \gamma_{t-1}) + L_\sigma(b_i)] \qquad (16)\\
& \odot [L_{\text{tanh}}(W_g x_t) + L_{\text{tanh}}(V_g \beta_{t-1}) + L_{\text{tanh}}(V_g \gamma_{t-1}) + L_{\text{tanh}}(b_g)]\\
=& [L_\sigma(W_i x_t) \odot [L_{\text{tanh}}(W_g x_t) + L_{\text{tanh}}(V_g \beta_{t-1}) + L_{\text{tanh}}(b_g)] \qquad (17)\\
& + L_\sigma(V_i \beta_{t-1}) \odot [L_{\text{tanh}}(W_g x_t) + L_{\text{tanh}}(V_g \beta_{t-1}) + L_{\text{tanh}}(b_g)]\\
& + L_\sigma(b_i) \odot [L_{\text{tanh}}(W_g x_t) + L_{\text{tanh}}(V_g \beta_{t-1})]]\\
& + [L_\sigma(V_i \gamma_{t-1}) \odot g_t + i_t \odot L_{\text{tanh}}(V_g \gamma_{t-1}) - L_\sigma(V_i \gamma_{t-1}) \odot L_{\text{tanh}}(\\
& + L_\sigma(b_i) \odot L_{\text{tanh}}(b_g)]\\
=& \beta_t^u + \gamma_t^u \qquad (18)
\end{aligned}
$$

Cotextual Decomposition

$$\beta_t^c = \beta_t^f + \beta_t^u \tag{19}$$

$$\gamma_t^c = \gamma_t^f + \gamma_t^u \tag{20}$$

$$
\begin{aligned}
h_t &= o_t \odot \tanh(c_t) && (21)\\
&= o_t \odot [L_{\tanh}(\beta_t^c) + L_{\tanh}(\gamma_t^c)] && (22)\\
&= o_t \odot L_{\tanh}(\beta_t^c) + o_t \odot L_{\tanh}(\gamma_t^c) && (23)\\
&= \beta_t + \gamma_t && (24)
\end{aligned}
$$

Linearization of Activation functions

$$g_t = tanh(W_g x_t + V_g h_{t-1} + b_g) \tag{25}$$

Required:

$$g_t = L_{tanh}(W_g x_t) + L_{tanh}(V_g h_{t-1}) + L_{tanh}(b_g) \tag{26}$$

$$tanh(\sum y_i) = (\sum L_{tanh}(y_i)) \tag{27}$$

Linearization of Activation functions

summarization of partial sums as a linearization technique if $y_1, \ldots, y_n$ are ordered

$$L'_{tanh}(y_k) = tanh(\sum_{j=1}^{k} y_j) - tanh(\sum_{j=1}^{k-1} y_j) \tag{28}$$

But no ordering, compute an average over all orderings

$$L_{\text{tanh}}(y_k) = \frac{1}{M_N} \sum_{i=1}^{M_N} [\text{tanh}(\sum_{j=1}^{\pi_i^{-1}(k)} y_{\pi_i(j)}) - \text{tanh}(\sum_{j=1}^{\pi_i^{-1}(k)-1} y_{\pi_i(j)})] \tag{29}$$

This linearization technique is an approximation to the Shapley Values.(?)

Experiments: Stanford Sentiment Tree Bank

- Unigram Word Scores: Correlation with the logistic regression coefficient

Attribution Method	Heat Map
Gradient	It's easy to love Robin Tunney – she's pretty and she can act – but it gets harder and harder to understand her choices.
Leave one out (Li et al., 2016)	It's easy to love Robin Tunney – she's pretty and she can act – but it gets harder and harder to understand her choices.
Cell decomposition (Murdoch & Szlam, 2017)	It's easy to love Robin Tunney – she's pretty and she can act – but it gets harder and harder to understand her choices.
Integrated gradients (Sundararajan et al., 2017)	It's easy to love Robin Tunney – she's pretty and she can act – but it gets harder and harder to understand her choices.
Contextual decomposition	It's easy to love Robin Tunney – she's pretty and she can act – but it gets harder and harder to understand her choices.

Legend: Very Negative | Negative | Neutral | Positive | Very Positive

Table 2: Heat maps for portion of review from SST with different attribution techniques. Only CD captures that the first phrase is positive.

cdpositive

Identifying Dissenting Subphrases

- "used to be my favorite"
- favorite is positive, used to be is negative

Attribution Method	Heat Map								
Gradient	used	to	be	my	favorite	not	worth	the	time
Leave One Out (Li et al., 2016)	used	to	be	my	favorite	not	worth	the	time
Cell decomposition (Murdoch & Szlam, 2017)	used	to	be	my	favorite	not	worth	the	time
Integrated gradients (Sundararajan et al., 2017)	used	to	be	my	favorite	not	worth	the	time
Contextual decomposition	used	to	be	my	favorite	not	worth	the	time

Legend: Very Negative | Negative | Neutral | Positive | Very Positive

Table 1: Heat maps for portion of yelp review with different attribution techniques. Only CD captures that "favorite" is positive.

Contextual Decomposition Captures Negation

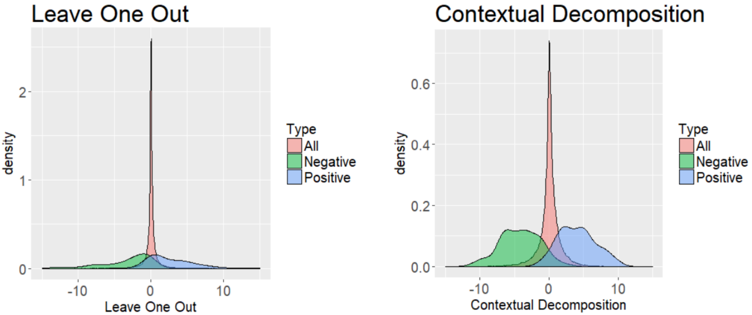

Figure 1: Distribution of scores for positive and negative negation coefficients relative to all interaction coefficients. Only leave one out and CD are capable of producing these interaction scores.

Identifying Similar Phrases

- Compare Dense embeddings β_T average for phrases across the training set and validation set
- Get nearest neghbors

not entertaining	**not bad**	**very funny**	**entertaining**	**bad**
not funny	never dull	well-put-together piece	intelligent	dull
not engaging	n't drag	entertaining romp	engaging	drag
never satisfactory	never fails	very good	satisfying	awful
not well	without sham	surprisingly sweet	admirable	tired
not fit	without missing	very well-written	funny	dreary

Table 3: Nearest neighbours for selected unigrams and interactions using CD embeddings