A Structured Self-Attentive Sentence Embedding

2018.11.30
김보섭

Agenda

1. Why this paper?
2. Abstract
3. Introduction
4. Approach (MODEL, PENALIZATION TERM)
5. Experimental results
6. Conclusion and Discussion

Why this paper?

NLU (Natural Language Understanding), NLP (Natural Language Processing) 등의 분야에서 항상 등장하는 self-attention

- 아래의 paper 뿐만 아니라 관련 논문들에서 자주 언급되고, **자주 사용됨 일종의 meme**

Attention Is All You Need

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin

(Submitted on 12 Jun 2017 (v1), last revised 6 Dec 2017 (this version, v5))

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks in an encoder-decoder configuration. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

Deep contextualized word representations

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, Luke Zettlemoyer

(Submitted on 15 Feb 2018 (v1), last revised 22 Mar 2018 (this version, v2))

We introduce a new type of deep contextualized word representation that models both (1) complex characteristics of word use (e.g., syntax and semantics), and (2) how these uses vary across linguistic contexts (i.e., to model polysemy). Our word vectors are learned functions of the internal states of a deep bidirectional language model (biLM), which is pre-trained on a large text corpus. We show that these representations can be easily added to existing models and significantly improve the state of the art across six challenging NLP problems, including question answering, textual entailment and sentiment analysis. We also present an analysis showing that exposing the deep internals of the pre-trained network is crucial, allowing downstream models to mix different types of semi-supervision signals.

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova

(Submitted on 11 Oct 2018)

We introduce a new language representation model called BERT, which stands for Bidirectional Encoder Representations from Transformers. Unlike recent language representation models, BERT is designed to pre-train deep bidirectional representations by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT representations can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications.
BERT is conceptually simple and empirically powerful. It obtains new state-of-the-art results on eleven natural language processing tasks, including pushing the GLUE benchmark to 80.4% (7.6% absolute improvement), MultiNLI accuracy to 86.7 (5.6% absolute improvement) and the SQuAD v1.1 question answering Test F1 to 93.2 (1.5% absolute improvement), outperforming human performance by 2.0%.

SOTA!

Why this paper?

self-attention은 특정 방법론에 국한되기 보다는 일종의 paradigm으로, 자기자신을 잘 표현하는 방법임

- 크게 보면 "A simple neural network module for relational reasoning"에서 소개하는 relational network의 개념이 self-attention에 녹아들어가 있음

A simple neural network module for relational reasoning

Adam Santoro, David Raposo, David G.T. Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, Timothy Lillicrap

(Submitted on 5 Jun 2017)

Relational reasoning is a central component of generally intelligent behavior, but has proven difficult for neural networks to learn. In this paper we describe how to use Relation Networks (RNs) as a simple plug-and-play module to solve problems that fundamentally hinge on relational reasoning. We tested RN-augmented networks on three tasks: visual question answering using a challenging dataset called CLEVR, on which we achieve state-of-the-art, super-human performance; text-based question answering using the bAbI suite of tasks; and complex reasoning about dynamic physical systems. Then, using a curated dataset called Sort-of-CLEVR we show that powerful convolutional networks do not have a general capacity to solve relational questions, but can gain this capacity when augmented with RNs. Our work shows how a deep learning architecture equipped with an RN module can implicitly discover and learn to reason about entities and their relations.

A Decomposable Attention Model for Natural Language Inference

Ankur P. Parikh, Oscar Täckström, Dipanjan Das, Jakob Uszkoreit

(Submitted on 6 Jun 2016 (v1), last revised 25 Sep 2016 (this version, v2))

We propose a simple neural architecture for natural language inference. Our approach uses attention to decompose the problem into subproblems that can be solved separately, thus making it trivially parallelizable. On the Stanford Natural Language Inference (SNLI) dataset, we obtain state-of-the-art results with almost an order of magnitude fewer parameters than previous work and without relying on any word-order information. Adding intra-sentence attention that takes a minimum amount of order into account yields further improvements.

A Structured Self-attentive Sentence Embedding

Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, Yoshua Bengio

(Submitted on 9 Mar 2017)

This paper proposes a new model for extracting an interpretable sentence embedding by introducing self-attention. Instead of using a vector, we use a 2-D matrix to represent the embedding, with each row of the matrix attending on a different part of the sentence. We also propose a self-attention mechanism and a special regularization term for the model. As a side effect, the embedding comes with an easy way of visualizing what specific parts of the sentence are encoded into the embedding. We evaluate our model on 3 different tasks: author profiling, sentiment classification, and textual entailment. Results show that our model yields a significant performance gain compared to other sentence embedding methods in all of the 3 tasks.

A Structured Self-Attentive Sentence Embedding

Abstract

A STRUCTURED SELF-ATTENTIVE SENTENCE EMBEDDING

Zhouhan Lin[‡◇*] **Minwei Feng**[◇], **Cicero Nogueira dos Santos**[◇], **Mo Yu**[◇],
Bing Xiang[◇], **Bowen Zhou**[◇] **& Yoshua Bengio**[‡†]
[◇]IBM Watson
[‡]Montreal Institute for Learning Algorithms (MILA), Université de Montréal
[†]CIFAR Senior Fellow
lin.zhouhan@gmail.com
{mfeng, cicerons, yum, bingxia, zhou}@us.ibm.com

ABSTRACT

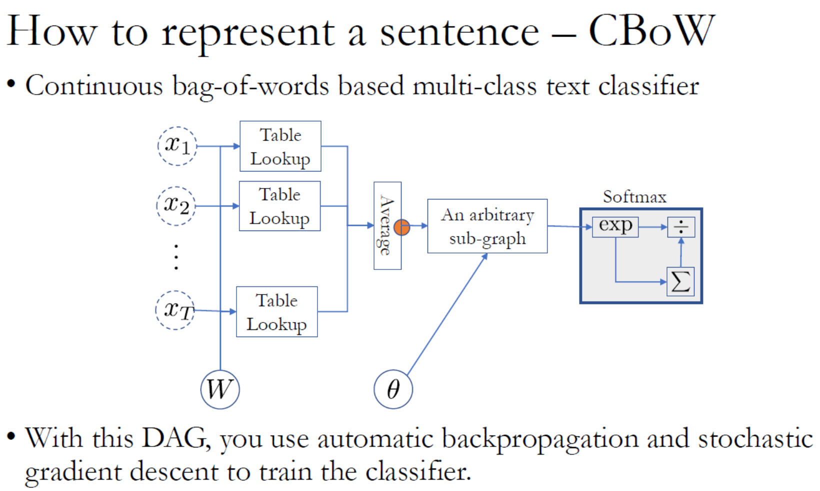
This paper proposes a new model for extracting an interpretable sentence embedding by introducing self-attention. Instead of using a vector, we use a 2-D matrix to represent the embedding, with each row of the matrix attending on a different part of the sentence. We also propose a self-attention mechanism and a special regularization term for the model. As a side effect, the embedding comes with an easy way of visualizing what specific parts of the sentence are encoded into the embedding. We evaluate our model on 3 different tasks: author profiling, sentiment classification and textual entailment. Results show that our model yields a significant performance gain compared to other sentence embedding methods in all of the 3 tasks.

Introduction (1/4)

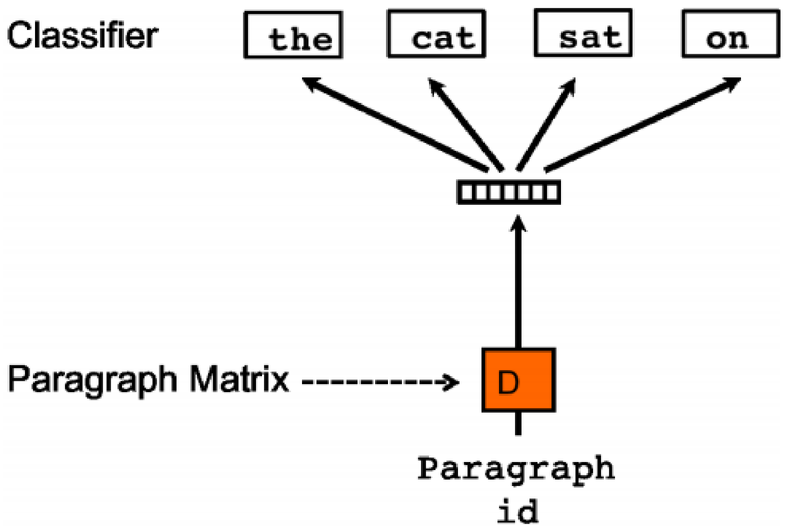
기존의 sentence embedding (not word embedding)은 크게 unsupervised, supervised 형태로 나눌 수 있으며, sentence를 vector로 표현함

- Unsupervised (Task agnostic)
  - ✓ CBOW, ParagraphVector(ie. Doc2vec), SkipThought, etc.

Continous Bag of Words



ParagraphVector



Introduction (2/4)

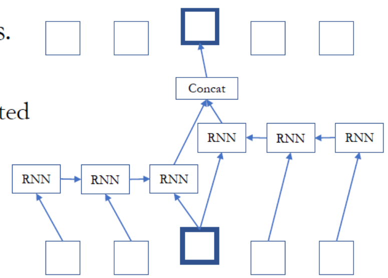
기존의 sentence embedding (not word embedding)은 크게 unsupervised, supervised 형태로 나눌 수 있으며, sentence를 vector로 표현함

- Supervised (Task specific)
  - ✓ Recurrent Neural Network, Convolution Neural Network (eg. 1D), Recursive Neural Network etc.

Recurrent Neural Network

How to represent a sentence – RNN

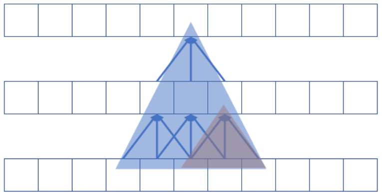
- Recurrent neural network: online compression of a sequence $O(T)$
  $h_t = \mathrm{RNN}(h_{t-1}, x_t)$, where $h_0 = 0$.
- Bidirectional RNN to account for both sides.
- Inherently sequential processing
  - Less desirable for modern, parallelized, distributed computing infrastructure.
- LSTM [Hochreiter&Schmidhuber, 1999] and GRU [Cho et al., 2014] have become de facto standard
  - All standard frameworks implement them.
  - Efficient GPU kernels are available.



Convolution Neural Network

How to represent a sentence – CNN

- Convolutional Networks [Kim, 2014; Kalchbrenner et al., 2015]
  - Captures k-grams hierarchically
  - One 1-D convolutional layer: considers all k-grams
    $h_t = \phi\left(\sum_{\tau=-k/2}^{k/2} W_\tau e_{t+\tau}\right)$, resulting in $H = (h_1, h_2, \ldots, h_T)$.
  - Stack more than one convolutional layers: progressively-growing window
  - Fits our intuition of how sentence is understood: **tokens→multi-word expressions→phrases→sentence**



Introduction (3/4)

task에 따라 extra information이 주어지는 경우 (eg. Neural Machine Translation)
attention을 적용하여, sentence를 vector로 표현함

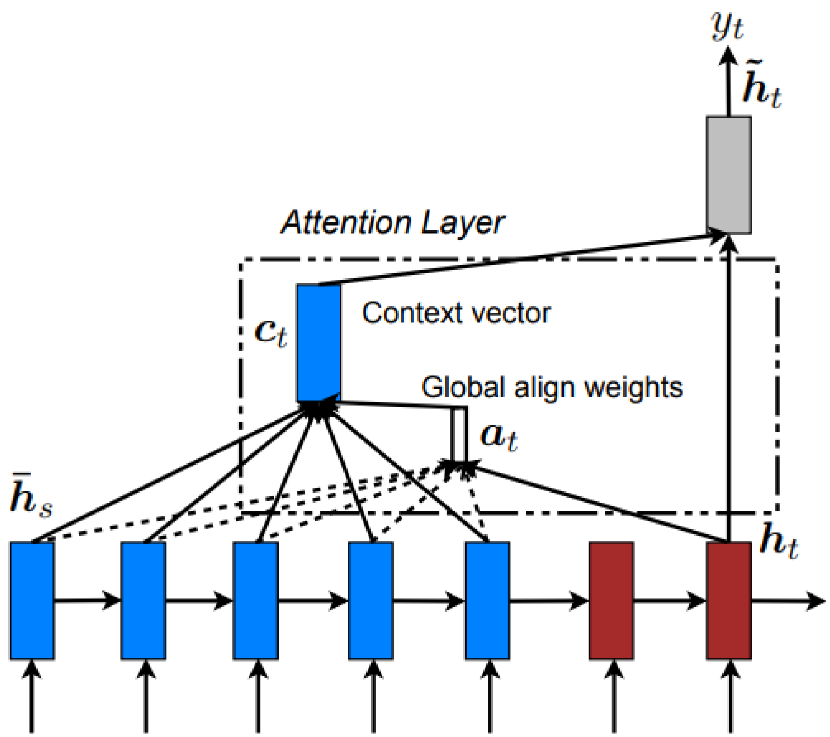


Figure 2: **Global attentional model** – at each time step t, the model infers a *variable-length* alignment weight vector a_t based on the current target state h_t and all source states $\bar{h}_s$. A global context vector c_t is then computed as the weighted average, according to a_t, over all the source states.

Luong, Minh-Thang, Hieu Pham, and Christopher D. Manning. "Effective approaches to attention-based neural machine translation." arXiv preprint arXiv:1508.04025 (2015).

Introduction (4/4)

extra information이 주어지지않은 task에 대해서는 attention이 불가능하지만 주어진 token들 간의 다양한 semantic을 볼 필요성이 존재 → self-attention을 제안

A common approach in many of the aforementioned methods consists of creating a simple vector representation by using the final hidden state of the RNN or the max (or average) pooling from either RNNs hidden states or convolved n-grams. We hypothesize that carrying the semantics along all time steps of a recurrent model is relatively hard and not necessary. We propose a self-attention mechanism for these sequential models to replace the max pooling or averaging step. Different from previous approaches, the proposed self-attention mechanism allows extracting different aspects of the sentence into multiple vector representations. It is performed on top of an LSTM in our sentence embedding model. This enables attention to be used in those cases when there are no extra inputs. In addition, due to its direct access to hidden representations from previous time steps, it relieves some long-term memorization burden from LSTM. As a side effect coming together with our proposed self-attentive sentence embedding, interpreting the extracted embedding becomes very easy and explicit.

Approach – MODEL (1/2)

논문에서 제안하는 model은 bidirectional lstm과 self-attention mechanism 두 module로 구성됨

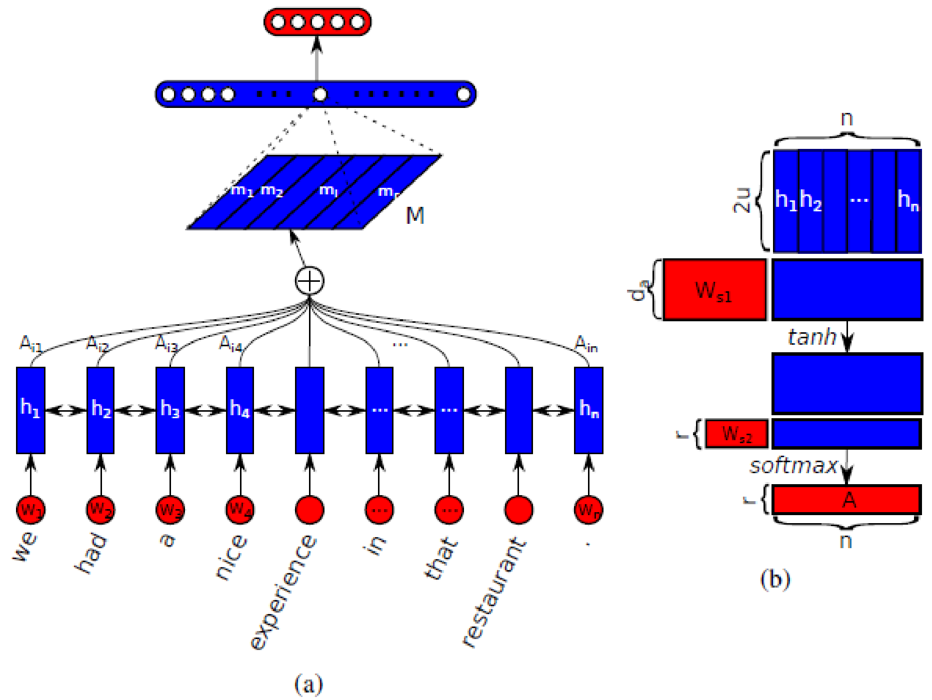
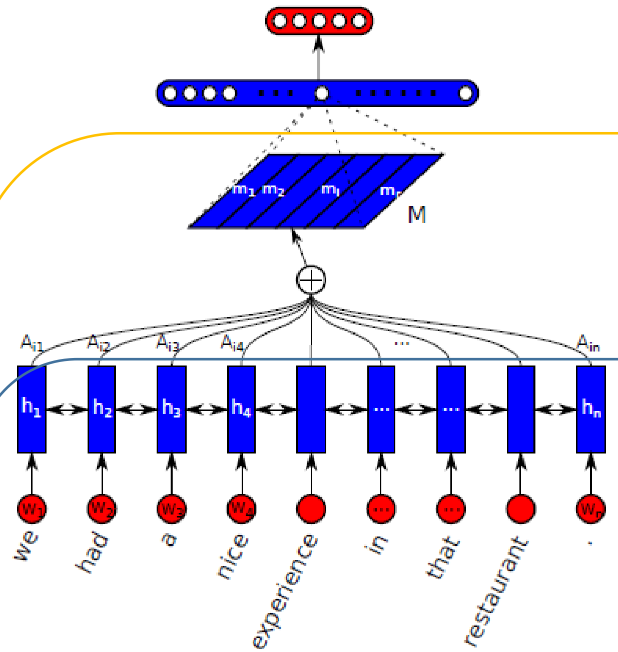
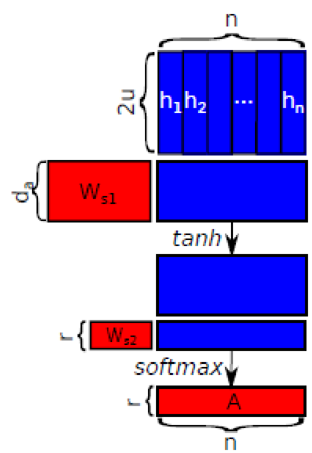


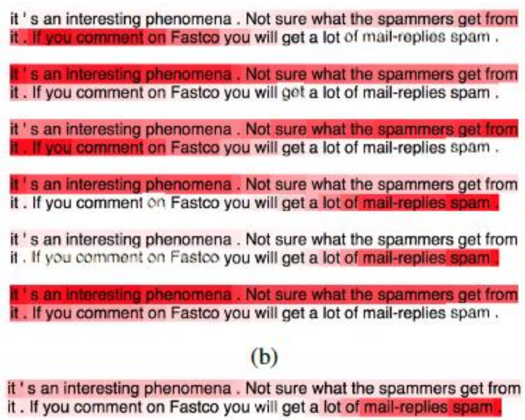
Figure 1: A sample model structure showing the sentence embedding model combined with a fully connected and softmax layer for sentiment analysis (a). The sentence embedding M is computed as multiple weighted sums of hidden states from a bidirectional LSTM ($\mathbf{h_1}$, ..., $\mathbf{h_n}$), where the summation weights (A_{i1}, ..., A_{in}) are computed in a way illustrated in (b). Blue colored shapes stand for hidden representations, and red colored shapes stand for weights, annotations, or input/output.

Approach – MODEL (2/2)

논문에서 제안하는 model은 bidirectional lstm과 self-attention mechanism 두 module로 구성됨







$S = (w_1, w_2, \dots, w_n),\quad w_i \in R^d, \dim(S) = (d, n)$

$\overrightarrow{h_t} = \overrightarrow{\mathrm{LSTM}}(w_t, \overrightarrow{h_{t-1}}),\quad \overrightarrow{h_t} \in R^u$

$\overleftarrow{h_t} = \overleftarrow{\mathrm{LSTM}}(w_t, \overleftarrow{h_{t+1}}),\quad \overleftarrow{h_t} \in R^u$

$h_t = \mathrm{concat}(\overrightarrow{h_t}, \overleftarrow{h_t}),\quad h_t \in R^{2u}$

$H = (h_1, h_2, \dots, h_n),\quad \dim(H) = (2u, n)$

$A = \mathrm{softmax}(W_{s2} \tanh(W_{s1} H))$

$M = AH^T$

$\dim(W_{s1}) = (d_a, 2u)$

$\dim(W_{s2}) = (r, d_a)$

$\dim(A) = (r, n)$

$\dim(M) = (r, 2u)$

Approach – PENALIZATION TERM

*row vector of A matrix*가 개별 aspect를 표현, 개별 aspect를 다양하게 보기위한 penalization term을 제안

$$P = \|(AA^T - I)\|_F^2$$

$$A = \begin{pmatrix} A_1^T \\ \vdots \\ A_r^T \end{pmatrix} = \begin{pmatrix} A_{11} & \cdots & A_{1n} \\ \vdots & \ddots & \vdots \\ A_{r1} & \cdots & A_{rn} \end{pmatrix}, A_i \in R^n \qquad \begin{pmatrix} A_1^T \\ \vdots \\ A_r^T \end{pmatrix} (A_1 \quad \dots \quad A_r) = \begin{pmatrix} A_1^T A_1 & \cdots & A_1^T A_r \\ \vdots & \ddots & \vdots \\ A_r^T A_1 & \cdots & A_r^T A_r \end{pmatrix}$$

$$\mathrm{w.r.t} \sum_{j=1}^{n} A_{ij} = 1, A_{ij} \geq 0$$

penalization term을 task loss(eg. cross entropy loss)에 더해서 최소화
→ *row vector of A matrix*간에 서로 달라지도록 학습이 이루어짐

Experimental results (1/4)

Author profiling, Sentiment Analysis 등 Classification task에 대해서, 좋은 성능을 보임, self-attention으로 classification에 대한 근거를 대략적으로 확인이 가능

Table 1: Performance Comparision of Different Models on Yelp and Age Dataset

Models	Yelp	Age
BiLSTM + Max Pooling + MLP	61.99%	77.40%
CNN + Max Pooling + MLP	62.05%	78.15%
Our Model	**64.21%**	**80.45%**

- if I can give this restaurant a 0 I will we be just ask our waitress leave because someone with a reservation be wait for our table my father and father-in-law be still finish up their coffee and we have not yet finish our dessert I have never be so humiliated do not go to this restaurant their food be mediocre at best if you want excellent Italian in a small intimate restaurant go to dish on the South Side I will not be go back
- this place suck the food be gross and taste like grease I will never go here again ever sure the entrance look cool and the waiter can be very nice but the food simply be gross taste like cheap 99cent food do not go here the food shot out of me quick then it go in
- everything be pre cook and dry its crazy most Filipino people be used to very cheap ingredient and they do not know quality the food be disgusting I have eat at least 20 different Filipino family home this not even mediocre
- seriously f *** this place disgust food and shitty service ambience be great if you like dine in a hot cellar engulf in stagnate air truly it be over rate over price and they just under deliver forget try order a drink here it will take forever get and when it finally do arrive you will be ready pass out from heat exhaustion and lack of oxygen how be that a head change you do not even have pay for it I will not disgust you with the detailed review of everything I have try here but make it simple it all suck and after you get the bill you will be walk out with a sore ass save your money and spare your self the disappointment
- i be so angry about my horrible experience at Medusa today my previous visit be amaze 5/5 however my go to out of town and I land an appointment with Stephanie I go in with a picture of roughly what I want and come out look absolutely nothing like it my hair be a horrible ashy blonde not anywhere close to the platinum blonde I request she will not do any of the pop of colour I want and even after specifically tell her I do not like blunt cut my hair have lot of straight edge she do not listen to a single thing I want and when I tell her I be unhappy with the colour she basically tell me I be wrong and I have do it this way no no I do not if I can go from Little Mermaid red to golden blonde in 1 sitting that leave my hair fine I shall be able go from golden blonde to a shade of platinum blonde in 1 sitting thanks for ruin my New Year's with 1 the bad hair job I have ever have

(a) 1 star reviews

- i really enjoy Ashley and Ami salon she do a great job be friendly and professional I usually get my hair do when I go to MI because of the quality of the highlight and the price the price be very affordable the highlight fantastic thank Ashley i highly recommend you and ill be back
- love this place it really be my favorite restaurant in Charlotte they use charcoal for their grill and you can taste it steak with chimichurri be always perfect Fried yucca cilantro rice pork sandwich and the good tres lech I have had.The desert be all incredible if you do not like it you be a mutant if you will like diabeetus try the Inca Cola
- this place be so much fun I have never go at night because it seem a little too busy for my taste but that just prove how great this restaurant be they have amazing food and the staff definitely remember us every time we be in town I love when a waitress or waiter come over and ask if you want the cab or the Pinot even when there be a rush and the staff be run around like crazy whenever I grab someone they instantly smile acknowlegde us the food be also killer I love when everyone know the special and can tell you they have try them all and what they pair well with this be a first last stop whenever we be in Charlotte and I highly recommend them
- great food and good service .... what else can you ask for everything that I have ever try here have be great
- first off I hardly remember waiter name because its rare you have an unforgettable experience the day I go I be celebrate my birthday and let me say I leave feel extra special our waiter be the best ever Carlos and the staff as well I be with a party of 4 and we order the potato salad shrimp cocktail lobster amongst other thing and boy be the food great the lobster be the good lobster I have ever eat if you eat a dessert I will recommend the cheese cake that be also the good I have ever have it be expensive but so worth every penny I will definitely be back there go again for the second time in a week and it be even good ...... this place be amazing

(b) 5 star reviews

Figure 2: Heatmap of Yelp reviews with the two extreme score.

Experimental results (2/4)

두 가지의 문장이 서로 대조되는 지 아닌 지를 판단하는 Text entailment task에서 도 좋은 성능을 보임

Table 2: Test Set Performance Compared to other Sentence Encoding Based Methods in SNLI Datset

Model	Test Accuracy
300D LSTM encoders (Bowman et al., 2016)	80.6%
600D (300+300) BiLSTM encoders (Liu et al., 2016b)	83.3%
300D Tree-based CNN encoders (Mou et al., 2015a)	82.1%
300D SPINN-PI encoders (Bowman et al., 2016)	83.2%
300D NTI-SLSTM-LSTM encoders (Munkhdalai & Yu, 2016a)	83.4%
1024D GRU encoders with SkipThoughts pre-training (Vendrov et al., 2015)	81.4%
300D NSE encoders (Munkhdalai & Yu, 2016b)	**84.6%**
Our method	84.4%

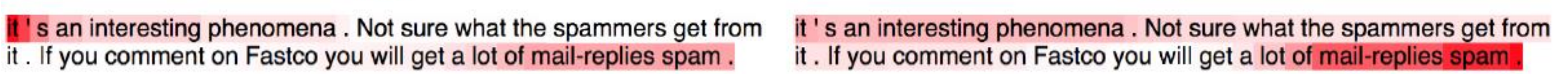
Experimental results (3/4)

서로 다른 aspect를 표현하기위해 제안한 penalization term이 제대로 작동함을 확인함

Table 3: Performance comparision regarding the penalization term

Penalization coefficient	Yelp	Age
1.0	64.21%	80.45%
0.0	61.74%	79.27%

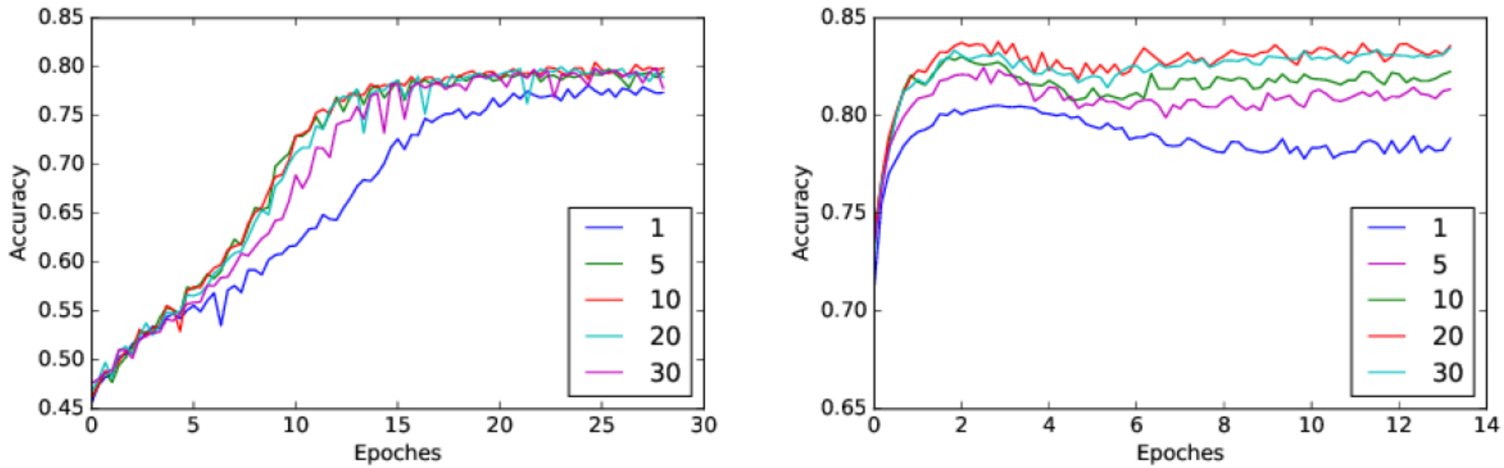
(a) (b)



(c) without penalization (d) with 1.0 penalization

Experimental results (4/4)

sentence를 vector로 representation 하는 것보다 다양한 aspect를 표현할 수 있는 matrix로 representation하는 것이 성능이 좋은 것을 확인



(a) (b)

Figure 5: Effect of the number of rows (r) in matrix sentence embedding. The vertical axes indicates test set accuracy and the horizontal axes indicates training epoches. Numbers in the legends stand for the corresponding values of r. (a) is conducted in Age dataset and (b) is conducted in SNLI dataset.

Conclusion and Discussion

본 논문에서는 sentence를 fixed sized matrix로 embedding을 하는 self-attention을 제안

- Higher level semantic(Long term dependency)을 LSTM에만 의지 X
  - ✓ LSTM doesn't need to carry every piece of information towards its last hidden state
  - ✓ Each LSTM hidden state is only expected to provide shorter tem context information around each word
  - ✓ Higher level semantics

- Future work
  - ✓ 제안한 Self-attention은 다른 supervised task에 기반한 방식으로, unsupervised로 연구가 진행되어야함

Q & A

감사합니다.