arXiv:1911.02711v1 [cs.CL] 7 Nov 2019

Exploring Hierarchical Interaction Between Review and Summary for Better Sentiment Analysis

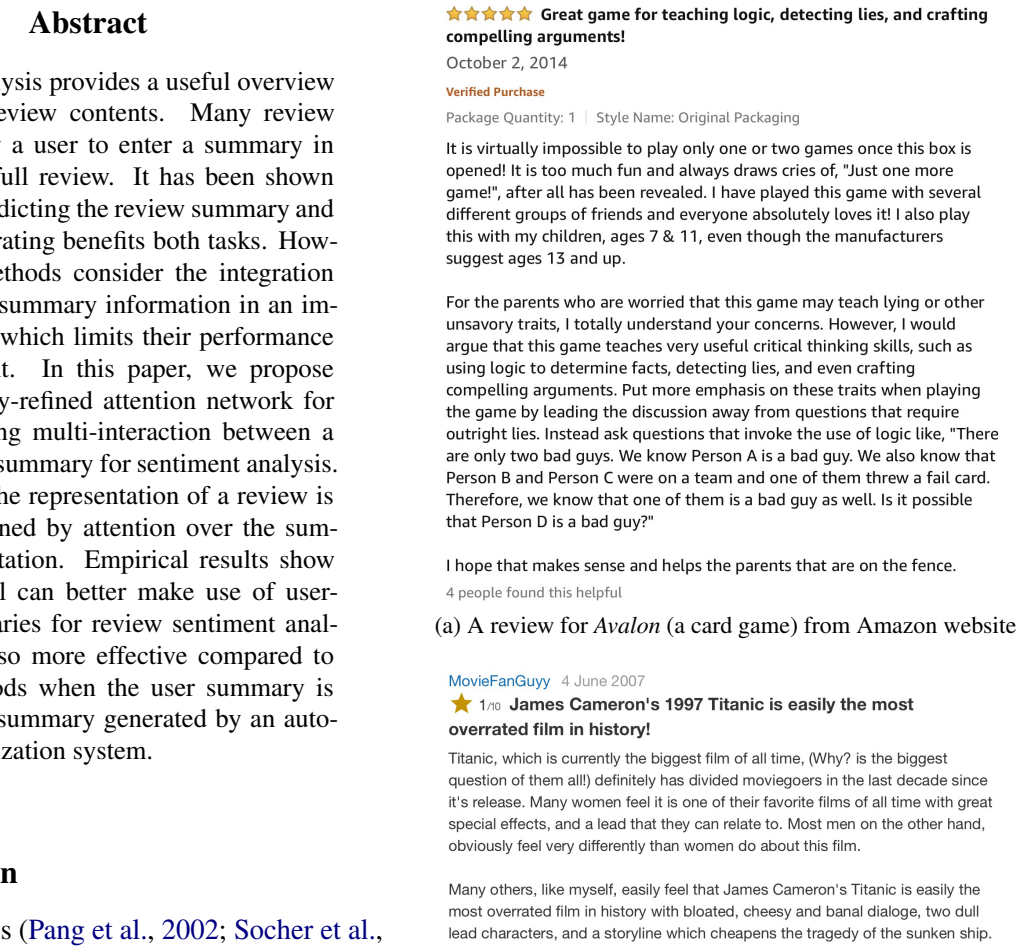
Sen Yang[*1], **Leyang Cui**[*1,2,3] and **Yue Zhang**[1,3]
[1]Westlake University
[2]Zhejiang University
[3]Westlake Institute for Advanced Study
senyang.stu@gmail.com cuileyang@westlake.edu.cn
yue.zhang@wias.org.cn

Abstract

Sentiment analysis provides a useful overview of customer review contents. Many review websites allow a user to enter a summary in addition to a full review. It has been shown that jointly predicting the review summary and the sentiment rating benefits both tasks. However, these methods consider the integration of review and summary information in an implicit manner, which limits their performance to some extent. In this paper, we propose a hierarchically-refined attention network for better exploiting multi-interaction between a review and its summary for sentiment analysis. In particular, the representation of a review is layer-wise refined by attention over the summary representation. Empirical results show that our model can better make use of user-written summaries for review sentiment analysis, and is also more effective compared to existing methods when the user summary is replaced with summary generated by an automatic summarization system.

1 Introduction

Sentiment analysis (Pang et al., 2002; Socher et al., 2013) is a fundamental task in natural language processing. In particular, sentiment analysis of user reviews has wide applications(Cui et al., 2006; Guan et al., 2016; Miyato et al., 2017; Johnson and Zhang, 2017). In many review websites such as Amazon and IMDb, the user is allowed to give a summary in addition to their review. Summaries usually contain more abstract information about the review. As shown in Figure 1, two screenshots of reviews were taken from Amazon and IMDb websites, respectively. The user-written summaries of these reviews can be highly indicative of the final polarity. As a result, it is worth considering them together with the review itself for making sentiment classification.



★★★★★ **Great game for teaching logic, detecting lies, and crafting compelling arguments!**

October 2, 2014

Verified Purchase

Package Quantity: 1 | Style Name: Original Packaging

It is virtually impossible to play only one or two games once this box is opened! It is too much fun and always draws cries of, "Just one more game!", after all has been revealed. I have played this game with several different groups of friends and everyone absolutely loves it! I also play this with my children, ages 7 & 11, even though the manufacturers suggest ages 13 and up.

For the parents who are worried that this game may teach lying or other unsavory traits, I totally understand your concerns. However, I would argue that this game teaches very useful critical thinking skills, such as using logic to determine facts, detecting lies, and even crafting compelling arguments. Put more emphasis on these traits when playing the game by leading the discussion away from questions that require outright lies. Instead ask questions that invoke the use of logic like, "There are only two bad guys. We know Person A is a bad guy. We also know that Person B and Person C were on a team and one of them threw a fail card. Therefore, we know that one of them is a bad guy as well. Is it possible that Person D is a bad guy?"

I hope that makes sense and helps the parents that are on the fence.

4 people found this helpful

(a) A review for *Avalon* (a card game) from Amazon website

MovieFanGuyy 4 June 2007

★ 1/10 **James Cameron's 1997 Titanic is easily the most overrated film in history!**

Titanic, which is currently the biggest film of all time, (Why? is the biggest question of them all!) definitely has divided moviegoers in the last decade since it's release. Many women feel it is one of their favorite films of all time with great special effects, and a lead that they can relate to. Most men on the other hand, obviously feel very differently than women do about this film.

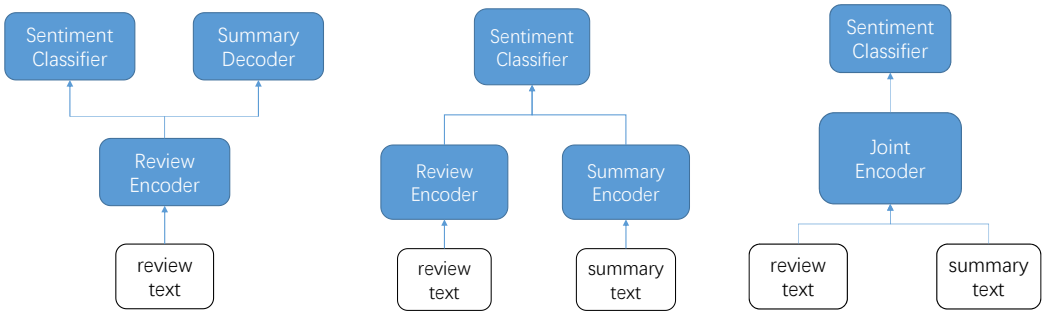
Many others, like myself, easily feel that James Cameron's Titanic is easily the most overrated film in history with bloated, cheesy and banal dialoge, two dull lead characters, and a storyline which cheapens the tragedy of the sunken ship.

(b) A review for the movie, *Titanic*, from IMDb website (Noted that we only include the first two paragraphs because the entire review is too long.)

Figure 1: Screenshots from two review websites. Each of them contains a brief summary along with the review text.

To this end, some recent work (Ma et al., 2018; Wang and Ren, 2018) exploits joint modeling. The model structure can be illustrated by Figure 2a. In

* Equal contribution.
** This work is still in progress.



(a) Joint modeling (b) Separate encoder (c) Joint encoder

Figure 2: Three model structures for incorporating summary into sentiment classification

particular, given a review input, a model is trained to simultaneously predict the sentiment and summary. As a result, both summary information and review information are integrated in the review encoder through back-propagation training. However, one limitation of this method is that it does not explicitly encode a summary during test time.

One solution, as shown in Figure 2b, is to train a separate summary generator, which learns to predict a summary given a review. This allows a sentiment classifier to simultaneously encode the review and its summary, before making a prediction using both representations. One further advantage of this model is that it can make use of a user-given summary if it is available with the review, which is the case for the review websites shown in Figure 1. We therefore investigate such a model. One limitation of this method, however, is that it does not capture interaction of review and summary information as thoroughly as the method shown in Figure 2a, since the review and the summary are encoded using two separate encoders.

To address this issue, we further investigate a joint encoder for review and summary, which is demonstrated in Figure 2c. The model works by jointly encoding the review and the summary in a multi-layer structure, incrementally updating the representation of the review by consulting the summary representation at each layer. As shown in Figure 3, our model consists of a summary encoder, a hierarchically-refined review encoder and an output layer. The review encoder is composed of multiple attention layers, each consisting of a sequence encoding layer and an attention inference layer. Summary information is integrated into the representation of the review content at each attention layer, thus, a more abstract review representation is learned in subsequent layers based on a lower-layer representation. This mechanism allows the summary to better guide the representation of the review in a bottom-up manner for improved sentiment classification.

We evaluate our proposed model on the SNAP (Stanford Network Analysis Project) Amazon review datasets (He and McAuley, 2016), which contain not only reviews and ratings, but also golden summaries. In scenarios where there is no user-written summary for a review, we use pointer-generator network (See et al., 2017) to generate abstractive summaries. Empirical results show that our model significantly outperforms all strong baselines, including joint modeling, separate encoder and joint encoder methods. In addition, our model achieves new state-of-the-art performance, attaining 2.1% (with generated summary) and 4.8% (with golden summary) absolutely improvements compared to the previous best method on SNAP Amazon review benchmark.

2 Related Work

The majority of recent sentiment analysis models are based on either convolutional or recurrent neural networks to encode sequences (Kim, 2014; Tang et al., 2015).

In particular, attention-based models have been widely explored, which assign attention weights to hidden states to generate a representation of the input sequence. A hierarchical model with two levels of attention mechanisms was proposed for document classification (Yang et al., 2016). Self-attention mechanism has also been used in sentiment analysis (Lin et al., 2017; Letarte et al., 2018). However, Jain and Wallace (2019) empirically showed that self-attention mechanism does not consistently agree with the most salient features, which means that self-attention models may

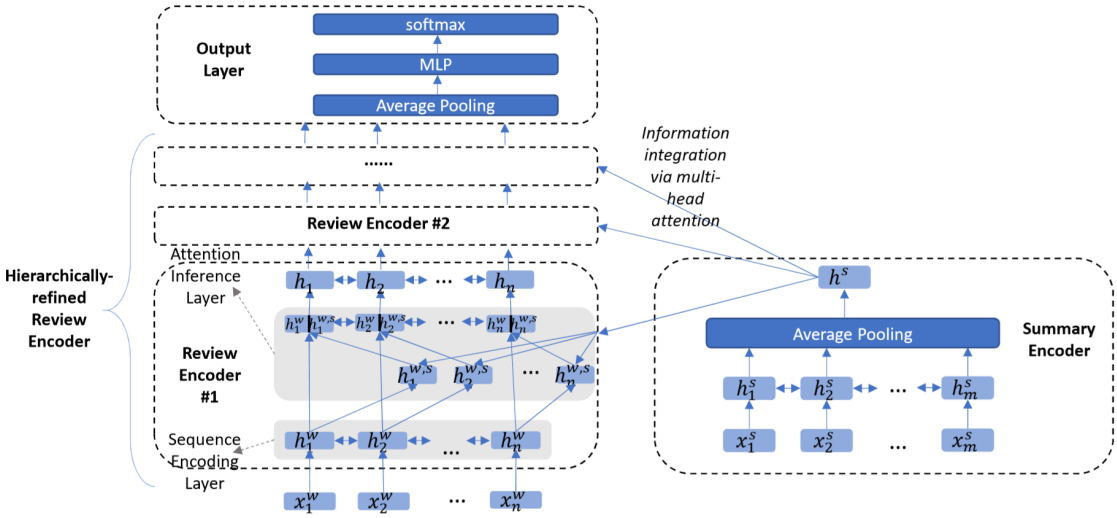


Figure 3: Architecture of proposed model ($X^w = x_1^w, x_2^w, ..., x_n^w$: review; $X^s = x_1^s, x_2^s, ..., x_m^s$: summary).

suffer from attending on explicit but irrelevant sentimental words.

Rationales were also introduced to sentiment analysis task. Bastings et al. (2019) proposed a unsupervised latent model that selects a rationale and then uses the rationale for sentiment analysis. A rationale-augmented CNN model (Zhang et al., 2016) was proposed, which regards golden rationales as additional input and uses the probability as rationale-level attention weights to generate the final representation for text classification.

There has also been work focusing on joint summarization and sentiment classification (Ma et al., 2018; Wang and Ren, 2018), whose general structures are illustrated in Figure 2a. These models can predict sentiment label and summary simultaneously. However, they do not encode summaries explicitly during test time, which makes their performance be limited to some extent.

3 Method

In this section, we introduce our proposed model in details. We first give the problem formulation, followed by an overview of the proposed model, and explain each layer of our model in details, before finally giving the loss function and training methods.

3.1 Problem Formulation

The input to our task is a pair (X^w, X^s), where $X^w = x_1^w, x_2^w, ..., x_n^w$ is a summary and $X^s = x_1^s, x_2^s, ..., x_m^s$ is a review, the task is to predict the sentiment label $y \in [1, 5]$, where 1 denotes the most negative sentiment and 5 denotes the most positive sentiment. n and m denote the size of the review and summary in the number of words, respectively. The training set is $D = \{(X_i^w, X_i^s, y_i)\}|_{i=1}^M$ where M is the total number of training examples.

3.2 Model Overview

Figure 3 gives the architecture of the proposed model, which consists of three modules: a summary encoder, a hierarchically-refined review encoder and an output layer. The summary encoder encodes the summary into a hidden state matrix. The review encoder consists of several layers for representing $\mathbf{x}^w$, each containing a sequence encoding sublayer and an attention inference sublayer. The sequence encoding sublayer encodes the review text as a word sequence. The attention inference layer acts as a key component, which takes the hidden states from both the original review and the summary as input calculating dot-product attention weights for original review under additional supervision from summary information. Multi-head attention (Vaswani et al., 2017) as well as residual connection are also adopted. The output layer predicts the potential sentiment label according to hidden states from the previous layer.

3.3 Summary Encoder

Input for the summary encoder is a sequence of summary word representations $\mathbf{x}^s = \mathbf{x}_1^s, \mathbf{x}_2^s, ..., \mathbf{x}_m^s = \{emb(x_1^s), ..., emb(x_m^s)\}$, where emb denotes a word embedding lookup table.

Word representations are fed into a standard BiLSTM. We adopt a standard LSTM formulation, where a sequence of hidden states $\mathbf{h}_t$ are calculated from a sequence of $\mathbf{x}_t (t \in [1, ..., m])$.

A forward left-to-right LSTM layer and a backward right-to-left LSTM yield a sequence of forward hidden states $\{\overrightarrow{\mathbf{h}}_1^s, ..., \overrightarrow{\mathbf{h}}_n^s\}$ and a sequence of backward hidden states $\{\overleftarrow{\mathbf{h}}_1^s, ..., \overleftarrow{\mathbf{h}}_n^s\}$, respectively. The two hidden states are concatenated to form a final representation:

$$\begin{aligned} \mathbf{h}_i^s &= [\overrightarrow{\mathbf{h}}_i^s; \overleftarrow{\mathbf{h}}_i^s] \\ \mathbf{H}^s &= \{\mathbf{h}_1^s, ..., \mathbf{h}_m^s\} \end{aligned} \tag{1}$$

We then apply an average-pooling operation over the hidden and take $\mathbf{h}^s = avg\_pooling(\mathbf{h}_1^s, \mathbf{h}_2^s, ..., \mathbf{h}_n^s)$ as the final representation of summary text.

3.4 Hierarchically-Refined Review Encoder

The hierarchically-refined review encoder consists of several review encoder layers, each of which is composed of a sequence encoding layer and an attention inference layer.

3.4.1 Sequence Encoding Layer

Given a review $\mathbf{x}^w = \{emb(x_1^w), ..., emb(x_n^w)\}$, another BiLSTM is adopted (the same equation with different parameters compared to the one used in the summary encoder), deriving a sequence of review hidden states $\mathbf{H}^w = \{\mathbf{h}_1^w, \mathbf{h}_2^w, ..., \mathbf{h}_n^s\}$.

3.4.2 Attention Inference Layer

In the attention inference layer, we model the dependencies between the original review and the summary with multi-head dot-product attention.Each head produces an attention matrix $\boldsymbol{\alpha} \in \mathbb{R}^{d_h \times 1}$ consisting of a set of similarity scores between the hidden state of each token of the review text and the summary representation. The hidden state outputs are calculated by

$$\begin{aligned} \boldsymbol{\alpha} &= \mathrm{softmax}(\mathbf{H}^w \mathbf{W}_i^Q (\mathbf{h}^s \mathbf{W}_i^K)^T) \\ \mathrm{head}_i &= \boldsymbol{\alpha} \mathbf{h}^s \mathbf{W}_i^V \\ \mathbf{H}^{w,s} &= \mathrm{concat}(\mathrm{head}_1, \ldots, \mathrm{head}_k) \end{aligned} \tag{2}$$

where $\mathbf{W}_i^Q \in \mathbb{R}^{d_h \times \frac{d_h}{k}}$, $\mathbf{W}_i^K \in \mathbb{R}^{d_h \times \frac{d_h}{k}}$ and $\mathbf{W}_i^V \in \mathbb{R}^{d_h \times \frac{d_h}{k}}$ are model parameters. Q, K and V represent *Query*, *Key* and *Value*, respectively. k is the number of parallel heads and $i \in [1, k]$ indicates which head is being processed.

Following (Vaswani et al., 2017), we adopt a residual connection around each attention inference layer, followed by layer normalization (Ba et al., 2016) :

$$\mathbf{H} = \mathrm{LayerNorm}(\mathbf{H}^w + \mathbf{H}^{w,s}) \tag{3}$$

$\mathbf{H}$ is then fed to the subsequent sequence encoding layer as input, if any.

According to the equations of standard LSTM and Equation 2, tokens of the original review that are the most relevant to the summary are focused on more by consulting summary representation. The hidden states $\mathbf{H}^{w,s}$ are thus a representation matrix of the review text that encompass key features of summary representation. Multi-head attention mechanism ensures that multi-faced semantic dependency features can be captured during the process, which is beneficial for scenarios where several key points exist in one review.

Note also that our design of the review encoding part of the hierarchically-refined attention network is similar to the Transformer architecture in the use of multi-head attention, residual connection and layer normalization (Vaswani et al., 2017). However, our experiments show that bi-directional LSTM works better compared to self-attention network as a basic layer structure. This may result from the fact that Transformer requires a larger amount of training data for the most effectiveness.

3.5 Output Layer

Finally, global average pooling is applied after the previous layer, and then followed by a classifier layer:

$$\begin{aligned} \mathbf{h}^{avg} &= avg\_pooling(\mathbf{h}_1, ..., \mathbf{h}_n) \\ \mathbf{p} &= \mathrm{softmax}(\mathbf{W}\mathbf{h}^w + \mathbf{b}) \\ \hat{y} &= \mathrm{argmax}\, \mathbf{p} \end{aligned} \tag{4}$$

where $\hat{y}$ is the predicted sentiment label; $\mathbf{W}$ and $\mathbf{b}$ are parameters to be learned.

3.6 Training

Given a dataset $D = \{(X_t^w, X_t^s, y_t)\}|_{t=1}^{|T|}$, our model can be trained by minimizing the cross-entropy loss between

$$L = -\sum_{t=1}^{|T|} log(\mathbf{p}^{y_t}) \tag{5}$$

where $\mathbf{p}^{y_t}$ denotes the value of the label in $\mathbf{p}$ that corresponds to y_t.

Domain	Size	#Review	#Summary
Toys & Games	168k	99.9	4.4
Sports & Outdoors	296k	87.2	4.2
Movies & TV	1698k	161.6	4.8

Table 1: Data statistics. Size: number of samples, #Review: the average length of review, #Summary: the average length of summary.

Domain	#Recall	#ROUGE		
Toys & Games	0.34	18.44	5.00	17.69
Sports & Outdoors	0.33	17.85	4.77	17.59
Movies & TV	0.33	14.52	4.84	13.42

Table 2: Statistics of generated summary. #Recall refers to the percentage of words in a summary that occur in the corresponding review. #ROUGE (Lin, 2004) indicates the abstractive summarization experimental result reported in HSSC (Ma et al., 2018), including ROUGE-1, ROUGE-2, ROUGE-L, respectively.

4 Experiments

We compare our model with several strong baselines and previous state-of-the-art methods, investigating its main effects.

4.1 Datasets

We empirically compare different methods using Amazon SNAP Review Dataset (McAuley and Leskovec, 2013), which is a part of Stanford Network Analysis Project. The raw dataset consists of around 34 millions Amazon reviews in different domains, such as books, games, sports and movies. Each review mainly contains a product ID, a piece of user information, a plain text review, a review summary and an overall sentiment rating which ranges from 1 to 5. The statistics of our adopted dataset is shown in Table 1. For fair comparison with previous work, we adopt the same partitions used by previous work (Ma et al., 2018; Wang and Ren, 2018), which is, for each domain, the first 1000 samples are taken as the development set, the following 1000 samples as the test set, and the rest as the training set.

4.2 Experimental Settings

We use GloVe (Pennington et al., 2014) 300-dimensional embeddings as pretrained word vectors. A LSTM hidden size of 256 and four heads for multi-head attention mechanism are adopted. We use Adam (Kingma and Ba, 2015) to optimize our model, with an initial learning rate of 0.0003, a decay rate of 0.97, momentum parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1 \times 10^{-8}$. The dropout rate is set depending on the size of each dataset, which is 0.5 for both Toys & Games and Sports & Outdoors and 0.2 for Movies & TV.

We conduct experiments with both golden summaries and generated summaries. For generating automatic-decoded summaries, we train a pointer-generator network (PG-Net) with coverage mechanism (See et al., 2017), which is a specially designed sequence-to-sequence attention-based model that can generate the summary by copying words from the text document or generating words from a fixed vocabulary set at the same time. We generally follow the experimental settings in the original paper except for some minor adjustments specially made for our datasets. Noted that in our work PG-Net can be replaced by any other summarization model.

4.3 Baselines

HSSC (Ma et al., 2018). This model adopts encoder parameter sharing for jointly sentiment classification and summarization. It predicts the sentiment label using a highway layer, concatenating the hidden state in summary decoder and the original text representation in encoder.

SAHSSC (Wang and Ren, 2018). This work also adopts encoder parameter sharing for jointly sentiment classification and summarization. They use two separate BiLSTMs with self-attention mechanism for generating review and summary representations.

BiLSTM+Pooling. For this baseline, we use a BiLSTM with hidden sizes of 256 in both directions, and average pooling across all hidden states to form the representation. This method serves as a naive baseline for making use of both review and summary in sentiment classification. It can also be used to compare the effectiveness of the review itself, the summary itself and the combination of both when used as inputs to the problem.

BiLSTM+Self-attention (Lin et al., 2017). This baseline uses a BiLSTM with hidden size of 256 in both directions. On the top of BiLSTM, self-attention is used to provide a set of summation weight vectors for the final representation. This method is conceptually simple yet gives the state-of-the-art results for many classification and text matching tasks. Its main difference to our model lies in the fact that attention is performed only in

Model	#Hidden	#Layer	Acc	#Param
BiLSTM +self-attention	128	2	76.3	1M
	256	1	76.7	1.3M
	256	2	76.6	2.6M
	256	3	76.4	3.9M
	360	2	76.7	4.3M
Our model	128	2	77.1	2.7M
	256	1	77.3	4.2M
	256	2	77.6	5.3M
	256	3	77.3	6.4M
	360	2	77.7	9.3M

Table 3: Results (with golden summary) on the development set of Toys&Games. #Hidden: LSTM hidden size, # Layer: number of layers, Acc: accuracy, # Param: number of parameters

the top hidden layer in this method, yet in every layer in ours.

BiLSTM+Hard Attention To demonstrate the efficiency of our model structure, we also adopt hard attention (Xu et al., 2015) for comparison, which is supervised using an extractive summarization objective. In particular, words in the original review that match to the corresponding summary are treated as the summary in their original order. In the case of Figure 1b, the extractive summaries for the review are "James Cameron's Titanic is easily the most overrated film in history", which corresponds to the user-written summary "James Cameron's 1997 Titanic is easily the most overrated film in history!". The model also calculates another loss between attention weights and extractive summary labels, so that the hard attention weights are trained to strictly follow the extractive summary.

For baselines that adopt the separate encoder structure, we generally calculate the representations of review and summary separately with two encoders that hold their own parameters, and then concatenate the two representations alongside the hidden-size dimension. For the joint encoder baselines, we first concatenate the review and summary text, and then encode the concatenated text with one single encoder.

4.4 Development Experiments

We use the Toys & Games development set to investigate different key configurations of our model. The results are shown in Table 3.

Self-attention Baseline We compare different numbers of BiLSTM layers and hidden sizes in BiLSTM self-attention. As can be seen, with more layers a stacked BiLSTM with larger hidden sizes does not give better results compared to a hidden size of 256 either.

Hidden Size We see an evident improvement of our model when the hidden size increases from 128 to 256. However, the improvement becomes relatively small compared to a large increase in the number of parameters when the hidden size is further increased to 360. Therefore, we adopt 256 as the hidden size in our experiments.

Number of Layers As Table 3 shows, the accuracy increases when increasing layer numbers from 1 to 2. More layers do not increase the accuracy on development set. We thus set 2 as the number of review encoder layers in the experiments. The best performing model size is comparable to that of the BiLSTM self-attention, demonstrating that the number of parameters is not the key factor to models' performance.

4.5 Results

Table 4 and Table 5 show the final results. Our model outperforms all the baseline models and the top-performing models with both generated summary and golden summary, for all the three datasets. In the scenario where golden summaries are used, BiLSTM+self-attention performs the best among all the baselines, which shows that attention is a useful way to integrate summary and review information. Hard-attention receives more supervision information compared with soft-attention, by supervision signals from extractive summaries. However, it underperforms the soft attention model, which indicates that the most salient words for making sentiment classification may not strictly overlap with *extractive* summaries. This justifies the importance of user written or automatic-generated summary.

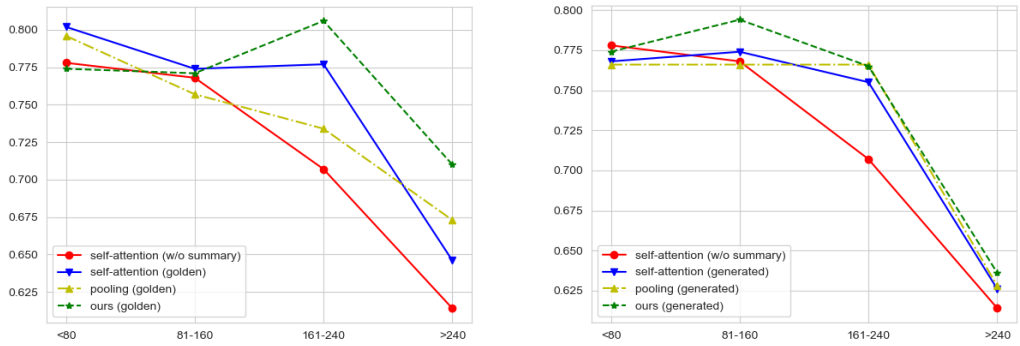
A comparison between models that use summary information and those that do not use summary information shows that the review summary is useful for sentiment classification. In addition, the same models work consistently better when the user written gold summary is used compared to a system generated summary, which is intuitively reasonable since the current state-of-the-art abstractive summarization models are far from perfect. Interestingly, as shown in the second section of the table, the gold summary itself does not lead to better sentiment accuracy compared with the review

Structure	Model	Toys & Games	Sports & Outdoors	Movies & TV	Average
Joint Modeling	HSSC (Ma et al., 2018)	71.9	73.2	68.9	71.3
	SAHSSC (Wang and Ren, 2018)	72.5	–	69.2	70.9
Separate Encoder	BiLSTM+pooling (*Predicted*)	–	–	–	–
	BiLSTM+self-attention (*Predicted*)	68.3	–	–	–
Joint Encoder	BiLSTM+hard attention*	73.4	72.1	73.9	73.1
	BiLSTM+pooling (*Predicted*)	73.8	72.0	72.0	72.6
	BiLSTM+self-attention (*Predicted*)	73.9	71.6	72.4	72.6
	Our model (*Predicted*)	**74.8**	**72.6**	**72.8**	**73.4**

Table 4: Experimental results. *Predicted* indicates the use of system-predicted summaries. Star (*) indicates that hard attention model is trained with golden summaries but does not require golden summaries during inference.

Structure	Model	Toys & Games	Sports & Outdoors	Movies & TV	Average
Separate Encoder	BiLSTM+pooling (*Golden*)	71.2	–	–	–
	BiLSTM+self-attention (*Golden*)	73.0	–	–	–
Joint Encoder	BiLSTM+pooling (*Golden*)	75.4	73.4	73.2	74.0
	BiLSTM+self-attention (*Golden*)	75.8	74.3	75.3	75.1
	Our model (*Golden*)	**77.0**	**75.7**	**75.6**	**76.1**

Table 5: Experimental results. *Golden* indicates the use of user-written (golden) summaries. Noted that joint modeling methods, such as HSSC (Ma et al., 2018) and SAHSSC (Wang and Ren, 2018), cannot make use of golden summaries during inference time, so their results are excluded in this table.



(a) Accuracy with golden summary

(b) Accuracy with generated summary

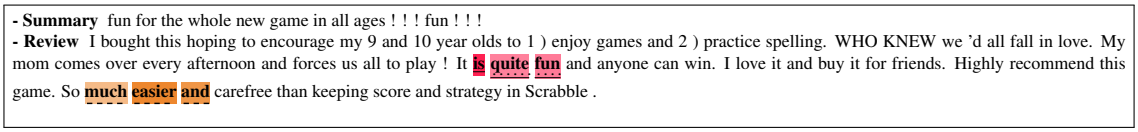
Figure 4: Accuracy against the review length

itself, which shows that summaries better serve as auxiliary information sources to review contents.

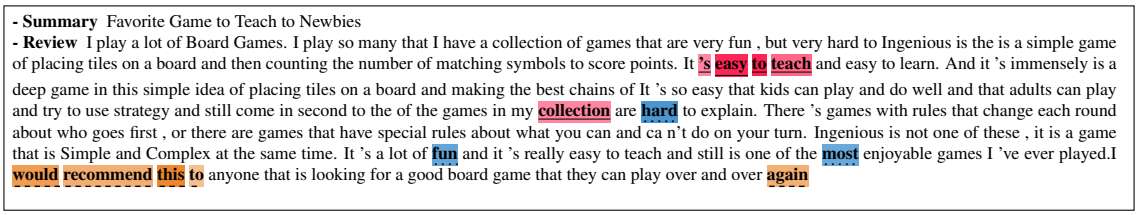
With both gold summaries and automatic-generated summaries, our model gives better results as compared to BiLSTM+self-attention. The latter integrates information from reviews and summaries only in the top representation layer, which is also the standard practice in question answering (Chen et al., 2016) and machine translation (Bahdanau et al., 2015) models. In contrast, our model integrates summary information into the review representation in each layer, thereby allowing the integrated representation to be hierarchically refined, leading to more abstract hidden states.

Finally, the fact that with gold summary, our baseline and final models outperforms the state-of-the-art methods by jointly training shows the importance of making use of user written summaries when they are available. Even with system summary, out models still outperforms HSSC and SAHSSC, showing that our network is more effective than parameter sharing under the same setting without input summaries.

Review Length Figure 4 consists of line graphs on the accuracy of BiLSTM+self-attention, BiLSTM+pooling and our model against the review length. As the review length increases, the performance of all models decreases. BiLSTM+self-attention does not outperform BiLSTM+pooling on long text. Our method gives better results compared to two baseline models for long reviews,



(a) Attention heatmap with generated summary



(b) Attention heatmap with golden summary

Figure 5: Visualizations of self-attention and hierarchically-refined attention, one with generated summary and the other with golden summary. (1) BiLSTM+self-attention: dot line / blue color; (2) First layer of our model: straight line / pink color; (3) Second layer of our model: dash line / yellow color. Deeper colors indicates higher attention weights. Noted that there exist attention visualization overlaps among different layers.

demonstrating that our model is effective for capturing long-term dependency. This is likely because hierarchically-refined attention maintains the most salient information while ignoring the redundant parts of the original review text. Our model can thus be more robust when review has irrelevant sentimental words, which usually exists in larger reviews such as the example in Figure 1a. The hierarchical architecture allows the lower layers to encode local information, while the higher layers can capture long-term dependency and thus better encode global information.

Case Study Our model has a natural advantage of interpretability thanks to the use of attention inference layer. We visualize the hierarchically-refined attention of two samples from the test set of Toys & Games. We also visualize self-attention distribution for fair comparison. To make the visualizations clear and to avoid confusion, we choose to visualize the most salient parts, by rescaling all attention weights into an interval of $[0, 100]$ and adopting 50 as a threshold for attention visualization, showing only attention weights ≥ 50.

As shown in Figure 5a, the example with generated summary has 5 stars as its golden rating score. The summary text is "fun for the whole new game in all ages ! ! ! fun ! ! !", which suggests that the game is (1) fun (from word "fun") and (2) not difficult to learn (from phrase "all ages"). It can be seen that both the self-attention model and the first layer of our model attend to the strongly positive phrase "quite fun", which is relevant to the word "fun" in the summary. In comparisons the second layer attends to the phrase "much easier", which is relevant to the phrase "in all ages" in the summary. This verifies our model's effectiveness of leveraging abstractive summary information.

Figure 5b illustrates a 5-star-rating example with golden summary. The summary text is "Favorite Game to Teach to Newbies". As shown in the heatmap, self-attention can only attend to some general sentimental words, such as "hard", "fun", "immensely" and "most", which deviates from the main idea of the document text. In comparison, the first layer of our model attends to phrases like "easy to teach", which is a perfect match of the phrase "teach to newbies" in the summary. This shows that the shallow sequence inference layer can learn direct similarity matching information under the supervision of summarization. In addition, the second layer of our model attends to phrases including "would recommend this to anyone", which links to "easy to teach" and "Teach to Newbies", showing that the deeper sequence inference layer of our model can learn potential connections between the review and the summary.

5 Conclusion

We investigated a hierarchically-refined attention network for better sentiment prediction. Our model allows multi-interaction between summary and review representation in a hierarchical manner. Empirical results show that the proposed method outperforms all strong baselines and previous work and achieves new state-of-the-art performance on

SNAP Amazon Review dataset.

References

Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *CoRR*, abs/1607.06450.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Joost Bastings, Wilker Aziz, and Ivan Titov. 2019. Interpretable neural predictions with differentiable binary variables. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2963–2977.

Danqi Chen, Jason Bolton, and Christopher D. Manning. 2016. A thorough examination of the CNN/Daily Mail reading comprehension task. In *Association for Computational Linguistics (ACL)*.

Hang Cui, Vibhu O. Mittal, and Mayur Datar. 2006. Comparative experiments on sentiment classification for online product reviews. In *Proceedings, The Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference, July 16-20, 2006, Boston, Massachusetts, USA*, pages 1265–1270.

Ziyu Guan, Long Chen, Wei Zhao, Yi Zheng, Shulong Tan, and Deng Cai. 2016. Weakly-supervised deep learning for customer review sentiment classification. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, pages 3719–3725.

Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 507–517, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 3543–3556.

Rie Johnson and Tong Zhang. 2017. Deep pyramid convolutional neural networks for text categorization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 562–570.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1746–1751.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Gaël Letarte, Frédérik Paradis, Philippe Giguère, and François Laviolette. 2018. Importance of self-attention for sentiment analysis. In *Proceedings of the Workshop: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP 2018, Brussels, Belgium, November 1, 2018*, pages 267–275.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Zhouhan Lin, Minwei Feng, Cícero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. *CoRR*, abs/1703.03130.

Shuming Ma, Xu Sun, Junyang Lin, and Xuancheng Ren. 2018. A hierarchical end-to-end model for jointly improving text summarization and sentiment classification. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden.*, pages 4251–4257.

Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: Understanding rating dimensions with review text. In *Proceedings of the 7th ACM Conference on Recommender Systems*, RecSys '13, pages 165–172, New York, NY, USA. ACM.

Takeru Miyato, Andrew M. Dai, and Ian J. Goodfellow. 2017. Adversarial training methods for semi-supervised text classification. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, EMNLP 2002, Philadelphia, PA, USA, July 6-7, 2002*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1073–1083.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1631–1642.

Duyu Tang, Bing Qin, and Ting Liu. 2015. Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1422–1432.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008.

Hongli Wang and Jiangtao Ren. 2018. A self-attentive hierarchical model for jointly improving text summarization and sentiment classification. In *Proceedings of The 10th Asian Conference on Machine Learning, ACML 2018, Beijing, China, November 14-16, 2018.*, pages 630–645.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 2048–2057.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alexander J. Smola, and Eduard H. Hovy. 2016. Hierarchical attention networks for document classification. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 1480–1489.

Ye Zhang, Iain James Marshall, and Byron C. Wallace. 2016. Rationale-augmented convolutional neural networks for text classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 795–804.