ICIC Express Letters
Volume **16**, Number **5**, May **2022**

 ISSN 1881-803X
pp. **497–504**

GCNN WITH SELF-ATTENTION IS BETTER THAN GRU WITH SELF-ATTENTION FOR SENTIMENT ANALYSIS

Hidekazu Yanagimoto[1,*] and Kiyota Hashimoto[2]

[1]College of Sustainable System Sciences
Osaka Metropolitan University
1-1 Gakuen-cho, Naka-ku, Sakai, Osaka 599-8531, Japan
[*]Corresponding author: hidekazu@kis.osakafu-u.ac.jp

[2]Faculty of Technology and Environment
Prince of Songkla University
80 Moo 1, Vichitsongkram Road, Kathu, Phuket 83120, Thailand
kiyota.h@phuket.psu.ac.th

Received September 2021; accepted November 2021

Abstract. *A relatively simple task of natural language processing like sentiment analysis wants efficient but sufficiently effective methods and GCNN with Self-Attention and GRU with Self-Attention are such candidates. This paper compared these two models with several datasets and with different conditions. The results show that GCNN with Self-Attention is always better by approximately 1% or more of accuracy than GRU with Self-Attention, let alone GCNN without Self-Attention, and it may well indicate that not only prior but also posterior information in text should be employed and that such employment of both information should not be too far. GCNN is known for faster computing time than GRU, but our results urge more investigations on how textual context should be considered.*

Keywords: Natural language processing, Deep learning, Sentiment analysis, Self-Attention mechanism

1. **Introduction.** Sentiment Analysis (SA) is one of the long intrigued topics of natural language processing to estimate the positive or negative rate of texts [1,2]. Its methods are roughly divided into two directions: summing up the sentiment value of keywords included and regarding it as a machine learning task of classification. The sentiment value, or even the polarity, of a keyword can be different in contexts, and texts are bound to syntactic and discourse dependencies. Thus, effective considerations of linguistic contexts are a key to improving the accuracy of SA, where deep learning approaches have been expected to contribute. Different from images and other fixed-sized data, the lengths of text in each data entry is varying and thus recurrent neural networks have been mainly employed, including Long Short-Term Memory (LSTM) models and Gated Recurrent Networks (GRU) models [3,4]. Gated Convolutional Neural Networks (GCNN) were also proposed later, which can employ parallel GPU computing with a comparable accuracy [5].

To capture key factors in contexts more effectively, attention mechanisms [6], particularly Self-Attention mechanism [7], were proposed originally for machine translation [8], and it has been improved rapidly and used for many tasks [9]. In terms of natural language processing, BERT, Transformer, and other methods have been cultivated. In the case of SA, however, the size of the target texts is often limited, and the simple nature of its tasks of binary classification is still in need of efficient as well as effective learning

DOI: 10.24507/icicel.16.05.497

models. GCNN with Self-Attention and GRU with Self-Attention are two of such candidates. Our previous study [9] showed their sufficient performance, but a more deliberate comparison is waited to reveal their characteristics.

The aim of this paper is to clarify the performance differences among various settings of GCNN with Self-Attention in comparison with GRU with Self-Attention in terms of SA. GCNN with Self-Attention was also compared with different kernel sizes of 2, 3, and 4. The results showed that Self-Attention is indeed effective for all models of GCNN but not necessarily for GRU, that GCNN with Self-Attention, particularly with the kernel size of 4, achieved better accuracies than GRU with Self-Attention for all datasets, and that their computing time is four to six times faster. The result of better accuracy also strongly suggests that Self-Attention with both prior and posterior information in texts should be employed and that in due course, too long prior and posterior information may decrease accuracy. This indicates that simple one-way recurrent neural networks may not capture necessary linguistic dependencies at an appropriate level at least for SA.

The organization of this paper is as follows. Section 2 describes our models of GCNN with Self-Attention and GRU with Self-Attention. Section 3 reports the experiments: 3.1 for the experimental condition, 3.2 for the experimental results, and 3.3 for discussion. Section 4 gives the conclusion.

2. Sentiment Analysis Systems (SAS's).

Two types of SAS's are compared: a GRU-based one and a GCNN-based one. Both have rarely been applied for SA in the way adopted in this study.

2.1. GRU-based sentiment analysis with Self-Attention.

SA is one of natural language processing applications and thus various recurrent neural networks have been employed to process a text. In this system, the GRU, which is one of the recurrent neural networks, processes an input text as a word sequence and generates a feature vector from it. Figure 1 shows the architecture of GRU.

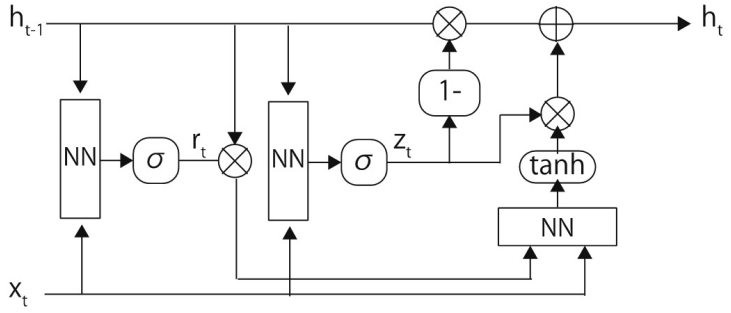

FIGURE 1. The architecture of GRU

The GRU has two gates, an update gate and a reset gate, and controls the data flow. The GRU is defined as follows:

$$
\begin{aligned}
z_t &= \sigma(W_z x_t + U_z h_{t-1} + b_z) \\
r_t &= \sigma(W_r x_t + U_r h_{t-1} + b_r) \\
h_t &= (1 - z_t) \odot h_{t-1} + z_t \odot \tanh(W_h x_t + U_h(r_t \odot h_{t-1}) + b_h)
\end{aligned}
$$

where z_t works as a switch for information update while r_t is used as a switch to reset past information, and the third formulation chooses necessary information according to input words in which $\odot$ means an element-wise product, W_* and U_* are adjustable parameters in the neural networks, and b_* is a bias parameter in the neural networks.

Self-attention is one of the attention mechanisms and calculates an attention weight only with the input data. Attention mechanisms are defined in general as follows:

$$Attention(Q, K, V) = Softmax(f(Q, K))V$$

Usual attention mechanisms have different variables for Q, K, and V, while Self-Attention sets the same variable to Q, K, and V. Figure 2 shows the architecture of Self-Attention, where the hidden states in GRU are input data and $f(Q, K)$ is defined with a 3-layer fully-connected neural network.

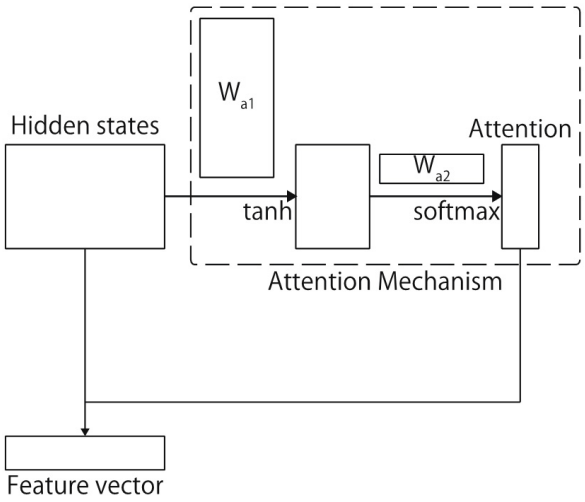

Figure 2. The architecture of Self-Attention

The GRU-based SAS is constructed with the GRU and Self-Attention. The architecture of the system is shown in Figure 3.

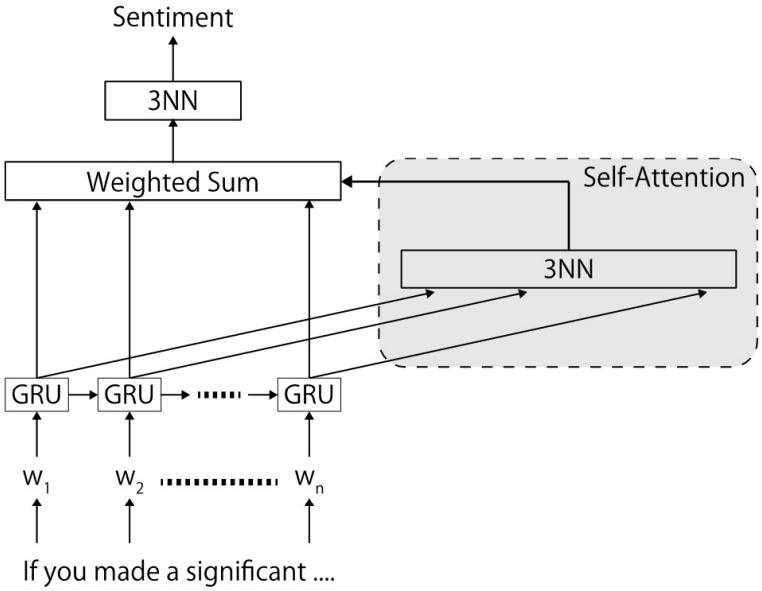

Figure 3. The architecture of GRU-based SAS

The system receives a sequence of words and embeds them into continuous vectors with the GRU. The system generates the final feature vectors for SA with Self-Attention. Finally, the 3-layer fully-connected neural network predicts the sentiment polarity as a classifier.

2.2. **GCNN-based sentiment analysis system.** We construct another SAS with Gated Convolutional Neural Networks (GCNN) with Self-Attention. GCNN can consider only the small size of the prior and posterior context defined as the kernel size and is different from GRU, which considers a text as sequential data with only prior contexts. GCNN consists of a convolutional module and a gate module, both implemented with convolutional neural networks. Figure 4 shows the architecture of GCNN.

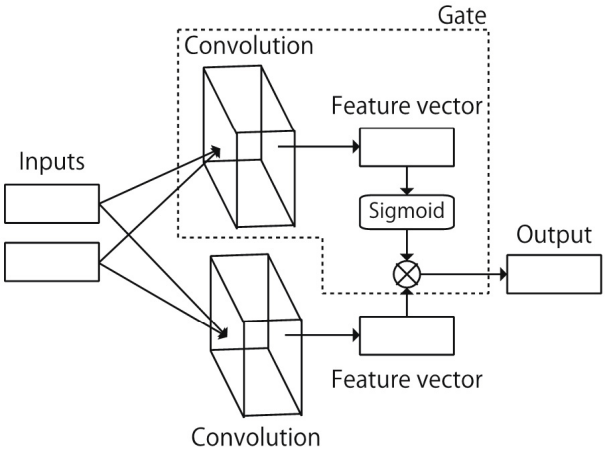

FIGURE 4. The architecture of GCNN

The GCNN consists of two convolutional layers and is defined as follows:

$$o = K_c * x_{i:i+k} \odot sigmoid(K_G * x_{i:i+k})$$

where K_c is a convolution kernel for word embedding; K_G is another for the gate; $*$ denotes a convolutional operation; and k is a manually defined kernel size in the convolutional neural network.

In the GCNN-based SAS, we combine the GCNN with Self-Attention. Figure 5 shows the whole architecture with GCNN with Self-Attention.

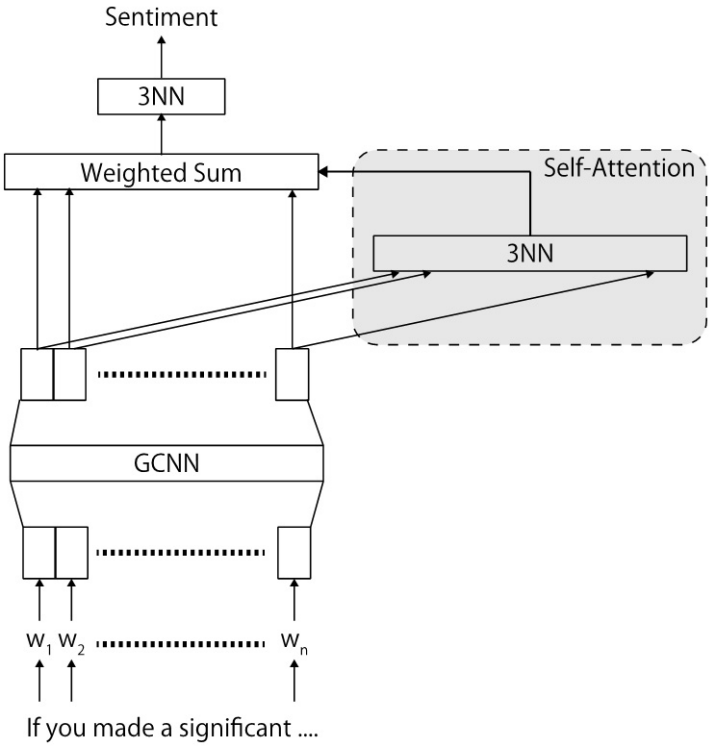

FIGURE 5. The architecture of GCNN-based SAS

The system receives a sequence of words and embeds them into continuous vectors with the GCNN. The system generates the final feature vectors of the input text with Self-Attention. Finally, the 3-layer fully-connected neural networks predict the sentiment polarity.

3. **Experiments.** We employed Amazon Product Reviews [10,11] to evaluate the two methods described in the previous section. In experiments, we measure performances for GCNN with three different kernel sizes and GRU with three different datasets.

3.1. **Experimental condition.** We used Amazon Product Reviews for evaluation experiments as training, validation, and test data. The original dataset consists of 24 product categories. Among them, we used three product categories: Electronics, Home and Kitchen, and Sports and Outdoors. For each category, 10,000 positive reviews and 10,000 negative reviews, or totally 20,000 reviews, were randomly extracted: 7,200 for training, 1,800 for validation, and 1,000 for test, as shown in Table 1. As such, no review is shared among training, validation, and test datasets.

Table 1. The contents of Amazon Product Reviews

Product category	Data type	Positive	Negative
Electronics (elec)	Training	7,200	7,200
	Validation	1,800	1,800
	Test	1,000	1,000
Home and Kitchen (home)	Training	7,200	7,200
	Validation	1,800	1,800
	Test	1,000	1,000
Sports and Outdoors (sports)	Training	7,200	7,200
	Validation	1,800	1,800
	Test	1,000	1,000

Each dataset is designed as balanced in terms of positive and negative reviews in order to avoid any unexpected categorical biases with regard to the proportion between positive and negative reviews among categories. The training data is used to train the proposed methods, and the validation data is employed to determine the optimal learning epoch. Using the test data, we evaluate trained models from the viewpoint of prediction accuracy and discuss performance.

All the models constructed in this study share the same hyper-parameters as in Table 2. For GCNN's, different kernel sizes, 2, 3, and 4, are also investigated.

The systems were implemented with PyTorch and trained with nVidia GeForce GTX TITAN X with which the performances of all methods were measured.

Table 2. The proposed methods settings

Parameter	Value
Word embedding size	256
Neural network for Self-Attention	256-128-1
Neural network for Classification	256-128-1
Stride (only for GCNN's)	1
Padding (only for GCNN's)	No
Minimum word occurrence frequency	3
Learning epoch	1,000
Optimization algorithm	ADAM
Learning rate	1.0e-5

3.2. **Experimental results.** Figure 6 shows learning curves under different conditions.

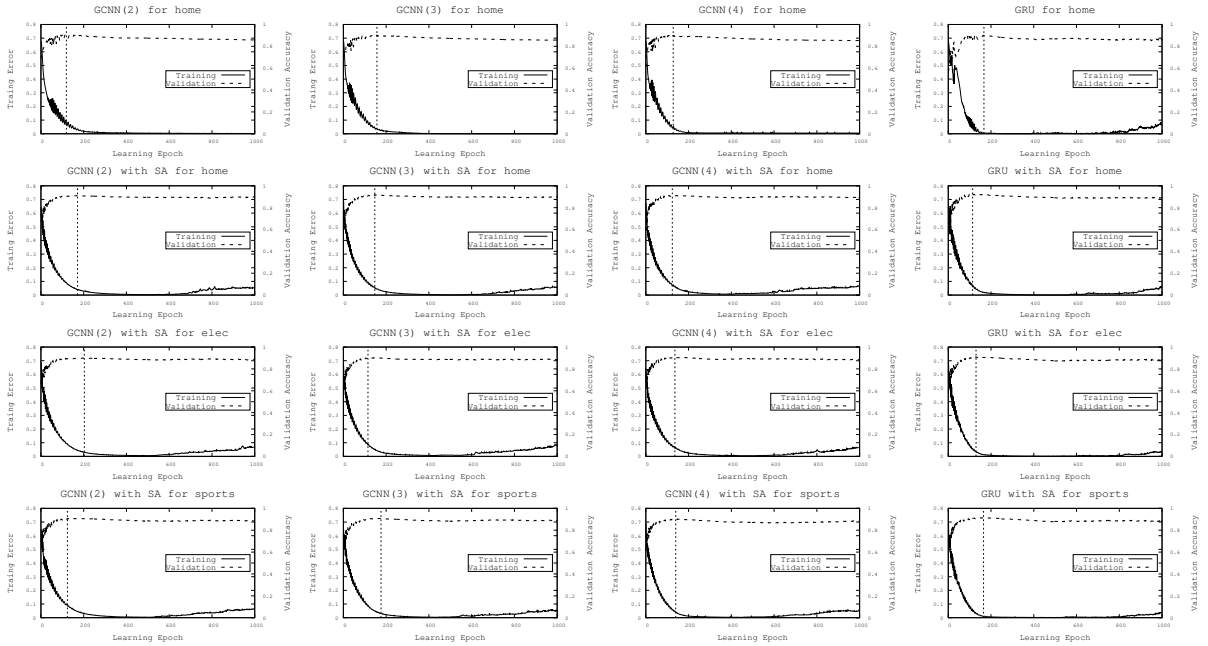

FIGURE 6. Learning curves in the proposed method for 4 product categories

The first row shows the performances of GCNN's and GRU without Self-Attention for Home and Kitchen. The vertical dashed line denotes the epoch where the training model achieves the highest accuracy in validation data. The second row shows the performances of GCNN's and GRU with Self-Attention for Home and Health. In GCNN's with Self-Attention, overfitting appears because the prediction accuracy for validation data is worse as epochs proceed. However, the model that achieved the highest validation accuracy is selected before overfitting. The third row shows the performances of GCNN's and GRU with Self-Attention for Electronics. All graphs are similar to the graphs with Home and Health dataset. All methods are trained enough from the viewpoint of learning epochs. The final row shows the performances of GCNN's and GRU with Self-Attention for Sports and Outdoors. The results have the same tendency as the other product categories.

The first, second, and third columns show the performances of GCNN with the kernel sizes of 2, 3, and 4, respectively. The fourth column shows the performances for GRU. Note that bigger kernel sizes for GCNN tend to produce overfitting in earlier epochs because bigger kernel sizes capture more detailed characteristics in the training dataset. The appropriate kernel size depends on the task and data. In this study, we selected a trained model based on validation accuracy. Using validation data, we can avoid overfitting and select the appropriate model from models in each epoch.

Table 3 shows the prediction accuracy of SAS's without Self-Attention. As shown, all four models achieved almost the same result.

On the other hand, bigger differences were seen with Self-Attention as in Table 4.

TABLE 3. Prediction accuracy in SAS's without Self-Attention

	Home and Kitchen
GCNN with the kernel size of 2	0.895
GCNN with the kernel size of 3	0.893
GCNN with the kernel size of 4	0.895
GRU	0.894

Table 4. Prediction accuracy in SAS's with Self-Attention

	Home and Kitchen	Electronics	Sports and Outdoors
GCNN with the kernel size of 2	0.901	0.891	0.879
GCNN with the kernel size of 3	0.904	0.896	**0.893**
GCNN with the kernel size of 4	**0.909**	0.895	0.882
GRU	0.883	**0.899**	0.876

GCNN's achieved the highest prediction accuracy in two datasets and the other dataset, Electronics, saw GRU be the best with only a slight difference. Home and Kitchen and Sports and Outdoors saw better results with GCNN with larger kernel sizes than GRU.

Note that GCNN's and GRU are totally different architectures, as described in Section 2. GCNN's process the data as a collection of small fragments defined with the kernel size and thus easy to be parallelized while GRU uses the input data as a sequence, and thus hard to be parallelized. This affects their learning time strongly. Table 5 shows each learning time for Home and Kitchen dataset. The smallest GCNN with the kernel size of 2 achieved approximately three times faster computation.

Table 5. Learning time in SAS's without Self-Attention

	Learning time (min)
GCNN with the kernel size of 2	468
GRU	1,315

3.3. **Discussion.** As Tables 3 and 4 show, Self-Attention improves the prediction accuracy for 1% or so both with GCNN's and GRU, which shows that Self-Attention works fine for the task of SA to add more complexity to improve function description ability. In particular, the improvement is clearer with GCNN's. Note that the Amazon Product Reviews contain some reviews that have very different ratings between the 5-scale evaluation, the ground-truth label in the dataset, and text reviews, which is why approximately 90% of the prediction accuracy is almost the realistic upper limit.

One of the reasons why GCNN's are better than GRU particularly with Self-Attention seems to come from their different architectures. As mentioned previously, GCNN's consider both prior and posterior contextual information within the kernel size while GRU considers the whole prior contextual information but not any posterior information. In the case of natural language processing, including SA, both prior and posterior contextual information that is mainly within a limited neighborhood is related to syntactic structure. Thus, the architecture of GCNN's seems to have contributed to their better performance. Note that GRU can also cope with posterior contextual information with a bi-directional GRU model, but its investigation is one of our future works.

With regard to GCNN's, wider kernel sizes mostly achieved better results, but it will not mean that the wider, the better. Crucial is whether the kernel size can contain contextual information enough for a task. SA, whose output is mainly the binary sentiment polarity, needs the proper size of context enough to include both a sentiment word and its modifying or predicative object, and the kernel size of 3 or 4 seems to be almost enough for this task. More detailed investigation on the appropriate kernel size will be tackled next.

4. **Conclusions.** This paper reported the comparison results of GCNN with Self-Attention with different kernel sizes and GRU with Self-Attention for SA of Amazon Review Datasets. The results showed that GCNN with Self-Attention is superior both in accuracy and in computation time. Interestingly, compared to models without Self-Attention, GCNN with Self-Attention is better as expected while GRU with Self-Attention is not

necessarily so. The overall results suggest that not only prior but also posterior information in text is desirable for sentiment classification, which is why GCNN is better than GRU, but that the appropriate size of context consideration may not always be the same.

Based on these results, more consideration on bi-directional GRU should be pursued, as well as consideration of the appropriate kernel size of GCNN with Self-Attention according to text types. It is also an issue of interest in how attention should be obtained. Currently, Self-Attention features are simply obtained through learning, but it is also possible to construct another level of learning to obtain other types of Self-Attention features by focusing on different, and presumably useful, information from what will be obtained through gated mechanisms.

REFERENCES

[1] L. Zhang, S. Wang and B. Liu, Deep learning for sentiment analysis: A survey, *WIREs Data Mining and Knowledge Discovery*, vol.8, no.4, 2018.

[2] B. Liu, *Sentinent Analysis and Opinion Mining,* Morgan & Claypool, 2012.

[3] S. Hochreiter and J. Schmidhuber, Long short-term memory, *SIAM J. Neural Computation*, vol.9, no.8, pp.1735-1780, 1997.

[4] J. Chung, C. Gulcehre, K. Cho and Y. Bengio, Empirical evaluation of gated recurrent neural networks on sequence modeling, *NIPS 2014 Workshop on Deep Learning*, 2014.

[5] N. Y. Dauphin, A. Fan, M. Auli and D. Grangier, Language modeling with gated convolutional networks, *Proc. of the 34th International Conference on Machine Learning*, vol.70, pp.933-941, 2017.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, N. A. Gomez, L. Kaiser and I. Polosukhin, Attention is all you need, *Advances in Neural Information Processing Systems*, vol.30, pp.5998-6008, 2017.

[7] Z. Lin, M. Feng, C. Nogueira dos Santos, M. Yu, B. Xiang, B. Zhou and Y. Bengio, A structured self-attentive sentence embedding, *Proc. of the 5th International Conference on Learning Representations (ICLR2017)*, 2017.

[8] D. Bahdanau, K. Cho and Y. Bengio, Neural machine translation by jointly to align and translate, *Proc. of the 3rd International Conference on Learning Representations (ICLR2015)*, 2015.

[9] M. Okada, H. Yanagimoto and K. Hashimoto, Sentiment analysis with gated CNN, *Proc. of ASEAN-AI2018*, 2018.

[10] R. He and J. McAuley, Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering, *Proc. of the 25th International Conference on World Wide Web (WWW'16)*, pp.507-517, DOI: 10.1145/2872427.2883037, 2016.

[11] J. McAuley, C. Targett, Q. Shi and A. van den Hengel, Image-based recommendations on styles and substitutes, *Proc. of SIGIR'15*, DOI: 10.1145/2766462.2767755, 2015.