Figure 1. An example of the image-based contents generated from a news document by News2Images. Left box includes an original online news document and right box represents the contents summarizing the news into three images. Red sentences in the left box are key sentences extracted by summarization and they are located in the black rectangle below the retrieved images in the right box.

key sentences, we define a score considering both the similarity to the core news contents and the diversity for the coverage on the entire contents of the news. The similarity and the diversity are computed using sentence embedding based on word2vec [10]. The image retrieval module searches the images semantically associated with the sentences extracted by the summarization module. The semantic association between a sentence and an image is defined as the cosine similarity between the sentence and the title of the news article which the image is attached in. Also, we use the hidden node values of the top fully connected layer of the convolutional neural networks (CNNs) [4] for each image as an image feature. Finally, the image-based content module generates a set of new images by synthesizing a retrieved image and the sentence corresponding to the image. These image-based contents generated can improve the readability and enhance the interests of mobile device users, compared to text-based news articles. The proposed News2Images has the originality in aspect of generating new contents suitable for mobile services by summarizing a long news document into not sentences but images even if there exist many methods for summarization [9] or text-to-image retrieval [1]. Figure 1 presents an example of the image-based content consisting of three synthesized images generated from a Korean online news article.

We evaluate the proposed News2Images on a big media data including more-than one million news articles served through a Korean media portal website, NAVER[2], in 2014. Experimental results show our method outperforms a baseline method based on word occurrence in terms of both quantitative and qualitative criteria. Moreover, we discuss some future directions for applying News2Images to personalized news recommender systems.

2. DEEP LEARNING-BASED FEATURE REPRESENTATION

Most news articles consist of a title, a document, and attached images. Mathematically, a news article x is defined as a triple $x = \{\mathbf{t}, S, V\}$, where **t**, S, and V denote a title, the set of document sentences, and an image set. V can be an empty set. A title **t** and a document sentence $\mathbf{s}$, $\mathbf{s} \in S$, are represented as a vector of word features such as occurrence frequency or word embedding. An image $\mathbf{v}$, $\mathbf{v} \in V$ is also defined as a vector of visual features such as Scale invariant feature transform (SIFT) [8] or CNN features. For representing a news article with a feature vector, we use deep learning in this study.

Many recent studies have reported that the hidden node values generated from deep learning models such as word embedding networks and CNNs are very useful for diverse problems including image classification [5], image descriptive sentence generation [14], and language models [12].

Formally, a word $\mathbf{w}$ is represented as a real-valued vector, $\mathbf{w} \in \Re^d$, where d is the dimension of a word vector. The vector value of each word is learned from a large corpus by word2vec [10]. This distributed word representation, called word embedding, is to not only characterize the semantic and the syntactic information but also overcome the data sparsity problem [6, 10]. It means that two words with similar meaning are located at a close position in the vector space. A sentence or a document can be represented as a real-valued vector as well. Sentence or document vectors can be generated by learning of deep networks, or they are calculated by pooling the word vectors included in the sentences. Here a sentence vector is calculated by average pooling:

$$s_i = \frac{1}{|s|} \sum_{w \in s} w_i , \quad (1)$$

where w and s denote a word and the set of words included in a sentence. Also, s_i and w_i are the i-th element of embedding vector $\mathbf{s}$ and $\mathbf{w}$ corresponding to s and w, respectively. Simple average pooling leads to lose sequence information of words. Therefore, the concatenation of multiple word vectors and the sliding window strategy can be used instead of simple pooling.

Image features can be generated for an input image by the CNNs learned from a large-scale image database. Typically, the hidden node values of the fully connected layer below the top softmax layer of CNNs are used as features. The CNN image features are also represented as a (non-negative) real-valued vector and they are known to be distinguishable for object recognition.

[2] www.naver.com

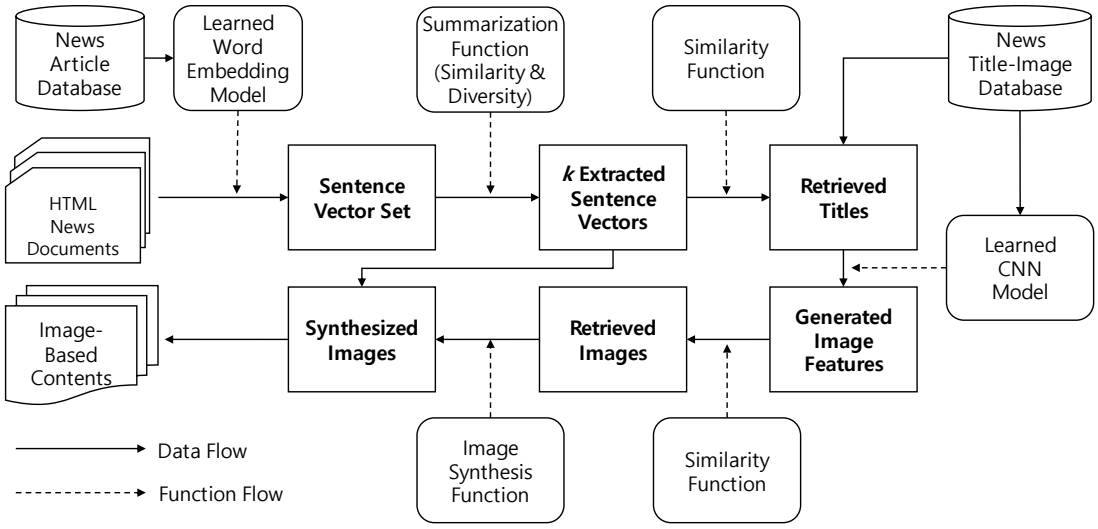

Figure 2. Overall flow of generating image-based contents from a news article via News2Images

3. NEWS-TO-IMAGES

News2Images is a method of generating image-based contents from a given news document using summarization and text-to-image retrieval. News2Images consists of three parts including key sentence extraction based on the single document summarization, key sentence-related image retrieval by associating images with sentences, and image-based content generation by synthesizing sentences and images. Figure 2 shows the overall framework of News2Images.

3.1 News Document Summarization

Document summarization is a task of automatically generating a minority of key sentences from an original document, minimizing loss of the content information [9]. Two approaches are mainly used for document summarization. One is abstraction which is to generate a few new sentences. Abstraction more precisely summarizes a document but still remains a challenging issue. The other is extraction, to select some core sentences from a document, and we use the extraction approach in this study. Also, the news summarization in this study belongs to single document summarization [7]. We assume two conditions for the summarization:

i) A news title is the best sentence consistently representing the entire content of the news.

ii) A news article consists of at least two sentences and the entire content is built up by composing its sentences' content.

For precisely summarizing a news document, thus, it is required that a summarized sentence set consists of the sentences not only semantically similar to its title but also covering the entire content with diverse words. We call the former similarity and the latter diversity.

Formally, a document S is defined as a set of its sentences, $S=\{\mathbf{s}_1,...,\mathbf{s}_M\}$, where M denotes the number of the sentences included in S. The i-th sentence $\mathbf{s}_i$ is represented as a real-valued vector, $\mathbf{s}_i \in \Re^d$, where d is the vector size, by word2vec and average pooling. Then, document summarization is formulated with

$$S_k^* = \arg\max_{S_k \subset S} \{\alpha \cdot f(S_k,S) + (1-\alpha)\cdot g(S_k,S)\} = \arg\max_{S_k \subset S} \{\alpha \cdot f(S_k,t) + (1-\alpha)\cdot g(S_k,S)\}, \quad (2)$$

$$\text{s.t. } f(S_k,S) = \sum_{\mathbf{s}\in S_k} f(\mathbf{s},S) \text{ and } g(S_k,S) = \sum_{\mathbf{s}\in S_k} g(\mathbf{s},S),$$

where $\mathbf{t}$ denotes the title of S, S_k and S_k^* are the set of k sentences extracted and an optimal set among S_k. $f(S_k,S)$ and $g(S_k,S)$ denote the similarity and the diversity functions, and α is the constant for moderating the ratio of two criteria.

The similarity $f(\mathbf{s},\mathbf{t})$ between a given sentence $\mathbf{s}$ and a news title $\mathbf{t}$ is defined as the cosine similarity between two sentence embedding vectors:

$$f(\mathbf{s},\mathbf{t}) = \frac{\mathbf{s}\cdot\mathbf{t}}{\|\mathbf{s}\|\|\mathbf{t}\|}. \quad (3)$$

For calculating the diversity, we partition the sentences of S into multiple subsets using a clustering method. Because a sentence vector implicitly reflects syntactic and semantic information, multiple semantically distinctive subsets are generated by clustering. For the j-th cluster C^j, we calculate the cosine similarity between all the sentences in C^j and the centroid of C^j. Because the cosine similarity can be negative, we consider a negative value as zero. This value is defined as the diversity:

$$g(\mathbf{s},C^j) = \frac{\mathbf{s}\cdot\mathbf{c}^j}{\|\mathbf{s}\|\|\mathbf{c}^j\|}, \quad (4)$$

where $\mathbf{c}^j$ denotes the centroid vector of C^j.

Finally, k sentences with the largest value defined in (2) are extracted as the summarization set for the given document. Here we set k to three, which means that a news article is summarized into three image-based contents.

3.2 Sentence-to-Image Retrieval

The second subtask is to retrieve the images representing semantics similar to the extracted sentences. Because we use the images attached in news articles, the title of a news including an image can be used as a description sentence of the image.

Therefore, the semantic similarity of an image to an extracted sentence is calculated by measuring the similarity between the image title vector and the sentence vector.

Formally, when an image feature vector set, $V=\{\mathbf{v}_1, \ldots, \mathbf{v}_N\}$, is given, the images similar to an extracted sentence $\hat{\mathbf{s}}$ are extracted:

$$\mathbf{v}^* = \arg\max_{\mathbf{v}\in V} \left\{ f\left(\hat{\mathbf{s}}, \mathbf{t}(\mathbf{v})\right) \right\} = \arg\max_{\mathbf{v}\in V} \left\{ \frac{\hat{\mathbf{s}} \cdot \mathbf{t}(\mathbf{v})}{|\hat{\mathbf{s}}||\mathbf{t}(\mathbf{v})|} \right\}, \quad (5)$$

where $\mathbf{t}(\mathbf{v})$ denotes the title of an image $\mathbf{v}$.

Due to the diversity, sentences which are not directly related to the title may be extracted as a core sentence. We assume that a title is "*Yuna Kim decided to participate in 2013 world figure skating championship*", and two extracted sentences are "*Yuna Kim will take part in the coming world figure skating championship*" and "*The competition will be held in February.*" In this case, the title is not semantically similar to the second sentence. Thus it is difficult to associate the second sentence with Yuna Kim's images. For overcoming this, we can additionally use the title vector of the news articles given as a query for pooling word vectors into a sentence vector. The use of the news title does not influence the summarization because the title vector is reflected on all the sentence vectors.

Instead of $\mathbf{v}^*$, we can generate a new image vector $\hat{\mathbf{v}}$ by averaging the vectors of top K images with the large similarity value. Then, $\mathbf{v}^*$ is selected as follows:

$$\mathbf{v}^* = \arg\max_{\mathbf{v}\in V} \left\{ f\left(\hat{\mathbf{v}}, \mathbf{v}\right) \right\}, \quad (6)$$

$$\hat{v}_i = \frac{R(\mathbf{v})}{\sum_{\mathbf{v}\in V_K} R(\mathbf{v})} v_i, \quad (7)$$

where v_i is the i-th element of $\mathbf{v}$ and $R(\mathbf{v})$ denotes a weight function proportional to the similarity rank. An image more similar to $\hat{\mathbf{v}}$ has a larger $R(\mathbf{v})$.

3.3 Image-Based Content Generation

Readability is a main issue of mobile content service. Therefore we generate new image-based contents instead of using the retrieved images for improving the readability and enhancing the users' interests. An image-based content includes continuous series of synthesized images where the retrieved images and their corresponding sentences are merged. Figure 1 illustrates an example of the image-based contents from a news document.

4. EXPERIMENTAL RESULTS

4.1 Data and Parameter Setting

We evaluate the proposed News2Images on a big media data including over one million Korean news articles, which are provided by a media portal site, NAVER, in 2014. In detail, the word vectors are learned from all the news documents and the CNN models for constructing image features are trained from approximately 220 thousands of news images, which are related to 100 famous entertainers, movie stars, and sports stars. Also, 6,967 news articles are used as the validation set for evaluating the performance. Three key sentences were extracted from a news article including more than three sentences and we used all the sentences in the news consisting of less than three sentences. Then, 20,224 image-based contents were generated from validation news data in total.

We used the word2vec for word embedding and modified GoogleNet implemented in Caffe for CNN features [4]. The word vector and image feature sizes are 100 and 1024, respectively. For error correction in learning CNNs, we set the label of an image to the person name in the image. Thus, the size of the class label set is 100. The learned CNN model for generating image features yields 0.56 and 0.79 as Top-1 and Top-5 classification accuracies, respectively. This indicates that the generated image features are distinguishable enough to be used for associating images and sentences. The number of clusters for the diversity in summarization was set to 3 and the constant moderating the similarity and the diversity is 0.9.

For comparisons, we used a word occurrence vector based on TF/IDF as a baseline in computing the similarity between sentences and titles, instead of a word embedding vector. TF/IDF has been widely used for text mining, and thus we can verify the effects of deep learning-based word features.

Table 1. Accuracy of the baseline method and News2Images

Classification	Baseline (TF/IDF)	News2Images
Correct #	14,020/20,224	**18,908/20,224**
Accuracy	0.693	**0.935**
Cosine Similarity	0.636	**0.866**

We set the number of images for averaging in (6), M to 1 both two methods. The window size of the words is 1. Both methods use news titles in pooling word vectors into sentence vectors.

4.2 Content Generation Accuracy

Human efforts are still essential for precisely measuring how similar the generated image-based contents are semantically to the news document given as a query. Instead of manual evaluation by

Table 2. Accuracies according to the usage of news titles

News title	No used	Used
Correct #	13,896/20,224	**18,908/20,224**
Accuracy	0.687	**0.935**

Table 3. Accuracies according to the size of retrieved images size for generating a new image feature

Image size	K=1	K=3
Correct #	**18,908/20,224**	18,791/20,224
Accuracy	**0.935**	0.929

Table 4. Accuracies according to the weight for proper nouns

Proper noun weight	PW = 1.0	PW=10.0
Correct #	18,908/20,224	**19,191/20,224**
Accuracy	0.935	**0.950**

PW denotes the weight of proper nouns.

Table 5. Accuracies according to word vector window sizes

Window size	$\|W\|$=1	$\|W\|$=3
Correct #	**18,908/20,224**	18,743/20,224
Accuracy	**0.935**	0.927
Cosine Similarity	**0.866**	0.833

$|W|$ denotes the number of concatenated word vectors.

humans, we consider a classification problem as the similarity evaluation. That is, for a given extracted news sentence, we consider that the retrieved image is similar to the sentence when the persons referred in the sentence exist in the image. It is reasonable because this means the method provides diverse images of a movie star for users when a user reads a news about the star.

Table 1 compares the classification accuracy of the baseline and the proposed method. As shown in Table 1, News2Images outperforms the baseline method. This indicates the word embedding features used in News2Images more precisely represent semantics, compared to TF/IDF-based features. Also, we compared the cosine similarity between the titles of the retrieved images and the extracted sentences using their word embedding vectors. The values are averaged on the titles of 20,224 retrieved images. We can find that our method retrieves the images more semantically similar to the extracted sentences.

4.3 Effects of Parameters on Performance

We compare the accuracies of the generated contents under four parameters including i) the use of news title for pooling word vectors into a sentence vector, ii) the number of retrieved images for an image feature, iii) the weight for proper nouns, and iv) the size of concatenated word vectors. Table 2 presents the accuracy improvement when the title of the summarized news documents is used. We found that the use of the news title dramatically improves the accuracy as 30% compared to the case in which the titles are not used. Interestingly, News2Images not using titles provides the similar performance to the baseline method using titles. Table 3 shows the effects of averaging multiple image features on sentence-to-image retrieval. This indicates that generating a new image feature from multiple image features has no effect on enhancing the performance. To give more weight to proper nouns can improve the quality of the image-based content generation because proper nouns are likely to be a key content of the news. The results in Table 4 support this hypothesis. The number of concatenated word vectors rarely influences the accuracy. We indicate that the information on word sequences is not essential to classify the person in the images from Table 5.

4.4 Image-Based Contents as News Summarization

Figure 3 illustrates good and bad examples of image-based contents from news articles. Most of the images are related to the

Sentences	News2Images	Baseline
Park, the home run leader of KBO, hit the 34th home run in this season.		
Son of Leverkusen played as a starter forward in this game for 60 minutes until substituted with Yurchenko		
Today, Ryu pitched 7 innings, allowed two runs and 9 hits, and got 7 kills against the Chicago Cubs at the home game, and thus ERA becomes 3.39.		
Lee, Hyori is practicing yoga with a grave look in the released photo.		
Chu, Soohyun showed her bodyline at the swimming pool scene in the 18th episode of the drama.		

Figure 3. Examples of image-based contents generated from the summarization sentences extracted from news articles by News2Images and the baseline method. Images with a red border are very similar to the sentences. Blue bordered images include the persons referred in the given sentences but represent contents different from the sentences.

news contents but the sentences including polysemy or too many words are occasionally linked to images not relevant to the sentences. This is caused that one word is represented as only one vector regardless of its meaning. Also, the representation power of pooling-based sentence embedding can be weaken due to the property of average pooling when a sentence consists of too many words.

5. DISCUSSION

We proposed a new method for summarizing news articles into image-based contents, News2Images. These image-based contents are useful for providing the news for mobile device users while enhancing the readability and interests. Deep learning-based text and image features used in the proposed method improved the performance as approximately 24% of the classification accuracy and 0.23 of the cosine similarity compared to the TF/IDF baseline method. Our study has an originality in aspect of generating new image contents from news documents even if many studies on summarization or text-to-image retrieval have been reported.

This method can be applied to a personalized news recommender system adding user preference information such as subject categories and persons preferred by a user and feedback information into the method. In detail, we can give a weight to words related to subjects or persons preferred by a user when generating sentence vectors. This strategy allows the sentences which the user is likely to feel an interest in to have higher score in summarization and retrieval, thus exposing the photos which the user prefers.

Evaluation should be also improved. Although we evaluate the proposed method with the cosine similarity-based measure and the classification accuracy, it has a limitation for precisely measuring the similarity between the news articles and the image contents generated. It is required to make a ground truth dataset by humans, which not only helps to more precisely evaluate the model performance and can be used as a good dataset for recommendation as well as image-text multimodal learning. Furthermore, we will verify the effects of News2Images on the improvements of the readability through human experiments as future work.

The proposed method can be improved by adding the module of efficiently learning a common semantic hypothesis represented with sentences and images using a unified model [14].

ACKNOWLEDGMENTS

6. REFERENCES

[1] Datta, R., Joshi, D., Li, J. and Wang, J. Z. 2008. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys* (*CSUR*). 40, 2. 5.

[2] Hinton, G. et al. 2012. Deep neural networks for acoustic modeling in speech recognition, *IEEE Signal Processing Magazine*. 29, 6. 82-97.

[3] Irsoy, O. and Cardie C., Deep recursive neural networks for compositionality in language. In *Advances in Neural Information Processing Systems* 2014. 2096-2104.

[4] Jia, Y. et al. 2014. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia* 2014. 675-678.

[5] Krizhevsky, A., Sutskever, I., and Hinton, G. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems* 2012. 1097-1105.

[6] LeCun, Y., Bengio, Y., and Hinton, G. 2015. Deep learning. *Nature*. 521, 7553. 436-444.

[7] Lin, C.-Y. and Hovy, E. 2002. From single to multi-document summarization: a prototype system and its evaluation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (*ACL* '02). 457-464.

[8] Lowe, D. G. 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*. 60, 2. 91-110.

[9] McDonald, R. 2007. *A study of global inference algorithms in multi-document summarization*. Springer Berlin Heidelberg. 557-564.

[10] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems* 2013. 3111-3119.

[11] Salakhutdinov, R., Mnih, A., and Hinton, G. 2007. Restricted Boltzmann machines for collaborative filtering. In *Proceedings of the 24th International Conference on Machine Learning* (*ICML* 2007). 791-798.

[12] Socher, R., Lin, C. C.-Y., Ng, A., and Manning, C. 2011. Parsing natural scenes and natural language with recursive neural networks. In *Proceedings of the 28th International Conference on Machine Learning* (*ICML*-11). 129-136.

[13] Van den Oord, A., Dieleman, S., and Schrauwen, B. 2013. Deep content-based music recommendation, In *Advances in Neural Information Processing Systems* 2013. 2643-2651.

[14] Xu, K., Ba, J., Kiros, R., Courville, A., Salakhutdinov, R., Zemel, R., and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of 32th International Conference on Machine Learning* (*ICML'15*).