Received: 5 December 2018 | Accepted: 7 May 2019

DOI: 10.4218/etrij.2018-0621

ORIGINAL ARTICLE

ETRI Journal WILEY

# Image classification and captioning model considering a CAM-based disagreement loss

**Yeo Chan Yoon**[1,2] | **So Young Park**[3] | **Soo Myoung Park**[1] | **Heuiseok Lim**[2]

[1]SW Content Research Laboratory, Electronics and Technology Research Institute, Daejeon, Rep. of Korea

[2]Department of Computer Science and Engineering, Korea University, Seoul, Rep. of Korea

[3]Department of Game Design and Development, Sangmyung University, Seoul, Rep. of Korea

**Correspondence**
Heuiseok Lim, Department of Computer Science and Engineering, Korea University, Seoul, Rep. of Korea.
Email: limhseok@korea.ac.kr

**Funding information**
This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (no. 2016-0-00010-003, Digital Content In-House R&D).

Image captioning has received significant interest in recent years, and notable results have been achieved. Most previous approaches have focused on generating visual descriptions from images, whereas a few approaches have exploited visual descriptions for image classification. This study demonstrates that a good performance can be achieved for both description generation and image classification through an end-to-end joint learning approach with a loss function, which encourages each task to reach a consensus. When given images and visual descriptions, the proposed model learns a multimodal intermediate embedding, which can represent both the textual and visual characteristics of an object. The performance can be improved for both tasks by sharing the multimodal embedding. Through a novel loss function based on class activation mapping, which localizes the discriminative image region of a model, we achieve a higher score when the captioning and classification model reaches a consensus on the key parts of the object. Using the proposed model, we established a substantially improved performance for each task on the UCSD Birds and Oxford Flowers datasets.

**KEYWORDS**
deep learning, image captioning, image classification

## 1 | INTRODUCTION

Computer vision and natural language processing have frequently been tackled as independent studies during the past decade. Humans often describe objects using text and images. Whereas images provide comprehensive information, texts summarize the descriptive features of objects, and both types of information help people to recognize objects, such as birds and flowers. Intuitively, visual descriptions and images are strongly connected and can therefore be used together to better understand images. Image captioning is a task to describe objects in images, and current state-of-the-art image captioning approaches [1–13] have achieved successful results by connecting computer vision and natural language with deep learning approaches. These approaches employ a convolutional neural network (CNN) encoder–recurrent neural network (RNN) decoder method, which utilizes a CNN to generate a fixed-length vector representation [14] and an RNN to generate a visual description using this vector representation. Image-to-text encoder-decoder approaches, which translate image features into textual descriptions, yield fascinating results. This approach has achieved success in many recent related studies, such as visual QA [8–10], generating explanations of deep learning algorithms [4], and zero-shot retrieval [15,16]. Conversely, text information can be utilized for image processing, such as image classification [15,17]. Most previous methods have only focused on one task with two types of modalities. However, closely linked tasks, such

as image captioning and image classification, can have a positive impact on each other.

The main objective of this study is to simultaneously improve the performances of both visual description generation and image classification using the descriptions. Intuitively, the visual descriptions of images can be exploited to help classify an object in images and vice versa. Fine-grained classification is an attractive method of demonstrating our approach, because distinctive visual descriptions help in recognizing objects from images. We apply our method to fine-grained image classification and an image-captioning task to demonstrate its validity.

In this study, we achieve our goal by proposing an end-to-end joint learning method with a novel loss function that mediates each task. An intermediate layer is utilized to represent multiple forms of information on an object in the target images. Through a disagreement loss function based on class activation mapping (CAM), which localizes the discriminative image region of a model, we achieve a higher score when the captioning and classification model reaches a consensus on the key parts of the object, further improving the results. By sharing an intermediate representation, we achieve a state-of-the-art fine-grained image classification performance and improve the performance for a caption generation task.

To the best of our knowledge, this study represents the first attempt to improve both image recognition and image captioning using an end-to end joint learning model.

## 2 | RELATED WORK

### 2.1 | Deep image captioning

The author of [1] proposed an image-to-text encoder-decoder model for image captioning tasks. The encoder-decoder model first extracts high-level visual features from a CNN trained on the image classification task, and then feeds the visual features into an RNN model to predict subsequent words of a caption for a given image. In recent years, a variety of successive models [2–16,18–20] have achieved promising results. Semantic concept analysis, or attribute prediction [17,21], is a task closely related to image captioning, because attributes can be interpreted as a basis for descriptions. To generate captions, semantic concepts or attributes of objects in images are detected and utilized as inputs of the RNN decoder [3,6,12,20,22]. Latent topics [6], cross domains [22], and inter-attribute correlations [12] are considered to improve the results. Meanwhile, some approaches [5,15,17–19] have adopted multimodal embedding, which represents multiple aspects of objects with pictures and descriptions as the latent semantics of objects. Language features from an RNN decoder and image features from a CNN encoder are embedded in a multimodal space. The learned embedding is then utilized to guide the caption generation or zero-shot retrieval. In this study, we also adopt a multimodal unit to represent both the images and descriptions. Unlike the previous studies outlined above, our method allows multimodal embedding learning to mediate between two different goals: image classification and image captioning.

Because a loss function is one of the key factors in a deep learning model, various loss functions have been introduced for image captioning tasks. These loss functions are designed to suit their own algorithms. To integrate topic representation into the training process, the authors of [3] introduced an interpretive loss, which helps to improve the interpretability of the learned features. A loss function that encourages generated sentences to include class discriminative information was designed to explain class discriminative characteristics for bird images [4]. We also designed our own loss function for the multimodal layer, which represents the aspects of both images and text. Objects belonging to the same class share many common features in both images and visual descriptions. Because a multimodal embedding represents the latent semantics of an input image with the aid of descriptions and image contents, it is desirable for the key visual object parts of each model's predictions to be close. To address this problem, we apply the CAM [23] method with a cosine distance loss [24] for image embedding.

### 2.2 | Fine-grained classification

Fine-grained classification is a challenging task, which assigns objects to subordinate classes. The objects are visually similar to each other, and can only be discriminated through subtle details. Most fine-grained classification systems employ visual features of images to classify objects using a CNN [25–29], and subordinate classes from various domains such as flowers, birds, dogs, aircrafts, and cars can be successfully recognized using these approaches. To improve the classification performance, some approaches employ hierarchical semantic information such as a taxonomic rank [30], the semantic distance of WordNet [31], and text [15,17]. In this study, we employ the visual descriptions of images to classify each visually similar object, because the visual descriptions contain discriminative and summarized characteristics of objects.

### 2.3 | Multitask learning

Multitask learning is a method of improving the performance by simultaneously learning several related tasks, such as face and gender detection [32], or POS tagging and dependency parsing [33]. This approach was inspired by the human learning process, which easily learns similar tasks with little experience. There are two types of approaches to multitask learning: hard parameter sharing

and soft parameter sharing. In hard parameter sharing, all hidden layers are shared between all tasks, while keeping several task-specific output layers [34]. In contrast, in soft parameter sharing each task has its own layers and weights. Rather than sharing parameters, each task affects the others by comparing or transferring knowledge [32,35]. In this study, we employ hard parameter sharing with a disagreement loss mediated between two related tasks: image captioning and image classification.

## 3 | METHODOLOGY

In this section, we present the key components of our image captioning and classification system. We first overview the proposed model, which aims to learn image classification and captioning at the same time. We then present a detailed end-to-end training algorithm, which incorporates a disagreement loss to improve the performance with the aid of multimodal features. The disagreement loss is based on CAM, which provides the rationale for the model predictions. The details are introduced in the next section.

### 3.1 | Image classification and captioning

The proposed model consists of an image encoder, intermediate layer, image classifier, and image captioner, as shown in Figure 1. The image encoder represents a CNN-based encoder with the weights pre-trained by ImageNet. This converts a given image into a fixed-length image feature vector. The intermediate layer transforms the image feature vector into another image feature vector, creating an influence between the image classifier and image captioner. The image classifier is a single-layer perceptron-based classifier, which assigns the class label to the image. The image captioner is an RNN-based captioner, which generates the caption describing the image. Like the image encoder-decoder model [1], the proposed model includes both a CNN model to transform an image into its image feature vector and an RNN model to generate a caption from the image feature vector. Whereas the image classifier and image captioner learn independently of each other in the image encoder-decoder model, the proposed model leads to both learning at the same time. Therefore, the proposed model can obtain both the class label and caption for a given image in parallel. In addition, the intermediate layer mediates between the image classifier and image captioner in the proposed model.

### 3.2 | End-to-end learning algorithm

The proposed image classification and captioning model utilizes the intermediate multimodal layer as a key component of joint learning. According to the end-to-end learning algorithm presented in Algorithm 1, the proposed model updates the intermediate layer $f_{\text{medi}}$, image classifier $f_{\text{class}}$, and image captioner $f_{\text{caption}}$. For the learning algorithm using mini-batch training, the training set $S^*$ is divided into $M$ mini-batch training subsets, where each mini-batch training set $S_i$ consists of $N$ pairs of an image, its class label, and its caption, as $S_i = \{(img_j, class_j, caption_j)\}$. Given a mini-batch training set $S_i$, the proposed model encodes each input image based on both the image encoder $f_{\text{encode}}$ and intermediate layer $f_{\text{medi}}$, as described in lines 6 and 7 of Algorithm 1. The encoded vector can potentially represent the relations between the image, its class, and its caption by embedding both the latent image representation and latent text representation simultaneously. As presented in lines 8 and 9 of Algorithm 1, the image classifier $f_{\text{class}}$ and image captioner $f_{\text{caption}}$ predict the class label and caption candidates, respectively. We optimize the intermediate layer in terms of three criteria. First, the intermediate

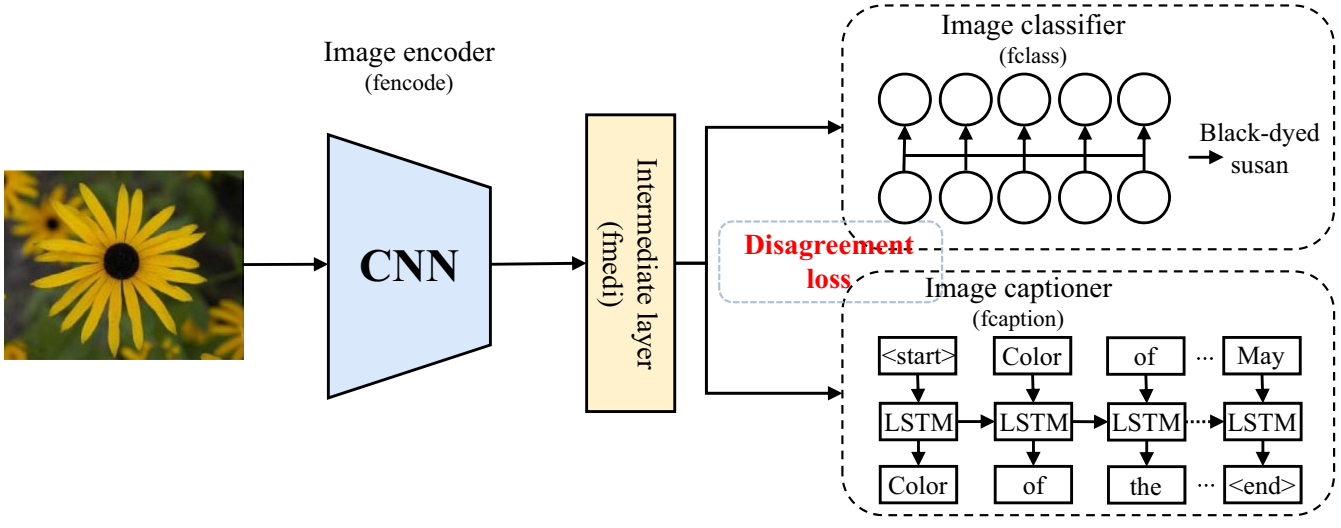


**FIGURE 1** Proposed image classification and captioning models

layer represents the aspects of an image. Second, it also represents the aspects of the image's matching visual descriptions. Third, the class discriminative information should be contained in the embedding. An objective function reflecting these criteria is defined as follows:

$$\frac{1}{N}\sum_{j=1}^{N}\log P\left(class_j \mid img_j\right), \tag{1}$$

$$\frac{1}{N}\sum_{j=1}^{N}\log P\left(caption_j \mid img_j\right), \tag{2}$$

$$\begin{aligned}\frac{1}{N}\sum_{j=1}^{N}&\log P\left(cam_a^c\left(img_j\right) \mid cam_p^l\left(img_j\right)\right)\\ &+\log P\left(cam_a^l\left(img_j\right) \mid cam_p^c\left(img_j\right)\right).\end{aligned} \tag{3}$$

The respective summations denote objective functions for the image classification, image captioning, and disagreement function. In Algorithm 1 and summations (1), (2), and (3), $N$ indicates the number of pairs in the batch training set. In addition, $class_j$ and $caption_j$ represent the predicted class label and caption for the $j$th image $img_j$, respectively. Here, $cam(img_j)$ is the CAM of $img_j$. The subscript $a$ indicates that the target class of the CAM output is the answer class, and $p$ indicates that the target class is the model prediction. The superscripts $c$ and $l$ indicate that the CAM output is generated by the captioning or classification model, respectively. That is, $cam_a^c(img_j)$ and $cam_p^l(img_j)$ represent the CAMs of the correct caption and the classification model prediction for the $j$th image, respectively.

**Algorithm 1** pseudocode for our algorithm

1: Input $M \times N$ Pairs of (img, class, caption)
2: Initialize $f_{\text{encode}}$ with the ImageNet pretrained model.
3: Initialize $f_{\text{class}}$, $f_{\text{caption}}$, $f_{\text{medi}}$ with uniform distribution
4: **for** $i = 1$ to $M$ do
5: **for** $j = 1$ to $N$ do
6: $img_j^{'} \Leftarrow f_{\text{encode}}(img_j)$
7: $img_j^{''} \Leftarrow f_{\text{medi}}(img_j^{'})$
8: $class_j \Leftarrow f_{\text{class}}(img_j^{''})$
9: $caption_j \Leftarrow f_{\text{caption}}(img_j^{''})$
10: **end for**
11: $L_{\text{class}} \Leftarrow Equation(1)$
12: $L_{\text{caption}} \Leftarrow Equation(2)$
13: $L'_{\text{medi}} \Leftarrow Equation(3)$
14: $L_{\text{medi}} \Leftarrow L_{\text{class}} + L_{\text{caption}} + L_{\text{medi}}{}'$
15: $f_{\text{class}} \Leftarrow f_{\text{class}} + \partial L_{\text{class}}/\partial f_{\text{class}}$
16: $f_{\text{caption}} \Leftarrow f_{\text{caption}} + \partial L_{\text{caption}}/\partial f_{\text{captoin}}$
17: $f_{\text{medi}} \Leftarrow f_{\text{medi}} + \partial L_{\text{medi}}/\partial f_{\text{medi}}$
18: **end for**

To consider class discriminative information, we ensure that the model's rationale is the same for the ground truth and predicted labels. Because the CAM outputs of the models highlight the most discriminating visual parts of each model's prediction, these can be utilized as the model rationale. To enable the two tasks to positively affect one another, we cross compared the rationale, as in summation (3).

To maximize the objective functions, the proposed model optimizes the intermediate layer, image classifier, and image captioner as described in lines 15–17 of Algorithm 1. Given the entire training set, the image classifier $f_{\text{class}}$ learns to maximize the log probability term in (1), and the image captioner $f_{\text{caption}}$ simultaneously learns to maximize the log probability term in (2). Likewise, the intermediate layer $f_{\text{medi}}$ learns to maximize the log probability term in (3), where the weight of the class discriminative neuron can be influenced by the text features and that of the text descriptive neuron can be influenced by the image features in the intermediate layer.

## 3.3 | CAM-based disagreement loss

For both the image classification and image captioning tasks, it is important to detect the discriminative parts of the given image. Nevertheless, it has proven difficult to reach a trainable consensus on the discriminative parts for the image classification and image captioning tasks. Fortunately, CAM [23] approaches can highlight the discriminative visual parts detected by the CNN. By comparing the heat maps of each model using CAM, we measure the agreement between the two models in the training direction. The neurons in the convolutional layers tend to seek class-specific information from the given image. To analyze which neurons conclusively influence the final decision, CAM utilizes the gradient information extracted from the final convolutional layer of the CNN.

Figure 2 presents CAM visualization examples, in which the image classifier and image captioner focus on different visual components for the same given image. In the top two images of Figure 2, the image classifier fails to classify the bird into the correct class, because the image classifier on focuses on the wings of the bird, although the image captioner appropriately focuses on its crown, bill, back, and wings. In the bottom two images of Figure 2, the image captioner fails to generate a suitable caption, because it only focuses on a partial component of the wrinkled petal of the flower, although the image classifier focuses on its yellow style and red petals.

To obtain the class activation map, the score $y^c$ for class $c$ is differentiated with respect to feature map $A^k$ of the last convolutional layer. These gradients are summed over to obtain the weight $a_k^c$ for the feature map $k$ and target class $c$:

$$a_k^c = \frac{1}{Z}\sum_i\sum_j\frac{\partial y^c}{\partial A_{ij}^k}. \tag{4}$$

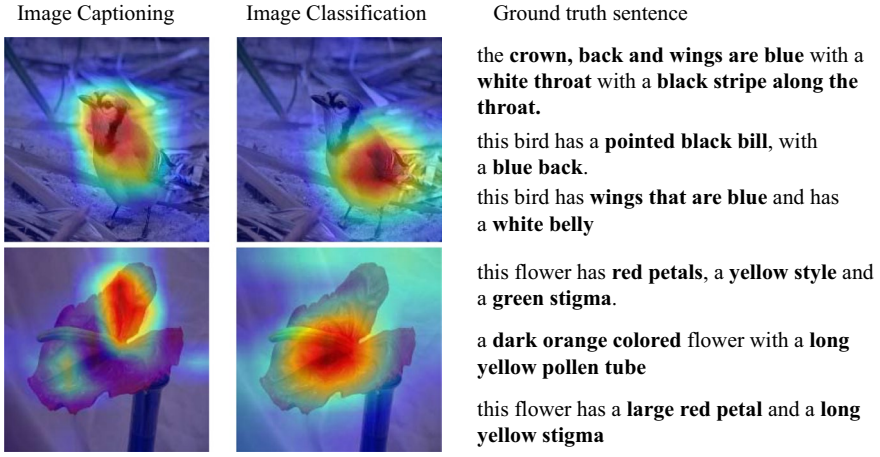


**FIGURE 2** CAM highlights the discriminative visual parts of each model. In the above example, the image captioning and image classification models fail to reach a consensus on the key parts of the object in each image. By imposing a penalty in such case, we can improve the performances of both models

Equation (5) represents a combination of forward activation maps using the weights, for the purpose of obtaining a class activation map value $L^c_{\text{CAM}}$ for class $c$:

$$L^c_{\text{CAM}} = \sum_k a^c_k A^k. \tag{5}$$

Given the same image, we assume that the same visual component is focused upon by both the image classifier and the image captioner. To reduce the disagreement between these two visual components, this assumption is represented by minimizing the disagreement loss from (6). This equation is calculated by comparing $cam^l_p(img_j)$ and $cam^c_a(img_j)$, as well as $cam^l_a(img_j)$ and $cam^c_p(img_j)$. For the disagreement loss optimization, the disagreement loss from (6) is calculated using the cosine distance loss [24].

$$\begin{aligned} L_{\cos} = &\left(1-\cos\left(cam^c_a\left(img_j\right), cam^l_p\left(img_j\right)\right)\right) \\ &+\left(1-\cos\left(cam^l_a\left(img_j\right), cam^c_p\left(img_j\right)\right)\right). \end{aligned} \tag{6}$$

In (6), $cam^l_p(img_j)$ and $cam^c_a(img_j)$ represent the CAMs of the predicted class label of the image and the correct caption, respectively. In addition, $cam^l_a(img_j)$ and $cam^c_p(img_j)$ represent the CAMs of the correct class label and the predicted caption, respectively.

$$\cos(\mathbf{l},\mathbf{c}) = \frac{\mathbf{l}\times\mathbf{c}}{\sqrt{\mathbf{l}^2\mathbf{c}^2}}, \tag{7}$$

$$\begin{aligned} \frac{\partial \cos\text{sim}(\mathbf{l},\mathbf{c})}{\partial l_1} &= \frac{\partial}{\partial l_1}\frac{l_1\times c_1+\cdots+l_n\times c_n}{|\mathbf{l}|\cdot|\mathbf{c}|} \\ &= \frac{\partial}{\partial l_1} l_1\times c_1\times\left(l_1^2+l_2^2+\cdots+l_n^2\right)^{-\frac{1}{2}}\times|\mathbf{c}|^{-1} \\ &= \frac{c_1}{|\mathbf{l}|\times|\mathbf{c}|} - \frac{l_1\times c_1}{|\mathbf{l}|\times|\mathbf{c}|}\times\frac{l_1}{|\mathbf{l}|^2}, \end{aligned} \tag{8}$$

$$\therefore \frac{\partial \cos\text{sim}(\mathbf{l},\mathbf{c})}{\partial \mathbf{l}} = \frac{\mathbf{l}}{|\mathbf{l}|\times|\mathbf{c}|} - \cos(\mathbf{l},\mathbf{c})\times\frac{\mathbf{l}}{|\mathbf{l}|^2}. \tag{9}$$

For the back propagation algorithm, the cosine similarity is derived as shown above. In these equations, $l_i$ represents the $i$-th value of the class label vector $\mathbf{l}$ and $c_i$ represents the $i$-th value of the caption vector $\mathbf{c}$.

## 4 | EVALUATION

### 4.1 | Experimental dataset

To verify the classification and captioning performance of the proposed model, we utilized two datasets: the Oxford Flowers 102 [15,36] and Caltech UCSD Birds 200–2011 [37] datasets. The Flowers dataset contains 8,189 flower images with 102 classes, while the Birds dataset includes 11,788 bird images with 200 different classes. Each dataset contains sentences describing each image, such as "this bird has a pointed black bill, with a blue back," or "this flower has red petals, a yellow style, and a green stigma." For a fair evaluation, the dataset is divided into a training, validation, and test set, as shown in Table 1, where each cell includes the number of images and the number of classes is written within parentheses.

### 4.2 | Evaluation measures

To evaluate the image classification performance of each model, we measure the top-1 and top-5 accuracies, where the accuracy indicates the number of correct candidate class labels in the top 1 or top 5, divided by the number of images in the test set. For the image captioning performance of each model, we measure the following four metrics: BLEU [38], ROUGE [39], METEOR [40], and CIDEr [41].

$$\log \text{BLEU} = \min\left(1-\frac{r}{c}, 0\right) + \sum_{n=1}^{N}(w_n \log p_n), \tag{10}$$

**TABLE 1** Training, validation, and test sets

| Set | Task | Train | Validation | Test |
|---|---|---|---|---|
| Flower | Image classification | 4222 (82) | 1406 (82) | 1406 (82) |
| | Image captioning | 5878 (62) | 1156 (20) | 1155 (20) |
| Bird | Image classification | 5313 (150) | 1771 (150) | 1771 (150) |
| | Image captioning | 5894 (100) | 2961 (50) | 2933 (50) |

$$P_n = \frac{\sum_{c \in C^*} \sum_{n\text{-gram} \in C} count_{\text{match}}(n\text{-gram})}{\sum_{C \in C^*} \sum_{n\text{-gram} \in C} count(n\text{-gram})}. \quad (11)$$

BLEU measures the geometric average of the n-gram precisions, based on the positive uniform weights $w_n$, precision $P_n$, length $r$ of the reference answer, and length $c$ of the candidate answer. The precision $P_n$ is calculated as the number of matched $n$-grams divided by the number of candidate $n$-grams in the candidate answer, where the precision $P_n$ is based on n-grams of up to length $n$.

$$R_n = \frac{\sum_{R \in R^*} \sum_{n\text{-gram} \in R} count_{\text{match}}(n\text{-gram})}{\sum_{R \in R^*} \sum_{n\text{-gram} \in R} count(n\text{-gram})} \quad (12)$$

ROUGE measures the geometric average of the $n$-gram recalls, where n denotes the length of the $n$-gram, and the recall $R_n$ is calculated as the number of matched $n$-grams divided by the number of candidate $n$-grams in the reference answer.

$$\begin{aligned} \text{METEOR} &= F_{\text{mean}} \times (1-p) \\ &= \frac{10 \times R \times P}{R + 9 \times P} \times \left(1 - 0.5 \times \left(\frac{chunk}{unigrams}\right)^3\right). \end{aligned} \quad (13)$$

METEOR measures the harmonic mean $F_{\text{mean}}$ between the precision $P$ and recall $R$ using a penalty $p$, as shown in (13). Unlike BLEU, METEOR accepts synonyms based on WordNet as matched words.

$$\text{CIDEr}_n(c_i, r_i) = \frac{1}{m} \sum_j \frac{\mathbf{g}^n(c_i) \times \mathbf{g}^n(r_{ij})}{\| \mathbf{g}^n(c_i) \| \times \| \mathbf{g}^n(r_{ij}) \|}. \quad (14)$$

CIDEr measures the average cosine similarity between the candidate and reference answers, based on the TF-IDF weighted $n$-grams. In (14), $\mathbf{g}^n(c_i)$ is the $n$-gram vector of the candidate answer for the $i$-th image, and $\| \mathbf{g}^n(c_i) \|$ is the norm of the vector. Likewise, $\mathbf{g}^n(r_{ij})$ is the $n$-gram vector of the $j$-th sentence in the reference answer for the $i$-th image, and $\| \mathbf{g}^n(r_{ij}) \|$ is the norm of the vector.

## 4.3 | Implementation

For the purpose of carefully examining the performance per module in the proposed image classification and captioning model, we implement four models: a CNN-based image classifier [42], an LRCN-based image captioner [1], the proposed model with an intermediate layer, and the proposed model with a CAM-based disagreement loss. First, the CNN-based image classifier consists of both a CNN-based image encoder and a single-layer perceptron-based image classifier, without an intermediate layer. Inception v.4 [42] is employed as the CNN-based image encoder, the image classifier consists of a 1024-dimensional single-layer perceptron, and an image is represented by 1024-dimensional image features. The image classifier is updated as shown in lines 11 and 15 of Algorithm 1.

Second, the LRCN-based image captioner [1] consists of both a CNN-based image encoder and an LSTM-based image captioner, without an intermediate layer. The LSTM-based image captioner is learned using 1024 hidden units. The image captioner is learned as described in lines 12 and 16 of Algorithm 1.

Third, the proposed model with an intermediate layer consists of a CNN-based image encoder, an intermediate layer, an image classifier, and an LSTM-based image captioner. The intermediate layer and image classifier is a 1024-dimensional single-layer perception. This model is updated as shown lines 11–17 of Algorithm 1. Unlike the proposed model with the CAM-based disagreement loss, this model replaces the equation in line 14 of algorithm 1 with $L_{\text{medi}} \Leftarrow L_{\text{class}} + L_{\text{caption}}$, without $L'_{\text{medi}}$ representing the CAM-based disagreement loss.

Finally, the proposed model with the CAM-based disagreement loss consists of a CNN-based image encoder, an intermediate layer, an image classifier, and an LSTM-based image captioner. This model is updated as described in lines 11–17 of Algorithm 1. Unlike the proposed model with an intermediate layer, this model utilizes the equation $L_{\text{medi}} \Leftarrow L_{\text{class}} + L_{\text{caption}} + L'_{\text{medi}}$ in line 14 of Algorithm 1.

These models are learned based on PyTorch [43], and the model hyperparameters are determined based on the validation set. Stochastic gradient descent is adopted with the base learning rate of 0.1, and the size of each mini-batch in training consists of 128 images.

## 4.4 | Quantitative experimental results

Table 2 shows the performances of the models on the Flowers and Birds datasets, as measured according to the evaluation metrics. The models comprise the CNN-based image classifier (CNN), LRCN-based image captioner (LRCN), proposed model with an intermediate layer, and proposed model with the CAM-based disagreement loss. These are measured in terms of BLEU, METEOR, ROUGE, and CIDEr.

For the image classification task, the baseline performance of the CNN-based image classifier is already very high, and thus it can be claimed to be significant that the proposed model with the CAM-based disagreement loss improves the performance by 1.53% for the top-1 accuracy on the Birds dataset. The best top-1 accuracy for the Flowers dataset (99.38%) is higher than the best top-1 accuracy for the Birds dataset (87.12%), because the Flowers dataset contains 102 classes, whereas the Birds dataset has 200 classes. In addition, the Flowers dataset consists of relatively easy to distinguish classes. Because the performance differences are within 0.8%, there are insignificant performance differences between the proposed model with the intermediate layer and that with the CAM-based disagreement loss. For the top-1 for the Flowers dataset, the performance of the proposed model with the intermediate layer is 0.07% higher than that of the proposed model with the CAM-based disagreement loss. For the top-1 for the Birds dataset, the performance of the proposed model with the CAM-based disagreement loss is 0.88% higher than that with the intermediate layer.

In the image captioning task, the proposed model with the intermediate layer yields a better performance on all evaluation metrics compared with the LRCN-based image captioner. Specifically, it achieves a 12.43% improvement in the CIDEr for the Birds dataset, and 3.33% for the Flowers dataset. This shows that the proposed model with the intermediate layer can effectively utilize important words, because the CIDEr metric measures the similarity between the TF-IDF weighted n-grams. Ultimately, the intermediate layer affects the image captioning task, although this model does not consider the CAM-based disagreement loss.

Compared with the proposed model with the intermediate layer, that with the CAM-based disagreement loss yields a better performance on all evaluation metrics. It achieves clear improvements in the CIDEr of 8.31% for the Birds dataset and 3.23% for the Flowers dataset. In addition, it achieves 20.74% and 6.56% improvements for Birds and Flowers datasets, respectively, compared with the LRCN-based image captioner. These results show that the proposed model with the CAM-based disagreement loss is highly effective for the image captioning task, because it generates more appropriate caption sentences for each image as the disagreement loss modifies the erroneous judgment criteria.

**TABLE 2** Image classification and captioning performance on the test set

| Metric | Top-1 (%) | Top-5 (%) | BLUE1 (%) | BLUE2 (%) | BLEU3 (%) | BLEU4 (%) | METEOR (%) | ROUGE (%) | CIDEr (%) |
|---|---|---|---|---|---|---|---|---|---|
| Flower | | | | | | | | | |
| CNN [42] | 98.62 | **99.93** | - | - | - | - | - | - | - |
| LRCN[1] | - | - | 88.32 | 81.51 | 73.74 | 67.95 | 39.45 | 76.74 | 42.47 |
| Proposed with intermediate layer | **99.38** | 99.85 | 89.51 | 83.05 | 75.53 | 69.89 | 40.27 | 77.16 | 45.80 |
| Proposed with CAM-based disagreement loss | 99.31 | 99.86 | **90.21** | **83.56** | **76.06** | **70.48** | **41.37** | **78.54** | **49.03** |
| Bird | | | | | | | | | |
| CNN [42] | 85.59 | 95.58 | - | - | - | - | - | - | - |
| LRCN[1] | - | - | 88.50 | 75.92 | 64.34 | 53.69 | 32.56 | 66.07 | 40.21 |
| Proposed with intermediate layer | 86.24 | 96.18 | 89.41 | 79.90 | 70.31 | 61.02 | 36.23 | 71.65 | 52.64 |
| Proposed with CAM-based disagreement loss | **87.12** | **96.89** | **90.97** | **81.87** | **72.71** | **63.75** | **37.61** | **73.06** | **60.95** |

Boldface indicates best score among the methods.

In summary, by embedding both the latent image and text representations at the same time, the multimodal learning model with the intermediate layer consistently outperformed the baseline model. In addition, the disagreement loss also yields a consistent improvement in the performance, because the similar criteria help to improve the performance for related tasks such as image classification and image captioning. Specifically, the disagreement loss method with CAM and the cosine distance provides a similar useful criteria. In addition, the proposed model improved the performance on the image captioning task more compared with the image classification task, because the baseline performance is already very high on the image classification task. This indicates that the higher performance on the image classifier led to an improvement in the performance on the image captioner.

## 4.5 | Qualitative experimental results

To carefully analyze the causes of the performance differences among the models, Figures 3 and 4 present heat-map visualizations generated by CAM based on the following models: the CNN-based image classifier, LRCN-based image captioner, proposed classifier with the CAM-based disagreement loss, and proposed captioner with the CAM-based disagreement loss. Figures 3 and 4 distinguish between the proposed classifier and captioner, because each module yields an individual answer. The heat maps describe the areas in the given image focused upon by each model when yielding an answer. For both the bird and flower images, the proposed model focuses on more appropriate points compared with the other models.

In the first image in Figure 3, the CNN-based image classifier focuses on the abdomen of the bird, whereas the proposed classifier focuses on its yellow shoulders with the aid of the image captioner. Intuitively, the yellow shoulders are more discriminative features of the bird in the image. In the second image, the CNN-based image classifier focuses on the wing of the bird, whereas the proposed classifier focuses on both the black stripe on the throat and the wing. In the third image, the CNN-based image classifier focuses on the feet of the bird, whereas the proposed classifier focuses on the body. For the flower images, the proposed classifier focuses on the pistil, stamen, and petals, whereas the CNN-based image classifier does not focus on the petals at all. Because the proposed model includes an intermediate layer, considering the synergy between the image classifier and image captioner, the proposed classifier can focus on useful components for the image classification task.

In Figure 4, we qualitatively compare the results of the proposed and baseline captioners. In the first image of Figure 4, compared with the reference caption written by a human, the proposed captioner generates a caption describing the red crown and black wings of the bird, whereas the LRCN-based image captioner misses the "red crown" phrase

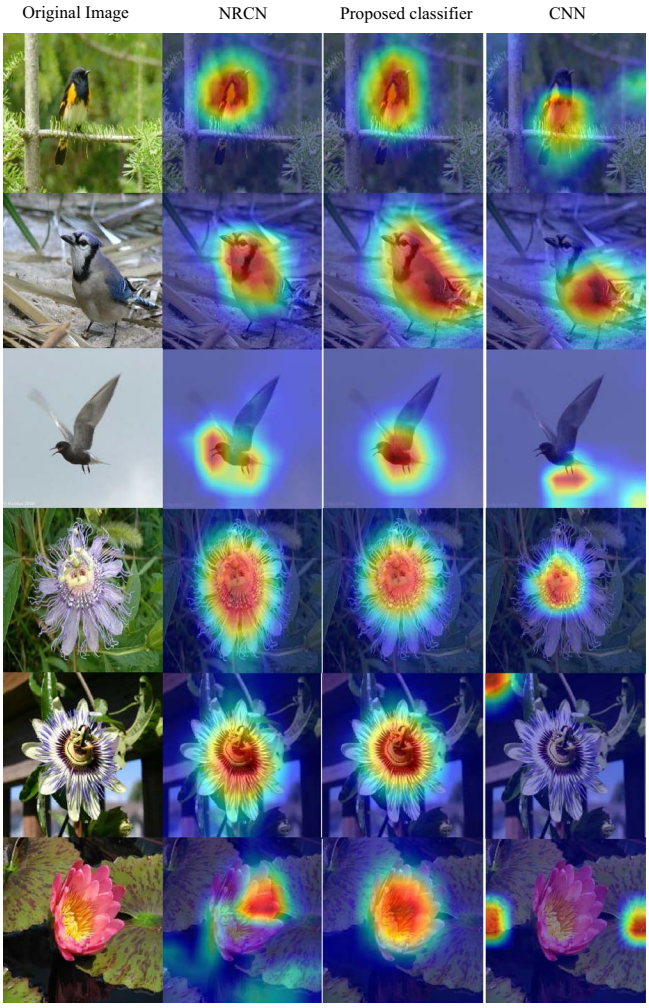


**FIGURE 3** CAM visualizations for the image classification task

because it focuses on the beak and head of the bird rather than the red crown. In the third image, the proposed captioner generates a caption describing the yellow style and red petals of the flower, whereas the LRCN-based image captioner cannot represent these significant words because it focuses on the component of the wrinkled petal rather than the yellow style. For the second bird and fourth flower images, the LRCN-based image captioner focuses on highly inappropriate components, whereas the proposed captioner focuses on appropriate components by utilizing the intermediate layer, which is influenced by both the image classifier and image captioner.

Ultimately, the proposed model can successfully perform both image classification and captioning tasks simultaneously, because it utilizes an intermediate layer that is influenced by feedback from both the image classification and image captioning tasks. The proposed classifier achieves a superior classification performance to the CNN-based classifier because it can be aided by the proposed image captioner through the intermediate layer. Compared with the LRCN-based captioner, the proposed captioner generates an image

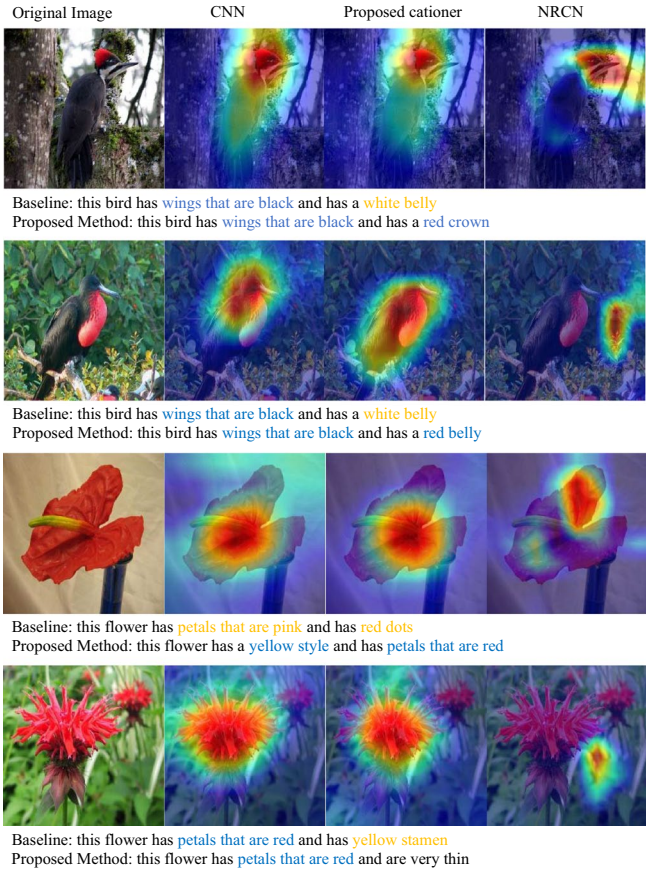


**FIGURE 4** Automatically generated text captions and CAM visualizations for the image captioning task. Yellow text indicates incorrectly generated text in the caption, whereas blue text indicates appropriately generated text

caption with fewer errors, because it focuses on the more distinctive components through the influence of the image classifier during training.

## 5 | CONCLUSION

In this study, we proposed an image classification and captioning model considering a CAM-based disagreement loss. The proposed model has the following characteristics.

First, it can learn the modules for different tasks concurrently and optimize the performance, because it utilizes an intermediate layer between the image classifier and image captioner. In the proposed model, the weight of the class discriminative neuron can be influenced by the text features, while that of the text descriptive neuron can be influenced by the image features. Experimental results demonstrate that the proposed model with the intermediate layer achieves 12.43% and 3.33% improvements in the CIDEr on the Birds and Flowers datasets, respectively, compared with the previous LRCN-based image captioner. In addition, the proposed model with the intermediate layer achieves 0.76% and 0.65% improvements in the top-1 accuracy for the Flowers and Birds datasets compared with the previous CNN-based image classifier.

Second, the proposed model using the CAM-based disagreement loss can consider the synergy in various relations: those between the image and its class, the image and its caption, and the class and its caption. Experimental results show that the proposed model with the CAM-based disagreement loss achieves 20.74% and 6.56% improvements in the CIDEr on the Birds and Flowers datasets, respectively, compared with the previous LRCN-based image captioner. In addition, the proposed model with the CAM-based disagreement loss achieves 1.53% and 0.69% improvements in the top-1 accuracy on the Birds and Flowers datasets compared with the previous CNN-based image classifier.

## ORCID

*Yeo Chan Yoon* https://orcid.org/0000-0002-5573-8964

## REFERENCES

1. J. Donahue et al., *Long-term recurrent convolutional networks for visual recognition and description*, in Proc. IEEE Conf. Comput. Vision Pattern Recogn., Boston, MA, USA, June 2015, pp. 2625–2634.
2. O. Vinyals et al., *Show and tell: A neural image caption generator*, in Proc. IEEE Conf. Comput. Vision Pattern Recogn., Boston, MA, USA, June 2015, pp. 3156–3164.
3. Y. Dong et al., *Improving interpretability of deep neural networks with semantic information*, arXiv preprint arXiv: 1703.04096 (2017), 3–19.
4. L.A. Hendricks et al., *Generating visual explanations*, in Eur. Conf. Comput. Vision, Amsterdam, The Netherlands, Oct. 2016, pp. 3–19.
5. L.A. Hendricks et al., *Deep compositional captioning: Describing novel object categories without paired training data*, in Proc. IEEE Conf. Comput. Vision Pattern Recogn., Las Vegas, NV, USA, June 2016, pp. 1–10.
6. Q. You et al., *Image captioning with semantic attention*, in Proc. IEEE Conf. Comput. Vision Pattern Recogn., Las Vegas, NV, USA, June 2016, pp. 4651–4659.
7. S.J. Rennie et al., *Self-critical sequence training for image captioning*, in IEEE Conf. Comput. Vision Pattern Recogn., Honolulu, HI, USA, July 2017, pp. 1179–1195.
8. W. Qi et al., *Image captioning and visual question answering based on attributes and external knowledge*, IEEE Trans. Pattern Anal. Mach. Intell. **40** (2018), no. 6, 1367–1381.
9. Y. Youngjae et al., *End-to-end concept word detection for video captioning, retrieval, and question answering* in IEEE Conf. Comput. Vision Pattern Recogn., Honolulu, HI, USA, July 2017, pp. 3261–3269.
10. P. Anderson et al., *Bottom-up and top-down attention for image captioning and VQA*, arXiv preprint arXiv: 1707.07998, 2017.
11. L. Jiasen et al., *Knowing when to look: Adaptive attention via a visual sentinel for image captioning*, in Proc. IEEE Conf. Comput. Vision Pattern Recogn., Honolulu, HI, USA, July 2017, pp. 3242–3250.

12. T. Yao et al., *Boosting image captioning with attributes*, in IEEE Int. Conf. Comput. Vision, Venice, Italy, Oct. 2017, pp. 22–29.
13. C. Wang, H. Yang, and C. Meinel, *Image captioning with deep bi-directional lstms and multi-task learning*, ACM Trans. Multimedia Comput., Commun., Applicat., **14** (2018), no. 2s, 1–20.
14. C. Szegedy et al., *Going deeper with convolutions*, in Proc. IEEE Conf. Computer Vision Pattern Recogn., Boston, MA, USA, June 2015, pp. 1–9.
15. S. Reed et al., *Learning deep representations of fine-grained visual descriptions*, in Proc. IEEE Conf. Comput. Vision Pattern Recogn., Las Vegas, NV, USA, June 2016, pp. 49–58.
16. L. Zhang et al., *Learning a deep embedding model for zero-shot learning*, in Proc. IEEE Conf. Comput. Vision Pattern Recogn., Honolulu, HI, USA, July 2017, pp. 3010–3019.
17. X. He and Y. Peng, *Fine-grained image classification via combining vision and language*, in Proc. IEEE Conf. Comput. Vision Pattern Recogn., Honolulu, HI, USA, July 2017, pp. 7332–7340.
18. R. Kiros, R. Salakhutdinov, and R.S. Zemel, *Unifying visual-semantic embeddings with multimodal neural language models*, arXiv preprint arXiv: abs/1411.2539, 2014.
19. J. Mao et al., *Learning like a child: Fast novel visual concept learning from sentence descriptions of images*, in Proc. IEEE Int. Conf. Comput. Vision, Santiago, Chile, 2015, pp. 2533–2541.
20. R. Vedantam et al., *Context-aware captions from context-agnostic supervision*, in Proc. IEEE, Conf. Comput. Vision Pattern Recogn., Honolulu, HI, USA, July 2017, pp. 1070–1079.
21. A.H. Abdulnabi et al., *Multi-task CNN model for attribute prediction*, IEEE Trans. Multimedia **17** (2015), no. 11, 1949–1959.
22. T.-H. Chen et al., *Show adapt and tell: Adversarial training of cross-domain image captioner*, in IEEE, Int. Conf. Comput. Vision, Venice, Italy, Oct. 2017, pp. 521–530.
23. R.R. Selvaraju et al., *Grad-CAM: Visual explanations from deep networks via gradient-based localization*, in IEEE Int. Conf. Comput. Vision, Venice, Italy, Oct. 2017, pp. 618–626.
24. Y.-C. Yoon et al., *Fine-grained mobile application clustering model using retrofitted document embedding*, ETRI J. **39** (2017), no. 4, 443–454.
25. S. Kong and C. Fowlkes, *Low-rank bilinear pooling for fine-grained classification*, in IEEE Comput. Vision Pattern Recogn., Honolulu, HI, USA, July 2017, pp. 7025–7034.
26. Y. Shaoyong et al., *A model for fine-grained vehicle classification based on deep learning*, Neurocomput. **257** (2017), 97–103.
27. X.-S. Wei et al., *Selective convolutional descriptor aggregation for fine-grained image retrieval*, IEEE Trans. Image Process. **26** (2017), no. 6, 2868–2881.
28. G.-S. Xie et al., *LG-CNN: from local parts to global discrimination for fine-grained recognition*, Pattern Recogn. **71** (2017), 118–131.
29. S.H. Lee, *HGO-CNN: Hybrid generic-organ convolutional neural network for multi-organ plant classification*, in IEEE Int. Conf. Image Process., Beijing, China, Sept. 2017, pp. 4462–4466.
30. A. Li et al., *Zero-shot fine-grained classification by deep feature learning with semantics*, arXiv preprint arXiv: abs/1707.00785, 2017.
31. Z. Akata et al., *Evaluation of output embeddings for fine-grained image classification*, in Proc. IEEE Conf. Comput. Vision Pattern Recogn., Boston, MA, USA, June 2015, pp. 2927–2936.
32. R. Ranjan, V. M. Patel, and R. Chellappa, *Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition*, IEEE Trans. Pattern Anal. Mach. Intell. **41** (2018), 121–135.
33. K. Hashimoto et al., *A joint many-task model: Growing a neural network for multiple NLP tasks*, arXiv preprint arXiv: abs/1611.01587, 2016.
34. R. Caruana, *Multitask learning: a knowledge-based source of inductive bias*, in Proc. Int. Conf. Mach. Learn., Amherst, MA, USA, June 1993, pp. 41 – 48.
35. L. Duong et al., *Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser*, in Proc. Annu. Meeting Association Computat. Linguistics Int. Joint Conf. Natural Language Process., Beijing, China, July 2015, pp. 845–850.
36. M. Nilsback and A. Zisserman, *Automated flower classification over a large number of classes*, in Proc. Indian Conf. Comput. Vision, Graphics Image Process., Bhubaneswar, India, Dec. 2008, pp. 722–729.
37. C. Wah et al., *The Caltech-UCSD Birds-200-2011 Dataset, Tech. Report CNS-TR-2011-001, California Institute of Technology*, 2011.
38. K. Papineni et al., *Bleu: A method for automatic evaluation of machine translation*, in Proc. Annu. Meeting Association Computat. Linguistics, Philadelphia, PA, USA, July 2002, pp. 311–318.
39. C.-Y. Lin, *Rouge: a package for automatic evaluation of summaries*, in Workshop Text Summarization Branches Out, Post-Conf. Workshop ACL, Barcelona, Spain, July 2004, pp. 74–81.
40. S. Banerjee and A. Lavie, *Meteor: an automatic metric for MT evaluation with improved correlation with human judgments*, in Proc. ACL Workshop Intrinsic Extrinsic Evaluation Measures Mach. Translation Summarization, Ann Arbor, MI, USA, 2005, pp. 65–72.
41. R. Lawrence, C.L. Zitnick, and D. Parikh, *Cider: Consensus-based image description evaluation*, arXiv preprint arXiv: abs/1411.5726 (2014).
42. C. Szegedy, S. Ioffe, and V. Vanhoucke, *Inception-v4, Inception-Resnet and the impact of residual connections on learning*, in Proc. AAAI Conf. Artif. Intell., San Francisco, CA, USA, Feb. 2017, pp. 2478-4284.
43. A. Paszke et al., *Automatic differentiation in PyTorch*, in Proc. NIPS, Long Beach, CA, USA, 2017.

## AUTHOR BIOGRAPHIES

**Yeo Chan Yoon** received BS and MS degrees in Computer Science and Engineering from Korea University, Seoul, Rep. of Korea, in 2004 and 2007, respectively. Currently, he is a senior researcher at the Electronics and Technology Research Institute (ETRI), Daejeon, Rep. of Korea, and also is a PhD student of Korea University, Seoul, Rep. of Korea. His research interests include digital content recommendation systems, natural language processing, machine learning, and big data analytics.

**So Young Park** received a BS degree in Computer Science and Engineering from Sangmyung University, Chungnam, Rep. of Korea, in 1997, and MS and PhD degrees in Computer Science from Korea University, Seoul, Rep. of Korea in 1999 and 2005, respectively. She has been a professor in the Department of Game Design and Development at Sangmyung University since 2007. Her current research interests include natural language understanding and data mining.

**Soo Myoung Park** received a BS degree in Computer Science Engineering from the University of Dankook, in 1990, and MS and PhD degrees in computer engineering from the University of Konkuk, in 1992 and 1999, respectively. Currently, he is the director of the Smart Content Research Section at the Electronics and Technology Research Institute (ETRI), Daejeon, Rep. of Korea.

**Heuiseok Lim** received BS, MS, and PhD degrees in Computer Science and Engineering from Korea University, Seoul, Rep. of Korea, in 1992, 1994, and 1997 respectively. Currently, he is a professor in the Department of Computer Science and Engineering at Korea University. His research interests include Cognitive-Neuro Language Processing in Human Brain, Computer Science Education, and Information Retrieval.