IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 13, 2020, accepted March 27, 2020, date of publication April 6, 2020, date of current version April 21, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2986055

Poster-Based Multiple Movie Genre Classification Using Inter-Channel Features

JEONG A. WI, (Student Member, IEEE), SOOJIN JANG, (Student Member, IEEE), AND YOUNGBIN KIM, (Member, IEEE)

Department of Image Science and Arts, Chung-Ang University, Seoul 06974, South Korea

Corresponding author: Youngbin Kim (ybkim85@cau.ac.kr)

This work was supported in part by the National Research Foundation of Korea (NRF) through the Korean Government (MSIT) under Grant NRF-2018R1C1B5046461.

ABSTRACT As the scale of the film industry grows, the demand for well-established movie databases is also growing. The genre of a movie supplies information on its overall content and has multiple values. Therefore, it should be well classified utilizing the characteristics of movies, without omissions in the database. In this study, we extract the optimal information and characteristics from movie posters to aid the classification of movies into genres and propose the use of a Gram layer in a convolutional neural network (CNN). The Gram layer first extracts style features by applying the Gram matrix to produce a feature map of a poster image. Using this as a style weight, the existing feature map is merged with style information to perform the genre classification task. The proposed Gram layer performed multi-genre classification tasks with higher efficiency than a residual neural network (ResNet), which is the current CNN model used for such tasks. We compared the activation map with the Squeeze-and-Excitation network, which gives weight to the image, and we confirmed that the introduction of the Gram layer actually focuses on the style of the movie poster. To classify the movie genres, we reconstructed the poster dataset into 12 multi-genres that emphasized the characteristics of each poster.

INDEX TERMS Classification, dataset, deep learning, gram matrix, multi-label classification, movie genre classification, movie poster.

I. INTRODUCTION

The scale of the global film industry is significantly growing every year, and the methods which it distributes content have undergone substantial changes. In the past, theater-oriented businesses have been the main focus, but the growth of online streaming services such as Netflix[1] and Amazon Prime[2] have markedly changed how audiences consume movie content. As a result, the content platform business has increased and multiple genre-oriented databases have become an essential element. Domestic and international movie databases, such as the Internet Movie Database (IMDB),[3] provide a variety of information related to movies. Such databases utilize the classification systems that divide movies into several dozens of genres.

The associate editor coordinating the review of this manuscript and approving it for publication was Liang-Bi Chen.

[1]https://www.netflix.com

[2]https://www.amazon.com/prime

[3]https://www.imdb.com

The genre is one of the most important characteristics of a movie; by itself, it implies the overall contents of the movie. Above all, the genre has a substantial influence on the choice of movies. Therefore, because the genre is the primary characteristic of a movie that can provide general information about the movie, it should be classified without omission or error and used in automated services such as well-established databases.

Movie genre classification may be very comprehensive or diverse based on the criteria; moreover, there are many genres which are similar, and one movie may belong to several of them, making accurate classification difficult. Although several databases use a combination of genre personality and movie characteristics to achieve genre classification, this method may create ambiguity regarding genres and inconsistency in the total number of genres. This method is also time-consuming and is exposed to the risk of subjective judgments.

To overcome these problems and perform genre classification efficiently, many previous studies utilized machine

learning and deep learning to attempt automatic genre classification, based on various data, such as movie posters [1], [2], plots [3], [4] and trailers [5], [6], which were used separately or in combination. However, the use of movie plots is limited by the fact that they express only the introductory portion of the main plot and not the entire content. Trailers contain various types of information, such as image frames and audio; however, trailers require high computational capacity, owing to their large data size. In contrast, a movie poster is a single image that may be easily processed; moreover, most movie posters are designed to have a similar size and ratio, thereby facilitating the creation of a consistent dataset.

Additionally, a movie poster is not simply an image but a well-planned delivery and promotion medium, designed to convey the overall content of the movie. It has the advantage of containing rich visual features about the movie, using various elements such as background, characters, and typography. For these reasons, many previous studies have attempted genre classification using posters. However, most of these studies used low-level feature extraction [7], [8] that relied on color histograms and textures. This limited the total number of genres to a minimal level [9] without considering the nature of the genre, and resulted in a single genre classification rather than multiple label classification [10].

In this study, we propose the use of the Gram layer with a Gram matrix for multi-label genre classification by poster. Gram matrices have been used in the deep learning field to extract information about the style of an image for tasks such as style transfer. The Gram layer extracts characteristics by applying the Gram matrix to the feature map and using it as a weight to focus on the characteristics of each genre contained in the poster. The Gram layer can be easily added to existing convolutional neural network (CNN)-based models, and the focus on the style of the poster allows this model to better perform the task of multiple movie genre classification, compared to existing models.

The movie poster data was crawled in the order of box office proceeds from the IMDB database and 12 genre types were selected. We then reconstructed the movie poster dataset with up to nine genres for multi-label genre classification.

Overall, this study suggests a novel approach for multiple movie genre classification with inter-channel features. This study makes the following contributions:

- We propose a Gram layer that captures the correlation between posters and genres: The Gram layer can be used for the performance of multiple movie genre classification models.
- We constructed a dataset for poster-based multi-genre classification: By identifying the characteristics of posters and genres, the number of genres and number of genre types were selected.
- Through feature map visualization, we were able to identify the characteristics of each genre of the movie posters and confirm that the Gram layer we proposed is well focused on the features of each genre.

The remainder of this paper is organized as follows. In Section 2, we present an overview of the related works. In Section 3, we propose our methods. In Section 4, we explain the experimental results. Finally, in Section 5, we conclude our work and discuss future avenues of research.

II. RELATED WORK

Many previous studies have attempted to classify movie genres using machine learning and deep learning. In this section, we review the studies that have used movie posters for genre classification, and those that have not. We also review studies that used Gram matrix for classification in conjunction with a visual attention mechanism.

A. POSTER-BASED MOVIE GENRE CLASSIFICATION

A poster expresses a variety of information about a movie via an image; therefore, many aspects related to the movie can be extracted from the poster.

Low-level visual characteristics, such as the colors and edges of the poster, were extracted using a color histogram and a GIST image descriptor [7], [8]. Genre classification was performed using conventional machine learning techniques [11], [12]. Multi-label k-nearest neighbor and Random K-labelsets algorithms were used for the classification of multiple label data; single label data were also classified using the Naive Bayes classifier.

In a similar study [8], characteristics such as colors, edges, and textures were extracted from the poster. Apart from the use of low-level features such as colors, edges, and textures, the number of human faces in the poster was used as an additional characteristic for classification. Additionally, the text characteristics of the synopsis were extracted using the vector space model [9]. The genre classification was performed using a support vector machine (SVM) based on the characteristics of the poster and the synopsis; however, the number of genres was limited to four, and multiple label classification was not performed. Most of the above studies attempted to classify genres with only a very small number of posters.

With the use of deep learning in genre classification, the number of posters used gradually increased [10], [13], [14]. However, many studies used both conventional machine learning techniques and statistical multi-label methods, rather than using only deep learning models [11]. Chu and Guo [1] performed genre classification beyond the existing machine learning techniques by applying a convolution network to the poster images. They also used additional characteristics for genre classification through object detection from poster images. A total of 23 genres were defined, more than those in previous studies, for poster genre classification. Moreover, this technique demonstrated better performance than the existing machine learning techniques. Since then, there have been studies using only deep learning, such as the one that applied transfer learning [10].

B. NON-POSTER-BASED MOVIE GENRE CLASSIFICATION

In addition to posters, other information, such as trailers and synopses that contain information about the movie, were used to identify the genre. In particular, movie trailers contain image frames and audio information, thereby providing various characteristics about a movie. Accordingly, many recent studies have used trailers for genre classification.

Similar to genre classification using posters, the use of low-level features for classification in movie trailers has also been studied [11], [15]. In addition, genre classification based on trailer information has been performed using convolutional neural networks [16]. A trailer dataset for genre classification was created; however, the number of genres was limited to four, and multiple label classification was not addressed. In addition, a CNN was built for motion recognition, including data augmentation and histograms at the scene level.

Similarly, Wehrmann and Barros [6] attempted genre classification by applying a convolutional network to trailers. In addition, their study used residual connections and built a convolution-through-time network for multi-label movie genre classification. The constitution of an additional network using audio information was also attempted. A total of nine genres were defined, and the Labeled Movie Trailer Dataset (LMTD) was used to create a new dataset, LMTD-9.

With the development of natural language processing research, genre classification using plots or synopses of movies is being actively conducted [4], [17]–[20]. Ertugrul and Karagoz [21] performed genre classification based on movie plot summaries using bidirectional long short-term memory. Only single label classification was used, and the number of genres was limited to four. Another study proposed a self-attention network based on synopses [3] for multiple label classification. This study used the LMTD-9 dataset and defined a total of nine genres.

There are many examples of genre classification using various types of data related to movies [13], [14], [22], [23], and there have been studies on classifying genres by combining them with posters [9]. However, this paper focuses only on poster data. Posters are easy to handle because they contain various types of information in one image, and they are also one of the main data sources available for most movies.

C. MACHINE LEARNING AND DEEP LEARNING USING A GRAM MATRIX AND VISUAL ATTENTION

The Gram matrix has been frequently addressed in the machine learning and deep learning fields. Some studies have calculated style representations using the Gram matrix [24], [25]. Using this technique, the style losses between the target (original) and generated images during the style transfer were calculated, such that the styles and textures were similar between the target and generated images. The Gram matrix was extracted as a feature for images—such as famous paintings with particular styles—and used for classification [26], [27]. The Gram matrix can also be used for medical images as well as images with specific styles [28].

The attention mechanism was first proposed in natural language processing [29] and is now widely used in various deep learning fields such as natural language processing [30], [31], image processing [32], [33], and recommendation systems [34]. It is known that the performance of models can be improved by inputting the content that should be emphasized to the deep learning model. Among the various attention models in the field of image processing, the Squeeze-and-Excitation Network (SENet) [35] is particularly well-known. SENet won the 2017 ImageNet competition, but unlike the previous CNN models, it did not develop a new architecture— it proposed an SE-block that could be applied to any CNN model. In other words, by only adding an SE-block, it can improve the performance of various models such as VGGNet [36], GoogLeNet [37], ResNet [38]. After the image feature map is squeezed, the channel is recalibrated by the excitation step. Since the introduction of SENet, many different image attention blocks similar to SENet have been studied [39].

However, these studies focus on the general classification of single label images. Inspired by SE-blocks that can be easily added, we have designed a Gram layer specifically for the multiple movie genre classification task. In this study, emphasis was focused on input tensors upon extraction of style features based on the Gram matrix. This allowed for a more precise genre classification owing to the focus on the correlation between style features among the posters of each genre. In the following section, we describe the method in detail.

III. METHOD

In this section, we first explore the dataset reconstructed for the multi-label genre classification task, and then explore the internal computational process of the Gram layer in detail.

A. RECONSTRUCTION OF MOVIE POSTER DATASET

A movie involves many data types, including posters, plots, and star ratings, as well as the image itself; therefore, movie datasets became one of the most important datasets extracted from existing databases for the purpose of machine learning and deep learning. Movie posters are also available from the MovieLens [40] and Kaggle movie datasets, among several others.

The direct usage of existing movie datasets for poster-based genre classification poses two challenges. First, existing datasets contain many movies from the 1800s. If a movie poster from this era is defined as ''an image intended for display in a movie theater during screening,'' instead of ''an image representing a movie,'' the movie poster should be excluded. The movie posters of this era in actual datasets were frequently missing images or consisted of simple movie stills. Because the composition of movie posters by genre was established in the late 20th century, the accuracy of poster-based genre classification is likely to be improved when movies from the late 20th century and early 2000s occupy a larger proportion of the dataset.

Second, most existing datasets define more than 20 genres. Although a more detailed genre is helpful in expressing precise information about a movie, the problem of obtaining an unbalanced ratio between movie genres arises. The imbalance is exaggerated by the fact that a movie may be classified into up to three genres.

Therefore, this study reconstructed the movie dataset for poster-based genre classification by solving these problems. First, to identify consistent poster features related to genre, poster images were crawled in the order of box office proceeds, by genre, in IMDB; movies with fewer than 50 ratings were excluded. Second, 12 subjective genres were defined, based on the definition of genres and guidelines in IMDB. The 12 genres were Action, Adventure, Animation, Comedy, Crime, Drama, Fantasy, Horror, Mystery, Romance, Sci-Fi, and Thriller. The maximum number of genres assigned to a movie was set as nine, based on the genres listed on the movie detail page in IMDB. A total of 20,764 movie posters were collected.

Figs. 1 and 2 show certain statistics on the poster dataset that was collected. Fig. 3 shows a co-occurrence matrix generated for each poster genre label. Fig. 1 indicates that the Drama genre contained the largest proportion of posters, which may be because it is a genre that spans a wide range.

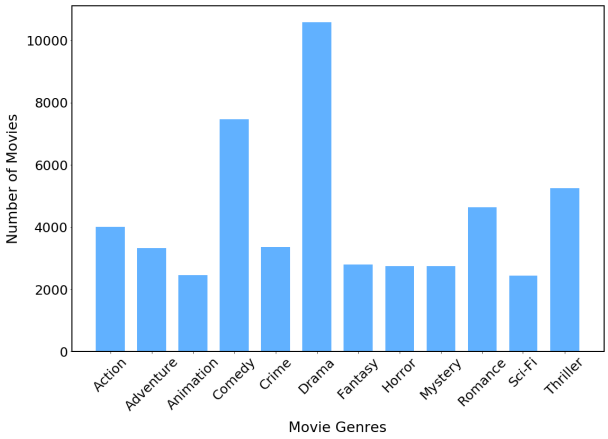

FIGURE 1. Number of movies by genre in our dataset.

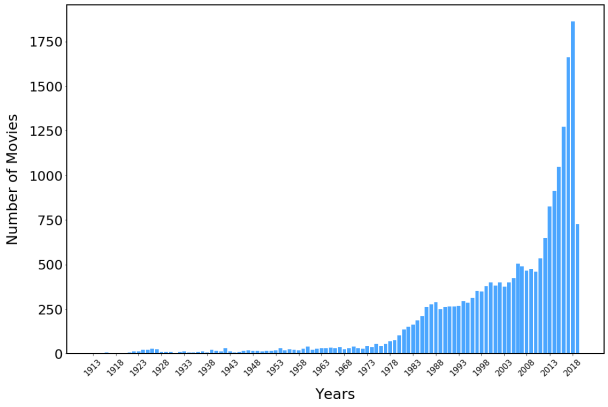

FIGURE 2. Number of movies by year in our dataset.

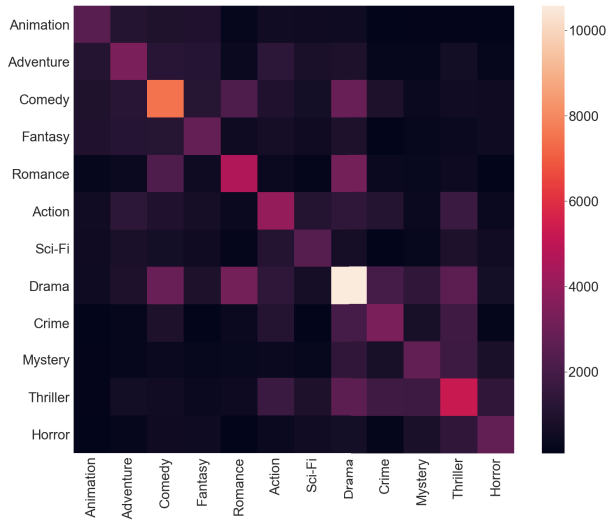

FIGURE 3. Co-occurrence matrix on genre label of our dataset.

Because only 12 genres were defined, the rare genres were excluded, alleviating the data imbalance. Fig. 2 shows that the posters in the newly reconstructed poster dataset used in this study were from movies released in the late 20th century and early 2000s.

B. GRAM LAYER FOCUSING ON THE STYLE OF MOVIE POSTERS

1) GRAM-STYLE FEATURE

The Gram layer is a single layer, broadly divided into two steps, which can be easily added to the bottleneck block and basic block of VGGNet [36] or ResNet [38]. Fig. 4 illustrates the overall internal structure of the Gram layer. The operations inside the Gram layer are two-fold. The first step is to create a gram matrix with an input feature map X^l to extract the style features from the input feature maps X^l. $X^l \in \mathbb{R}^{B\times C\times H\times W}$ is a feature map that is calculated by l's convolution layers. B is the mini batch size, C is the number of channels, and H and W are the sizes of the input feature map. The definition of the gram matrix is as follows:

$$G^l_{ij} = \sum_k F^l_{ik} F^l_{jk}. \tag{1}$$

Here, $F^l \in \mathbb{R}^{B\times C\times HW}$ represents the reshaping of X^l by flatting H and W, based on C. F^l_{ik} is the k-th output of the i-th filter. F^l_{jk} is the k-th output of the j-th filter and transpose. $G^l \in \mathbb{R}^{B\times C\times C}$ is an inner product between flattened feature maps F^l and contains information about the correlation among the outputs of different filters, thus summarizing the inter-channel relations associated with the interpretation of this image via various filters of the convolutional layer. Additionally, the flattened features were scalar-produced, leaving only the feature of style without information on space.

2) GRAM-WEIGHT AND WEIGHTED INPUT

The second operation in the Gram layer is to extract the style weight, used in multiplying the input tensor and weights in the

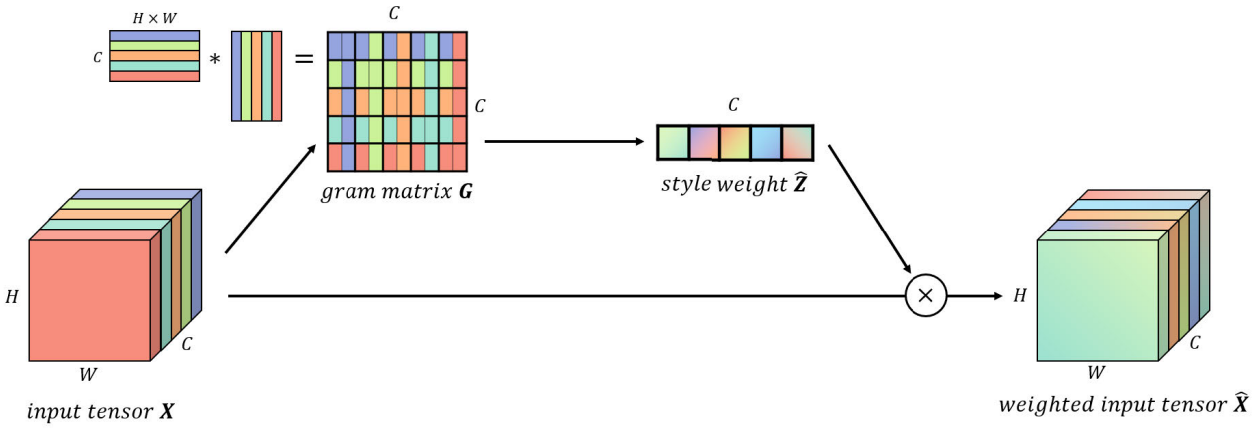

FIGURE 4. Structure of Gram layer: The feature map, the output of the convolution layer, is used as the input to the Gram layer. The Gram layer uses the input feature map to create a gram matrix and uses it to create a style weight. The style weight is multiplied by the input feature map so that the weighted input tensor becomes the output of the Gram layer.

style feature. The style feature G^l, which is extracted from X^l, through the fully connected weight layer $w^l \in \mathbb{R}^{C\times C}$, acquires the encoded style feature $Z^l \in \mathbb{R}^{B\times C\times C}$ by newly encoded features. w^l initially contains random values and trainable parameter values. Subsequently, $\bar{Z}^l \in \mathbb{R}^{B\times C\times 1}$ is acquired through the reduced mean. Next, through batch normalization and the sigmoid function, the style weight $\hat{Z}^l \in \mathbb{R}^{B\times C\times 1\times 1}$ is finally obtained.

This style weight is multiplied by the input tensor X^l to acquire a new weighted input tensor $\hat{X}^l \in \mathbb{R}^{B\times C\times H\times W}$. This process is represented in the following equations; the notation for the layer is omitted for convenience:

$$Z_{ij} = f_{fc}(G_{ij}) = w_{ij} \cdot G_{ij}, \tag{2}$$

$$\bar{Z}_{i1} = f_{rm}(Z_{ij}) = \frac{1}{C}\sum_{k=1}^{C} Z_{ik}, \tag{3}$$

$$\hat{Z} = \sigma(f_{bn}(\bar{Z})), \tag{4}$$

$$\hat{X} = \hat{Z} \cdot X. \tag{5}$$

where f_{rm} is a function that calculates a reduced mean of Z, and f_{bn} denotes a batch normalization function that calculates the mean and standard deviation of $\bar{Z}$ over the mini batches. $\sigma(x) = \frac{1}{1+e^{-x}}$ is a sigmoid function and utilizes the weight to represent the probability values.

3) COMPLETE GRAM LAYER PROCESS

The operation of each step is described in more detail as follows. The style feature G is calculated using the inter-product of two identical feature maps in the previous step, and therefore, it is a symmetric matrix. The fully connected layer is used to extract important genre classification information from between the inter-channel features contained in G. The batch normalization layer also represents style [41]; thus, it was added to improve the information about style.

The Gram matrix-based style weight $\hat{Z}$ was multiplied by the input tensor X to produce a new weighted input tensor $\hat{X}$, which served as a recalibration [42], representing an attention mechanism. This increased the performance of the network via assigning of higher weights to more useful information. The method proposed in this paper improves the performance of the genre classification task in that more weights are assigned to features focused on a particular style, according to genre.

IV. EXPERIMENT

A. DATASET

The dataset used in the experiment comprised 20,764 movie posters from movies released between 1913 and 2019; additionally, 18,684 training images, 1040 validation images, and 1040 test images were used. The genre labels included between one and nine films.

B. EVALUATION METRICS

Because movie genre classification is a multi-label classification task, it can be evaluated via two methods as follows [43], [44]. The sample-based method (Sb) calculates a score by evaluating one sample(movie poster) and averages the overall data. The label-based method (Lb) evaluates each genre label and averages all the genre labels.

We evaluated the performance of the model using four metrics from the two methods: accuracy, precision, recall, and f1-score. However, as the number of extracted prediction genre labels was identical to the number of ground truth genre labels, only accuracy and precision were used in the sample-based method because precision, recall, and f1-score have the same values. Therefore, a total of six scores were used: $Accuracy_{Sb}$, $Precision_{Sb}$, $Accuracy_{Lb}$, $Precision_{Lb}$, $Recall_{Lb}$, $F1score_{Lb}$. The metrics were defined as follows:

$$Accuracy_{Sb}, A_{Sb} = \frac{1}{n}\sum_{i=1}^{n}\frac{|\mathbb{Y}_i \cap Y_i|}{|\mathbb{Y}_i \cup Y_i|} \tag{6}$$

$$Precision_{Sb}, P_{Sb} = \frac{1}{n}\sum_{i=1}^{n}\frac{|\mathbb{Y}_i \cap Y_i|}{|Y_i|} \tag{7}$$

$$Accuracy_{Lb}, A_{Lb} = \frac{1}{k}\sum_{l=1}^{k}\frac{\sum_{i=1}^{n}\left|\mathbb{Y}_i^l \cap Y_i^l\right|}{\sum_{i=1}^{n}\left|\mathbb{Y}_i^l \cup Y_i^l\right|} \tag{8}$$

$$Precision_{Lb}, P_{Lb} = \frac{1}{k}\sum_{l=1}^{k}\frac{\sum_{i=1}^{n}\left|\mathbb{Y}_i^l \cap Y_i^l\right|}{\sum_{i=1}^{n}\left|Y_i^l\right|} \tag{9}$$

$$Recall_{Lb}, R_{Lb} = \frac{1}{k}\sum_{l=1}^{k}\frac{\sum_{i=1}^{n}\left|\mathbb{Y}_i^l \cap Y_i^l\right|}{\sum_{i=1}^{n}\left|\mathbb{Y}_i^l\right|} \tag{10}$$

$$F1score_{Lb}, F1_{Lb} = \frac{1}{k}\sum_{l=1}^{k}\frac{2\sum_{i=1}^{n}\left|\mathbb{Y}_i^l \cap Y_i^l\right|}{\sum_{i=1}^{n}\left|\mathbb{Y}_i^l\right| + \sum_{i=1}^{n}\left|Y_i^l\right|} \tag{11}$$

where n is the number of samples, and k is the number of labels. $\mathbb{Y}$ and Y represent the number of ground truth genre labels and prediction genre labels, respectively. The number of extracted prediction genre labels Y was identical to the number of ground truth genre labels; thus, the multi-label was applied. The evaluation metrics were calculated for the entire test set.

C. EXPERIMENTAL DETAILS

In this study, based on ResNet [38], experiments were conducted by attaching the proposed Gram layer to the basic block and bottleneck block of ResNet. For the structure of ResNet, the depth was set to 18, 34, 50, and 152, and the batch size was set to 64. For each model, training was performed for 40 epochs to prevent overfitting. Also, because overfitting easily occurs in the baselines, we applied early stopping depending on the training accuracy. We reported the highest scores based on the accuracy of the sample-based method(A_{Sb}). The stochastic gradient descent [45] was used as an optimizer, and the learning rate was set to 0.01. The loss function was used by binary cross entropy. These settings were equally applied to all the models used in the experiment, and the internal structure of each model was configured as reported in the original papers [36]–[38], [46]–[48]. The number of prediction labels extracted was equal to the number of ground truth labels in the movie poster, to make the baseline models take the form of multi-label classification tasks.

D. RESULTS

1) MULTIPLE GENRE CLASSIFICATION

First, we established ResNet as the reference model. This is because it is made up of both a residual block and an identity block, which allows other layers to be easily added. Rows 1 through 4 of Table 1 show the results of tests using ResNet18, ResNet34, ResNet50, and ResNet152. The tests were performed using the proposed movie test set.

When using the Gram layer, the accuracy and precision of the sample-based method were approximately 1∼2% higher than the baseline for all ResNet model depths. However, the same was not true for label-based metrics. This is thought to be due to an imbalance in genre labels. However, because it is more important to accurately classify genres for each movie poster, the importance of sample-based metrics is higher.

Second, we investigated multiple genre classification with various CNN models by adding Gram layers. VGGNet, GoogLeNet, InceptionV3, MobileNetV2 and DenseNet were selected in order to test various structures of CNN models. The tests were conducted with the same test set as above, and the performance was evaluated using the same metrics. Rows 5 through 10 of Table 1 show the results.

This experiment shows that the Gram layer can be used in various CNN structure-based models, apart from ResNet. In most cases, the Gram layer was inserted after the convolution layer of the building blocks of each model. Most models used in the multiple genre classification experiment demonstrated better performance when the Gram layer was used. Because of this, we believe the performance improvement is not greater than other models when using the Gram layer.

However, because the Gram layer can be freely inserted throughout the model, the performance can be changed, depending on its location and the number of uses. Therefore, if the position and number of Gram layers are properly adjusted, better performance will be achieved by the multiple genre classification task, according to the characteristics of each model.

Third, we experimented with the existing movie genre classification model [1], which is a combination of CNNs similar to AlexNet and Yolo version 2 pre-trained with the MSCOCO (Microsoft COCO: Common Objects in Context) dataset. The model was implemented as described in the paper, and we experimented with adding a Gram layer. These results are shown in Table 2. Chu's model attempted to classify genres by adding additional features to the CNN model through object detection. However, the number of object classes detected in the poster is not large. On the other hand, using the Gram layer can improve the multi-genre classification performance when only the CNN models are used.

2) ATTENTION TO MOVIE POSTER AND ITS STYLE

The Gram layer can be considered as giving attention to the image by applying the style weight to the feature map of the poster image. We compared the Gram layer, which is a type of attention method, to the image attention technique used frequently in the past and examined whether it was effective in the multiple movie genre classification task.

Among the existing image attention techniques, the SE layer of the SENet [35] was selected. The SE layer is an attention module that involves a squeeze operation, wherein spatial information is extracted using global average pooling, and an excitation operation, wherein recalibration is performed through a fully connected layer and sigmoid.

Table 3 displays the results of comparing the SE layer to the Gram layer. Three ResNet structures, ResNet18, ResNet 34, and ResNet50, were used. In ResNet18 and ResNet34, the Gram layer showed better performance than the SE layer in the overall evaluation metrics, and in ResNet50, the label-based accuracy, recall, and f1-score showed better performance than the SE layer.

TABLE 1. Comparison of evaluation metrics from the original model (Baseline) and the original model with a Gram layer (Gram layer) for ResNet and other CNN-based models.

index	CNN Model	Baseline						Gram layer (proposed)					
		A_{Sb}	P_{Sb}	A_{Lb}	P_{Lb}	R_{Lb}	$F1_{Lb}$	A_{Sb}	P_{Sb}	A_{Lb}	P_{Lb}	R_{Lb}	$F1_{Lb}$
(1)	ResNet18 [38]	0.4413	0.5305	0.3036	0.5581	0.4393	0.4478	0.4433	0.5342	0.2940	0.5711	0.4295	0.4344
(2)	ResNet34 [38]	0.4368	0.5275	0.3114	0.5339	0.4544	0.4611	0.4532	0.5398	0.3167	0.5645	0.4512	0.4660
(3)	ResNet50 [38]	0.4308	0.5185	0.3093	0.5358	0.4477	0.4602	0.4366	0.5264	0.3144	0.5445	0.4543	0.4662
(4)	ResNet152 [38]	0.4249	0.5130	0.2888	0.5168	0.4242	0.4347	0.4452	0.5334	0.3062	0.5659	0.4382	0.4543
(5)	VGG-16 [36]	0.4252	0.5168	0.2790	0.5153	0.4106	0.4180	0.4298	0.5163	0.2993	0.5333	0.4369	0.4471
(6)	VGG-19 [36]	0.4218	0.5093	0.2879	0.5323	0.4249	0.4317	0.4282	0.5171	0.2988	0.5040	0.4416	0.4427
(7)	GoogLeNet [37]	0.4646	0.5522	0.3330	0.5803	0.4692	0.4866	**0.4688**	**0.5565**	**0.3375**	**0.5638**	**0.4722**	**0.4918**
(8)	InceptionV3 [46]	0.4261	0.5140	0.2863	0.5385	0.4168	0.4307	0.4281	0.5115	0.2783	0.5209	0.4139	0.4181
(9)	MobileNetV2 [47]	0.4397	0.5279	0.3165	0.5399	0.4609	0.4673	0.4443	0.5310	0.3150	0.5487	0.4578	0.4631
(10)	DenseNet [48]	0.4575	0.5479	0.3322	0.5460	0.4809	0.4866	0.4597	0.5488	0.3369	0.5745	0.4798	0.4928

TABLE 2. Comparison of evaluation metrics from the existing movie genre classification model and ResNet34 with a Gram layer.

Model	A_{Sb}	P_{Sb}	A_{Lb}	P_{Lb}	R_{Lb}	$F1_{Lb}$
Chu et al. [1]	0.3576	0.4540	0.2027	0.4053	0.3322	0.3131
ResNet34 + Gram layer	0.4532	0.5398	0.3167	0.5645	0.4512	0.4660

TABLE 3. Comparison of evaluation metrics of SE layer and Gram layer for different ResNet models.

Model	SE layer						Gram layer (proposed)					
	A_{Sb}	P_{Sb}	A_{Lb}	P_{Lb}	R_{Lb}	$F1_{Lb}$	A_{Sb}	P_{Sb}	A_{Lb}	P_{Lb}	R_{Lb}	$F1_{Lb}$
ResNet18	0.4368	0.5253	0.3225	0.5319	0.4675	0.4765	0.4433	0.5342	0.2940	0.5711	0.4295	0.4344
ResNet34	0.4359	0.5244	0.3094	0.5494	0.4483	0.4597	0.4532	0.5398	0.3167	0.5645	0.4512	0.4660
ResNet50	0.4393	0.5267	0.3126	0.5473	0.4509	0.4614	0.4366	0.5264	0.3144	0.5445	0.4543	0.4662

The characteristic style of the genre was found throughout the poster, not only in specific parts. Although movie titles are often expressed in bold or large letters to draw attention from the observer, the title is not likely to be a major characteristic of a movie genre because the fonts and layouts of titles vary for different movies. To identify which part of the poster is concentrated on by the Gram layer and SE layer, we visualized and compared the activation maps of poster images.

Fig. 5 displays the results of visualizing the activation map extracted from layer 4 of the ResNet18 model, by selecting images with titles in different positions. It was observed that the Gram layer proposed in this paper was activated throughout the posters, showing the style of each genre rather than the title of the movie. The focus was on the features that could be used to determine the genre of the movie poster; however, all SE layers focused on the title of the movie and were activated locally rather than throughout the poster.

The proposed Gram layer is considered to be better at genre classification compared to the attention mechanism because it focused on the overall style of the posters, thereby recognizing the characteristics of each genre poster, while the attention mechanism focused only on structural features of the images.

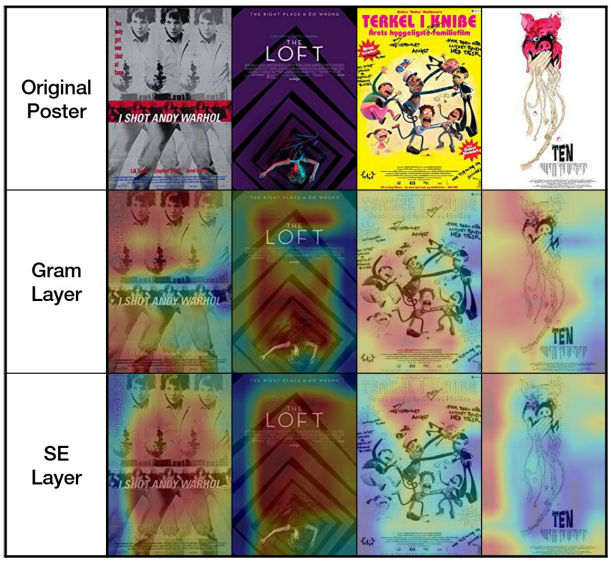

FIGURE 5. Comparison of SE layer and Gram layers for class activation mapping [49] visualization on movie poster.

3) FOCUS ON THE MOVIE POSTER BY GENRE

We conducted experiments to investigate which part of the poster was concentrated on by the network, using the Gram layer, by genre. Similar to the process described in the previous section, activation maps were extracted from layer 4 of the ResNet18 model using the Gram layer. After setting one genre as the ground truth, the extracted activation map was visualized, and the 12 genres were investigated. Fig. 6 displays the results of visualizing the activation map of posters that show the characteristics of the 12 genres in the test data.

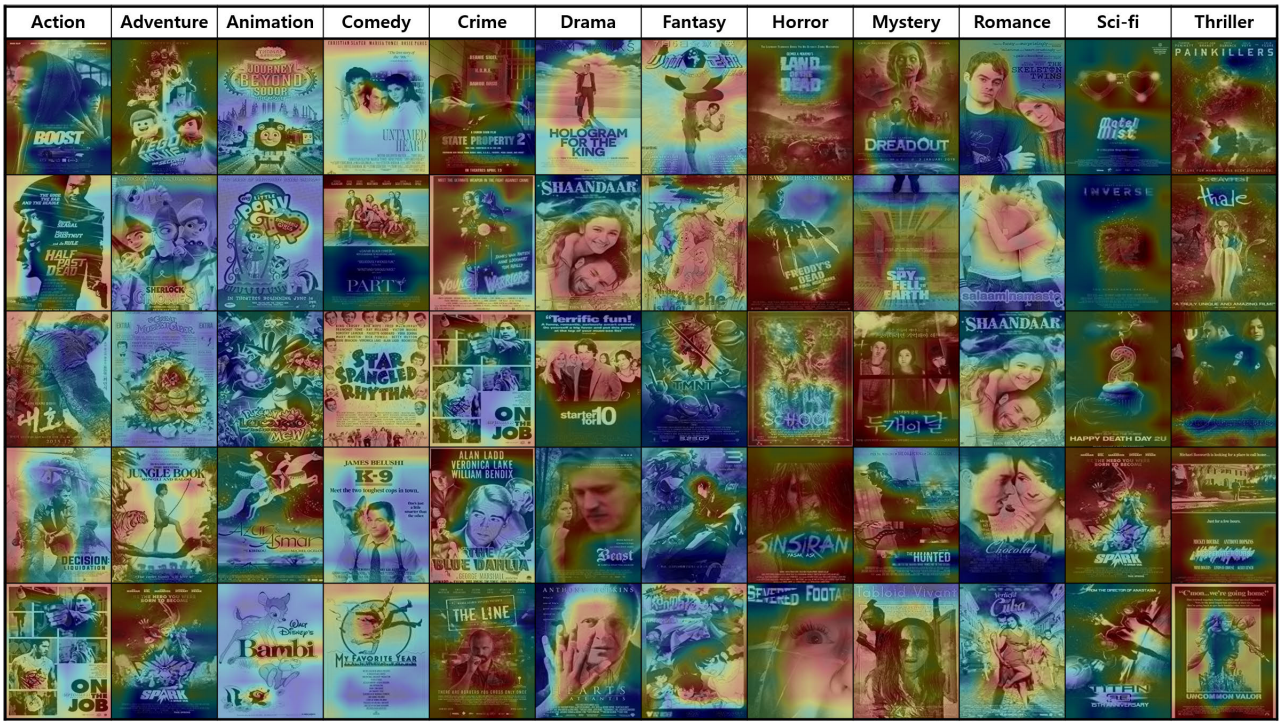

FIGURE 6. Visualization of activation map of 12 movie genres posters.

Notable points from the results are as follows: In the case of the Action genre, in the posters, we observed a focus on weapons such as guns and knives in the second column. For the Adventure genre, characters in the posters were primarily focused on; for the Animation genre, characters or titles that showed animation were focused on. The Fantasy genre also exhibited a focus on characters. The Comedy and Drama genres showed a general focus on people and atmosphere and were thus interpreted as the two genres that spanned the widest range. The Horror, Mystery, and Thriller genres emphasized on dark feelings throughout the poster to portray a scary atmosphere, which can also be interpreted as a focus on the background of the poster, representing the atmosphere. In the case of Romance, a man and woman occupied many of the posters, and the posters showed a concentration on the characters. The Sci-Fi genre showed a focus on the background and characters, representing the atmosphere of the poster.

Therefore, in the case of a network using a Gram layer, it can be observed that posters of various genres are classified, not only by the same part of the poster but also by the parts of the poster that characterize each genre.

V. CONCLUSION

This study proposes a method of classifying movie posters into multi-label genres by utilizing a Gram layer. The proposed Gram layer extracts style features using the feature summarizing inter-channel relations for input feature maps. Then, we generate the style weight through the fully connected layer for the extracted style features and generate the weighted input feature map through the input feature map and multiplication.

In experiments, the proposed Gram layer could identify the characteristics of the movie poster and the correlations between each genre, thus enabling more accurate multi-genre classification. Moreover, it can be applied to various CNN architectures. Activation map visualization showed that the proposed Gram layer was more focused on the content-based features on movie posters than on the attention mechanism, which focused on simple features of the images. We confirmed that the Gram layer reliably understands the characteristics of each genre because it uses the style features of the movie posters.

Although existing movie poster datasets lack many movies or have images that are insufficient to be regarded as posters, they are still used in poster genre classification. To use consistent style features from posters of each genre, a new dataset was created that primarily comprised movies from the late 20th century and the early 2000s. The data imbalance for rare genres was eliminated by limiting the number of defined genres to 12.

Future work will be directed at confirming that the method of weighting the style features, via insertion of Gram layers into the CNN architecture, may also be applied to other categorization tasks. This can be accomplished by conducting experiments with a variety of datasets containing, for example, cartoon and art works that possess a consistent style. In addition, a study will be conducted to determine

the correlation between the position and the number of Gram layers to achieve optimal performance.

REFERENCES

[1] W.-T. Chu and H.-J. Guo, "Movie genre classification based on poster images with deep neural networks," in *Proc. Workshop Multimodal Understand. Social, Affect. Subjective Attributes*, 2017, pp. 39–45.

[2] M. Pobar and M. Ivasic-Kos, "Multi-label poster classification into genres using different problem transformation methods," in *Proc. Int. Conf. Comput. Anal. Images Patterns*. Cham, Switzerland: Springer, 2017, pp. 367–378.

[3] J. Wehrmann, M. A. Lopes, and R. C. Barros, "Self-attention for synopsis-based multi-label movie genre classification," in *Proc. 31st Int. Flairs Conf.*, 2018, pp. 1–6.

[4] A. C. Saputra, A. B. Sitepu, Stanley, P. W. P. Y. Sigit, P. G. S. A. Tetuko, and G. C. Nugroho, "The classification of the movie genre based on synopsis of the Indonesian film," in *Proc. Int. Conf. Artif. Intell. Inf. Technol. (ICAIIT)*, Mar. 2019, pp. 201–204.

[5] K. Sivaraman and G. Somappa, "MovieScope: Movie trailer classification using deep neural networks," Univ. Virginia, Charlottesville, VA, USA, Tech. Rep., 2016.

[6] J. Wehrmann and R. C. Barros, "Movie genre classification: A multi-label approach based on convolutions through time," *Appl. Soft Comput.*, vol. 61, pp. 973–982, Dec. 2017.

[7] M. Ivasic-Kos, M. Pobar, and I. Ipsic, "Automatic movie posters classification into genres," in *Proc. Int. Conf. ICT Innov.* Cham, Switzerland: Springer, 2014, pp. 319–328.

[8] M. Ivasic-Kos, M. Pobar, and L. Mikec, "Movie posters classification into genres based on low-level features," in *Proc. 37th Int. Conv. Inf. Commun. Technol., Electron. Microelectron. (MIPRO)*, May 2014, pp. 1198–1203.

[9] Z. Fu, B. Li, J. Li, and S. Wei, "Fast film genres classification combining poster and synopsis," in *Proc. Int. Conf. Intell. Sci. Big Data Eng.* Cham, Switzerland: Springer, 2015, pp. 72–81.

[10] K. Kundalia, Y. Patel, and M. Shah, "Multi-label movie genre detection from a movie poster using knowledge transfer learning," *Augmented Hum. Res.*, vol. 5, no. 1, p. 11, Dec. 2020.

[11] F. Álvarez, F. Sánchez, G. Hernández-Peñaloza, D. Jiménez, J. M. Menéndez, and G. Cisneros, "On the influence of low-level visual features in film classification," *PLoS ONE*, vol. 14, no. 2, 2019, Art. no. e0211406.

[12] S. Sirattanajakarin and P. Thusaranon, "Movie genre in multi-label classification using semantic extraction from only movie poster," in *Proc. 7th Int. Conf. Comput. Commun. Manage.*, Jul. 2019, pp. 23–27.

[13] J.-M. Perez-Rua, V. Vielzeuf, S. Pateux, M. Baccouche, and F. Jurie, "MFAS: Multimodal fusion architecture search," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6966–6975.

[14] P. Cascante-Bonilla, K. Sitaraman, M. Luo, and V. Ordonez, "Moviescope: Large-scale analysis of movies using multiple modalities," 2019, *arXiv:1908.03180*. [Online]. Available: http://arxiv.org/abs/1908.03180

[15] H. Zhou, T. Hermans, A. V. Karandikar, and J. M. Rehg, "Movie genre classification via scene categorization," in *Proc. Int. Conf. Multimedia*, 2010, pp. 747–750.

[16] G. S. Simoes, J. Wehrmann, R. C. Barros, and D. D. Ruiz, "Movie genre classification with convolutional neural networks," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2016, pp. 259–266.

[17] G. Portolese, M. A. Domingues, and V. D. Feltrim, "Exploring textual features for multi-label classification of portuguese film synopses," in *Proc. EPIA Conf. Artif. Intell.* Cham, Switzerland: Springer, 2019, pp. 669–681.

[18] E. Noersasongko, D. S. Ervan, H. A. Santoso, C. Supriyato, F. Al Zami, and M. Soeleman, "The use of particle swarm optimization to obtain n-gram optimum value for movie genre classification," *J. Theor. Appl. Inf. Technol.*, vol. 97, no. 20, pp. 1–11, 2019.

[19] S. Saumya, J. Kumar, and J. P. Singh, "Genre fraction detection of a movie using text mining," in *Advanced Computing and Systems for Security*. Singapore: Springer, 2018, pp. 167–177.

[20] G. Portolese and V. D. Feltrin, "On the use of synopsis-based features for film genre classification," in *Proc. Anais do XV Encontro Nacional Inteligência Artif. Computacional (ENIAC)*, Oct. 2018, pp. 892–902.

[21] A. M. Ertugrul and P. Karagoz, "Movie genre classification from plot summaries using bidirectional LSTM," in *Proc. IEEE 12th Int. Conf. Semantic Comput. (ICSC)*, Jan. 2018, pp. 248–251.

[22] A. Austin, E. Moore, U. Gupta, and P. Chordia, "Characterization of movie genre based on music score," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2010, pp. 421–424.

[23] M. Sugano, R. Isaksson, Y. Nakajima, and H. Yanagihara, "Shot genre classification using compressed audio-visual features," in *Proc. Int. Conf. Image Process.*, vol. 2, Sep. 2003, p. II-17.

[24] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2414–2423.

[25] F. Luan, S. Paris, E. Shechtman, and K. Bala, "Deep photo style transfer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4990–4998.

[26] W.-T. Chu and Y.-L. Wu, "Image style classification based on learnt deep correlation features," *IEEE Trans. Multimedia*, vol. 20, no. 9, pp. 2491–2502, Sep. 2018.

[27] C. Sandoval, E. Pirogova, and M. Lech, "Two-stage deep learning approach to the classification of fine-art paintings," *IEEE Access*, vol. 7, pp. 41770–41781, 2019.

[28] S. Grützmacher, R. Kemkemer, and C. Curio, "Using deep correlation features to define the meta style of cell images for classification," *Current Directions Biomed. Eng.*, vol. 5, no. 1, pp. 227–230, Sep. 2019.

[29] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*. [Online]. Available: http://arxiv.org/abs/1409.0473

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[31] D. Hu, "An introductory survey on attention mechanisms in NLP problems," in *Proc. SAI Intell. Syst. Conf.* Cham, Switzerland: Springer, 2019, pp. 432–448.

[32] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.

[33] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3146–3154.

[34] X. Wang, X. He, Y. Cao, M. Liu, and T.-S. Chua, "KGAT: Knowledge graph attention network for recommendation," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2019, pp. 950–958.

[35] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[36] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[37] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[39] A. G. Roy, N. Navab, and C. Wachinger, "Recalibrating fully convolutional networks with spatial and channel-squeeze and excitation blocks," *IEEE Trans. Med. Imag.*, vol. 38, no. 2, pp. 540–549, Feb. 2019.

[40] F. M. Harper and J. A. Konstan, "The MovieLens datasets: History and context," *ACM Trans. Interact. Intell. Syst.*, vol. 5, no. 4, pp. 1–19, Jan. 2016.

[41] Y. Li, N. Wang, J. Liu, and X. Hou, "Demystifying neural style transfer," 2017, *arXiv:1701.01036*. [Online]. Available: http://arxiv.org/abs/1701.01036

[42] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3156–3164.

[43] S. Godbole and S. Sarawagi, "Discriminative methods for multi-labeled classification," in *Proc. Pacific–Asia Conf. Knowl. Discovery Data Mining*. Berlin, Germany: Springer, 2004, pp. 22–30.

[44] M. S. Sorower, "A literature survey on algorithms for multi-label learning," Oregon State Univ., Corvallis, OR, USA, Tech. Rep., 2010, pp. 1–25, vol. 18.

[45] T. Zhang, "Solving large scale linear prediction problems using stochastic gradient descent algorithms," in *Proc. 21st Int. Conf. Mach. Learn. (ICML)*, 2004, p. 116.

[46] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.
[47] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.
[48] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.
[49] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.

JEONG A. WI (Student Member, IEEE) received the B.S. degree in physics from Chung-Ang University, Seoul, South Korea, in 2019, where she is currently pursuing the M.S. degree in imaging engineering with the Graduate School of Advanced Imaging Science, Multimedia, and Film, Chung-Ang University. Her current research interests include deep learning and computer graphics.

SOOJIN JANG (Student Member, IEEE) received the B.S. degree in information and communication engineering from Sunmoon University, Chungnam, South Korea, in 2018. She is currently pursuing the M.S. degree in imaging engineering with the Graduate School of Advanced Imaging Science, Multimedia, and Film, Chung-Ang University. Her research interests include deep learning and computer vision.

YOUNGBIN KIM (Member, IEEE) received the B.S. and M.S. degrees in computer science and the Ph.D. degree in visual information processing from Korea University, in 2010, 2012, and 2017, respectively. From August 2017 to February 2018, he was a Principal Research Engineer at Linewalks. He is currently an Assistant Professor with the Graduate School of Advanced Imaging Science, Multimedia, and Film, Chung-Ang University. His current research interests include data mining and deep learning.