Assessment System of Presentation Slide Design Using Visual and Structural Features*

Shengzhou YI^{†a)}, Junichiro MATSUGAMI^{††b)}, Nonmembers, and Toshihiko YAMASAKI^{†c)}, Member

SUMMARY Developing well-designed presentation slides is challenging for many people, especially novices. The ability to build high quality slideshows is becoming more important in society. In this study, a neural network was used to identify novice vs. well-designed presentation slides based on visual and structural features. For such a purpose, a dataset containing 1,080 slide pairs was newly constructed. One of each pair was created by a novice, and the other was the improved one by the same person according to the experts' advice. Ten checkpoints frequently pointed out by professional consultants were extracted and set as prediction targets. The intrinsic problem was that the label distribution was imbalanced, because only a part of the samples had corresponding design problems. Therefore, re-sampling methods for addressing class imbalance were applied to improve the accuracy of the proposed model. Furthermore, we combined the target task with an assistant task for transfer and multi-task learning, which helped the proposed model achieve better performance. After the optimal settings were used for each checkpoint, the average accuracy of the proposed model rose up to 81.79%. With the advice provided by our assessment system, the novices significantly improved their slide design.

key words: presentation slide, class imbalance, feature fusion, transfer learning, multi-task learning

1. Introduction

Presentation slides are commonly used in business and academia. A well-designed slide can help a presenter fit their messages to their oral presentation, making the exposition very easy to understand. However, there remains a large gap in slide design skills between novices and experts, and there are a limited number of support systems available to assess the quality of slide design.

Although making well-designed slides is an important problem, there are only a few researches that tackles this problem [1], [2]. Besides, [1], [2] only classified or regressed the quality of the presentation slides. In other words, to the best of our knowledge, there has been no system that can give some feedback on how to improve the slides so far. In fact, our previous system [2] is already used in a business service**, and the most of the requests from the customers were more detailed advice on how to improve it.

By taking this into consideration, in order to analyze

how novices improve their slides with professional help, we collected the data from a training course under their consensus, which aims to help novices create professional presentation slides. In the course, the participants created single-page slides to describe some services or products. Then, they got modification advice from professional consultants, and they improved their slides by themselves according to the advice. For these informative slides, it is the most important to keep them easy to understand and well-structured. We used the collected data for our presentation-slide assessment system. It contains 1,080 slide pairs. For each slide pair, the original one was created by a novice without any help, and the other one was modified by the same person according to the advice from consultants. Based on all advice given to the novices, the consultants summarized the top-10 important checkpoints in slide design for the novices. They should check the lack of fundamental elements (e.g., title, messages, and subheading), whether the important words and areas are emphasized, whether the slides are well-structured, and so on. We would like to clarify that the slide designs are standard ones in business and academic areas. Some people can make very impressive presentations with unique-style slide design, but this is out of our focus.

Because the checkpoints proposed by the consultants are considered from the aspect of visual or structure, we applied a bimodal neural network for analyzing visual and structural features to check the slide design. The binary classification of each checkpoint exhibits a class imbalance, because the considered design problems only exist in a part of slide samples. Both of traditional and proposed re-sampling methods were used to address the imbalanced distribution. An assistant task, that of recognizing whether the slides are novice or modified, was proposed to enhance the capacity of the encoder in feature extraction. It was combined with the target task to conduct transfer and multi-task learning. Finally, we proposed to combine class-imbalanced, transfer learning, and multi-task learning together in an optimal way. According to the modification advice provided by our assessment system, the novices made better presentation slides. We are constructing an online service in order that the novices can upload their slides and get the advice conveniently.

Here, we would like to clarify the scope of our paper. We focus on visual quality of a single slide. Therefore, we

Manuscript received April 16, 2021.

Manuscript revised September 12, 2021.

Manuscript publicized December 1, 2021.

[†]The authors are with The University of Tokyo, Tokyo, 113–8656 Japan.

^{††}The author is with Rubato Co., Ltd., Tokyo, 102–0073 Japan.

*This is a paper on system development.

a) E-mail: yishengzhou@cvm.t.u-tokyo.ac.jp

b) E-mail: matsugami@rubato.co

c) E-mail: yamasaki@cvm.t.u-tokyo.ac.jp

DOI: 10.1587/transinf.2021HCK0001

**<https://www.presentationtrainer.jp/>

do not discuss the logical structure of the whole presentation nor linguistic expressions of the slides. The aim of our work is to give proper feedback to the users, not developing a machine learning system that can automatically modify the slides. In this work, only whether the checkpoints are satisfied or not is classified. One may think that it is better to show more concrete and detailed feedback to the user, but there are pros and cons. The most serious problem of giving concrete and detailed feedback is that users stop learning and they just rely on the feedback. The aim of our research is to effectively train the users by our system. Therefore, we believe that giving the classification results and letting the user to think about it is the best training tool.

The main contributions of this study are:

- A slide dataset assigned with various evaluation labels was created.
- A bimodal neural network was proposed to recognize the design problems of slides and provide modification advice to the novices.
- New sampling methods to resolve the unbalanced data problem are proposed: dynamic over-sampling (DOS) and dynamic under-sampling (DUS).
- Class-imbalanced, transfer learning, and multi-task learning methods were applied and combined in an optimal way, and they significantly enhanced the proposed neural network.
- We did a small-scale experiment on five novices to prove the effectiveness of the training system.

2. Related Works

2.1 Presentation Slide Analysis

A work on presentation slide analysis was presented in [1]. The work performed three class classifications (high, fair, and low) of slides with an accuracy of 62.2%. They classified the quality of multi-page slides using a small dataset in which 178 slides collected from SlideShare were evaluated by human using absolute scores. They used 28 quality indicators that were hand-crafted for slide features. Oyama and Yamasaki [2] proposed an evaluation system using convolutional neural networks (CNNs) to judge the quality of slides, but the ratings made by their system lacked of detailed criterion (i.e., only a single measure of goodness of the slides).

Yokota et al. [3] studied slide retrieval techniques that considered the relationship between slides, videos, voices, and laser pointer information. Winters and Mathewson [4] proposed an automatic generation system of presentation slides based on a topic suggestion. Yi et al. [5], [6] proposed a multimodal neural network to recognize the audiences' impressions on the presentation videos of TED Talks[†]. However, these studies cannot recognize the design problems of slides to help the novices improve their slide design skills.

The novelty of this paper is that we have clarified the important factors for the good slide design and developed an assessment tool that can check whether each of the checkpoints is satisfied.

2.2 Class-Imbalanced Learning

Re-sampling training data can balance classes consisting of various samples by either over-sampling the minority classes [7], [8] or under-sampling the majority classes [8], [9]. Another method was proposed to re-weight the loss using inverse class frequency [10]–[12]. However, this method usually results in overfitting the minority classes, leading to poor performance with extreme class imbalance. Consequently, Cui et al. [13] proposed a smoother re-weighting factor based on the concept of the effective number of samples used to mitigate overfitting. However, it is proposed for multi-class classification and not suitable for the binary classification task in our study. In this paper, we propose a new over-sampling and under-sampling strategies.

2.3 Transfer and Multi-Task Learning

Transfer learning improves the performance of machine-learning models by transferring knowledge from the source domain to the target. Consequently, the dependence on the large dataset size of the target domain can be reduced to construct the target model. Transfer learning is commonly performed in the research of natural language processing that uses distributed word representations learned from a large corpus [14], [15]. In computer vision, researchers commonly use neural networks that are pre-trained on large-scale image datasets, such as ImageNet [16] and Microsoft COCO [17].

Multi-task learning aims to jointly learn several related tasks to enhance the generalization of each. This is typically performed with either hard- or soft-parameter sharing of the hidden layers. Hard parameter sharing is the most commonly used method for multi-task learning and involving sharing the hidden layers between all tasks [18]. In soft-parameter sharing, each task has an individual model and parameters. The distance between these parameters in the model is regularized to encourage the detection of similar parameters [19].

3. Slide Improvement Dataset

3.1 Checkpoints

The dataset contains the slide data from a training course, where novices learned to create high quality slides used to describe the features of services or products. The slides created by the novices include much information, but most of them are not easy to understand and not well-structured. Five professional consultants with more than 10 years' experience provided their advice of improving the slides. According to the consultants' advice, the novices themselves

[†]<https://www.ted.com>

Fig. 1 Slide pair samples of top-10 checkpoints.

modified their slides. Note that the slides were not modified by the consultants to let the novices learn how to update their own slides. The slides created by the novices and the slides modified by the same person based on consultants' advice were collected as slide pairs. The dataset contains 1,080 such slide pairs in total. Different from the consultants' advice which is detailed and specific to the content of each slide, machine-learning methods can only provide relatively rough and general modification advice based on the relatively limited data. By analyzing the created slides and the advice given to the novices, the consultants summarized the top-10 important and common checkpoints in slide design, including

1. **Insert a pictogram.** Pictogram insertion can intuitively convey the slide content.
2. **Add a subheading.** Subheading addition makes the audiences convenient to find information.
3. **Emphasize words.** Emphasizing important words enhances a part of a text to make it noticeable.
4. **Emphasize areas.** Emphasizing important areas of the slides helps focus attention.
5. **Add T1 and T2.** Slide title (T1) and slide messages (T2) help the audiences more easily understand the main content of the exposition.
6. **Use the grid structure.** A grid structure helps organize target and comparison items in rows and columns, creating a table-like format.

7. **Itemize the text.** Text itemization is good in terms of readability and organization.
8. **Add a comment.** Comments aid in audiences' understanding.
9. **Correct flow.** Left-to-right top-to-bottom flow mimics human scanning methods.
10. **MECE.** Mutually exclusive and collectively exhaustive (MECE) objects should be aligned without omissions and duplication to make the slides convincing.

Figure 1 presents improved samples of the checkpoints listed above. In each block, the slide on the left is created by a novice, and the slide on the right is modified by the same person according to the consultants' advice. The red arrow indicates the modified areas at each checkpoint.

The considered checkpoints contain the most common problems when presenting a lot of information in a single-page slide. When creating the slides, many novices ignored to use the fundamental slide elements, such as slide title, slide messages, subheadings, and comments. Without these elements, it will be difficult for the audiences to capture the main point quickly. While pictogram and emphasis are commonly used for experienced presenters, most novices tend to present all things only by text, whose color is always black. Furthermore, it is more difficult for the novices to take care of the structure and logical relationship between the contents. Therefore, only a very small part of novices in the dataset used a structural layout and consider the flow di-

rection and MECE. The slide assessment system reminds the novices of the design problems in their slides according to the checkpoints above, and then they can understand the shortage of slide elements and how to improve their slides.

3.2 Label Distribution

The consultants assigned the labels of the checkpoints to the slide pairs created by the novices. There are four types of labels: (A) need modification and modified correctly; (B) no need to modify; (C) need modification but not modified properly yet; (D) not applicable. Table 1 shows the number of samples in each class of the slide pairs. The number of samples with the label “B” is larger than that with the label “C,” because only a relatively small part of slides has the corresponding design problem in the dataset.

Because the recognition of design problems is based on single-page slides, we mapped the labels of the slide pairs into the labels of the single-page slides using the following rules. A positive class represents a slide with a corresponding design problem. A negative class implies that the slide does not have the corresponding problem. If the label of

Table 1 Label distribution of slide pairs. The meanings of the labels are shown as follows: (A) the consultants pointed out that the slides need modification for the checkpoint, and the novices modified them correctly; (B) the consultants thought that the slides were well-designed and no need to modify for the checkpoint; (C) the consultants pointed out that the slides need modification, but the novices did not modified them properly yet; (D) the checkpoints are not applicable to the slides.

Check point	A	B	C	D
(1) Insert a pictogram	390	406	230	54
(2) Add a subheading	512	531	3	34
(3) Emphasize words	569	252	161	98
(4) Emphasize areas	291	95	83	611
(5) Add T1 and T2	526	494	39	21
(6) Use the grid structure	278	192	3	607
(7) Itemize the text	391	458	47	184
(8) Add a comment	90	43	47	900
(9) Correct flow	173	883	4	20
(10) MECE	391	683	4	2

the slide pair is “A,” the labels of the slides before and after modification are positive and negative, respectively. If the label of the slide pair is “B,” the labels of the slides before and after modification are both negative. If the label of the slide pair is “C,” the labels of the slides before and after modification are both positive. We do not consider the slide pair with the label “D.” Table 2 presents the number of positive and negative samples in the training and validation sets. The imbalance ratio (IR) represents the proportion of samples in the majority class to the number of minority class in the training set. Following the research of imbalanced learning on small-scale datasets [13], [20], the validation set maintained a balanced distribution.

4. Machine Learning on Presentation Slides

4.1 Feature Learning

The consultants gave their advice for slide improvement from the aspect of visual or structure, and therefore the checkpoints in this study were also proposed from these two aspects. We consider that some checkpoints are more easily considered by using the visual features, and some checkpoints are more relevant to the slide structure. For example, the pictograms have an unique look, and the emphasized words and areas usually have different colors. In contrast with these checkpoints, grid structure and itemization are more easily detected by focusing on the structural features. We built a bimodal neural network to capture both the visual and structural features of the slides to find whether the proposed checkpoints are suitable to be handled by machine-learning methods. Instead of artificially deciding the feature extraction strategies (i.e., visual or structural features only, or bimodal features) by our experience, we choose the optimal one according to the experimental results.

4.1.1 Overview

Figure 2 presents an overview of the proposed neural network. It shows the size of the representation immediately after each component. It contains a CNN architecture to ex-

Table 2 Label distribution of single-page slides.

Set	Training			Validation		
	Positive	Negative	IR	Positive	Negative	IR
(1) Insert a pictogram	680	1,032	1.52	170	170	1.00
(2) Add a subheading	415	1,471	3.54	103	103	1.00
(3) Emphasize words	713	895	1.26	178	178	1.00
(4) Emphasize areas	366	390	1.07	91	91	1.00
(5) Add T1 and T2	484	1,394	2.88	120	120	1.00
(6) Use the grid structure	228	606	2.66	56	56	1.00
(7) Itemize the text	388	1,210	3.12	97	97	1.00
(8) Add a comment	148	140	1.06	36	36	1.00
(9) Correct flow	145	1,903	13.12	36	36	1.00
(10) MECE	320	1,678	5.24	79	79	1.00

Fig. 2 Overview of the proposed neural network.

tract visual features. It also includes a sampling method to extract structural features. These features are then embedded to feature vectors of the same dimensions for additional feature fusion. Finally, the linear classifier separates the slide samples into two categories (positive with corresponding problems and negative without corresponding problems) based on the learned bimodal features.

4.1.2 Visual Features

The architecture for extracting the visual features of slides is based on a high-performance CNN, ResNet-50 [21]. We used the feature maps of the convolutional block, Conv3 of ResNet-50, to extract the image features. Considering the relatively small size of the slide dataset, we did not use ResNet-101 or ResNet-152, which have deep architectures but more easily cause overfitting on small-scale datasets. The convolution block is followed by a global average pooling (GAP) layer [22], which reduces the size of the preceding layers by taking the average of each feature map. GAP layers have shown great power in minimizing overfitting by reducing the total number of parameters in the model. GAP layers can reduce the spatial dimension of visual features, where the feature vectors having the shape of $h \times w \times d$ is reduced in size to have the shape of $1 \times 1 \times d$. Each $h \times w$ feature map is then mapped to a single number by taking the average.

4.1.3 Structural Features

We follow the procedure from [2]. The structural information of the presentation slides is described using the format

Fig. 3 Structural features.

of Office Open XML, which classifies the slide objects into five categories: background, cnvSp (connection), pic (picture), graphicFrame (table and figure), and sp (text). As shown in Fig. 3, we sampled $H \times W$ points from single-page slides. We used a one-hot vector to represent the category of the bounding box which the sampling point belongs to, and we concatenated all of the one-hot vectors together to represent the slide layout as a feature vector of $H \times W \times 5$ dimensions. Apart from the features of the sampling points, we also expressed the high-level features of the entire slide, including whether the slide is itemized list style, the number of objects, the number of words, and the maximum and minimum font sizes. Finally, the structural features are represented as a feature vector of $H \times W \times 5 + 5$ dimensions.

4.1.4 Feature Fusion

In our study, feature addition outperforms feature concatenation for feature fusion. To prepare the features for the feature fusion, we used a linear layer to embed the unimodal features to the feature space having the same dimension.

4.2 Class-Imbalanced Learning

Because the slide improvement dataset has an imbalanced distribution in the number of positive and negative samples, we used two traditional sampling methods and proposed two new re-sampling methods to explore their efficacy for improving classification accuracy. Traditional methods under-sample the majority classes and over-sample the minority classes, resulting in a number of samples remaining unlearned or weakly learned. To address the problem, we propose to re-sample the training data in a dynamic way, which we call dynamic over-sampling (DOS) and dynamic under-sampling (DUS). As shown in Fig. 4, compared with the traditional sampling that re-sample the data only once at the beginning of training stage, the dynamic re-sampling methods repeat re-sampling in every epoch and lasts for the whole training stage. The negative set is dynamically re-sampled, which helps the model learn from different data distribution (marked with yellow color). Consequently, the generalization of the model is improved.

Fig. 4 Sampling strategies of traditional under-sampling and dynamic under-sampling.

Fig. 5 Transfer and multi-task learning architecture.

4.3 Transfer Learning and Multi-Task Learning

We designed an assistant task (referred as Mod task), which is to recognize whether a slide is before or after the novices' modification. In contrast to the design-problem recognition task, Mod task has a balanced label distribution. The numbers of slides before and after modification are the same. Without the effects of imbalanced labels, the proposed model can focus on feature learning. Meanwhile, the Mod task is strongly related to the target task, because the slides after modification have a lower possibility of having the considered designed problems than the slides before modification.

Figure 5 presents the transfer and multi-task learning architecture. For transfer learning, the proposed neural net-

work was first trained for the source task (i.e., Mod task); then the encoder was fine-tuned, and the classifier was re-trained for the target task. We examined the performance of pre-training the encoder on a large-scale image dataset (i.e., ImageNet [16]). Comparing with the successive training of transfer learning, the prediction model for the source and target tasks were trained at the same time in multi-task learning. By the successive and joint learning of the tasks, the capacity of the encoder in feature learning is enhanced, and we consider it as a knowledge transfer.

5. Experiments

5.1 Implementation

PyTorch [23] was used as a deep learning framework in all

Table 3 Prediction accuracy (%) of different featur extraction strategies.

Checkpoint	Visual	Structural	Bimodal
(1) Insert a pictogram	82.29	69.06	82.41
(2) Add a subheading	73.06	75.82	76.50
(3) Emphasize words	88.03	71.74	86.86
(4) Emphasize areas	80.50	70.77	80.22
(5) Add T1 and T2	83.34	68.75	82.67
(6) Use the grid structure	70.00	78.04	75.18
(7) Itemize the text	73.20	75.36	75.05
(8) Add a comment	71.67	76.67	76.67
(9) Correct flow	56.67	53.06	57.78
(10) MECE	58.48	64.30	66.20
Average accuracy	73.72	70.36	75.95

experiments. We trained the model on the slide improvement dataset for 100 epochs with a batch size of 64. We used the stochastic gradient descent optimizer with a learning rate of 1×10^{-4} and a momentum of 0.9. The visual and structural features were mapped into a 256-dimensional feature space for the feature fusion. Although a single-page slide can have multiple labels, the labels of some checkpoints are not applicable for the slide. Therefore, we individually trained the proposed model for each checkpoint. The reported results were averaged over five runs.

5.2 Feature Learning

We performed binary classification for each checkpoint separately. We found that $H = W = 7$ was the optimal sampling rate for structural feature learning. Table 3 shows the prediction accuracy of using only visual features, only structural features, or bimodal features. The accuracy of using visual features was higher than that of using structural features for the checkpoints strongly related to the visualization (e.g., insert a pictogram and emphasize important words/areas). However, it was lower for some checkpoints considering the layout (e.g., use the grid structure). Because of the high prediction accuracy, the structural features were proved to be effective to capture the information of slide layout. For many checkpoints, the combination of visual and structural features are helpful to achieve better performance. The bimodal features improved the performance of the proposed model with an average accuracy of 75.95%.

We used class activation mapping [22] to visualize the feature map immediately after GAP layer. The labels of these slide samples were negative (not having the corresponding design problem), and the heat maps show the well-designed areas. The heat maps assigned with the weights to the negative class for checkpoints (1)-(4) are shown in Fig. 6. The images on the left are the original slides, and the images on the right are the heat maps. We can easily see that the highlighted areas are strongly related to the corresponding checkpoints and demonstrate the effectiveness of the proposed model.

We tried to consider the slide content and logical rela-

Fig. 6 Heat maps showing the well-designed areas.

tionship by extracting linguistic features, especially for the checkpoints of correct flow and MECE, whose prediction accuracy is relatively low. However, the text in a slide pair usually stays unchanged, but the labels of the checkpoints changes after modification. Therefore, we found that the use of linguistic features brought much redundant information and worsened the performance of our assessment system.

5.3 Class-Imbalanced Learning

We examined our proposal using the bimodal features on the imbalanced dataset with various sampling methods described in Sect. 4.2, including uniform sampling (Uniform), over-sampling (OS), under-sampling (US), proposed dynamic over-sampling (DOS), and dynamic under-sampling (DUS). As shown in Table 4, DUS achieved the highest average accuracy of 79.18%. Repeating re-sampling in each epoch improved the performance compared with re-sampling for only once. Generally, the performance gain of re-sampling is more significant as IR increases.

5.4 Transfer and Multi-Task Learning

We pre-trained the neural network on the Mod task and the large-scale dataset, and we fine-tuned the model on the design-problem recognition task. Table 5 shows the experimental results of scratch (without pre-training), Mod (recognizing before or after modification) pre-training, and ImageNet (large-scale image classification) pre-training. Because ImageNet contains only image data, we used only

Table 4 Prediction accuracy (%) of different sampling methods using **bimodal features, no transfer and multi-task learning**.

Checkpoint	IR	Uniform	OS	US	DOS	DUS
(1) Insert a pictogram	1.52	82.41	81.65	81.47	83.00	83.24
(2) Add a subheading	3.54	76.50	78.64	78.64	78.84	79.71
(3) Emphasize words	1.26	86.86	86.07	86.12	87.36	86.91
(4) Emphasize areas	1.07	80.22	79.12	79.23	81.21	81.98
(5) Add T1 and T2	2.88	82.67	83.08	82.50	87.75	84.75
(6) Use the grid structure	2.66	75.18	74.82	74.47	75.54	75.36
(7) Itemize the text	3.12	75.05	75.98	75.98	76.08	78.76
(8) Add a comment	1.06	76.67	76.39	75.83	76.67	76.95
(9) Correct flow	13.12	57.78	70.55	71.11	68.61	70.83
(10) MECE	5.24	66.20	71.65	71.39	71.65	73.29
Average accuracy	-	75.95	77.80	77.67	78.67	79.18

Table 5 Prediction accuracy (%) of transfer learning using **visual features only and no multi-task learning**.

Checkpoint	Scratch (visual only)	Mod	ImageNet
(1) Insert a pictogram	82.29	81.47	84.06
(2) Add a subheading	73.06	76.60	73.69
(3) Emphasize words	88.03	87.58	87.81
(4) Emphasize areas	80.50	79.34	80.11
(5) Add T1 and T2	83.34	83.58	84.33
(6) Use the grid structure	70.00	71.79	70.71
(7) Itemize the text	73.20	74.64	73.51
(8) Add a comment	71.67	70.83	78.61
(9) Correct flow	56.67	61.39	64.72
(10) MECE	58.48	59.11	59.49
Average accuracy	73.72	74.63	75.71

Table 6 Prediction accuracy (%) of transfer and multi-task learning using **bimodal features**.

Checkpoint	Scratch	Transfer learning	Multi-task learning
(1) Insert a pictogram	82.41	81.18	83.77
(2) Add a subheading	76.50	78.48	78.06
(3) Emphasize words	86.86	86.42	86.12
(4) Emphasize areas	80.22	79.12	80.55
(5) Add T1 and T2	82.67	83.89	82.92
(6) Use the grid structure	75.18	75.00	76.61
(7) Itemize the text	75.05	75.77	76.18
(8) Add a comment	76.67	76.39	76.95
(9) Correct flow	57.78	63.89	61.11
(10) MECE	66.20	66.14	65.44
Average accuracy	75.95	76.63	76.77

visual features when studying the effect of transfer learning. According to the results, transfer learning improved the performance of the proposed network for most checkpoints. Pre-training on ImageNet achieved the highest average accuracy of 75.71%.

We compared transfer and multi-task learning by combining the target and Mod task in different architecture using the bimodal features. As described in Table 6, both transfer and multi-task learning effectively improved the prediction accuracy, and multi-task learning achieved relatively better performance with an average accuracy of 76.77%.

5.5 Optimal Settings

According to the experimental results, we used the optimal settings of each checkpoint to combine feature, class-imbalanced, transfer learning, and multi-task learning together. Table 7 presents the optimal settings of each checkpoint and the corresponding prediction accuracy. For example, the optimal settings of checkpoint “(1) insert a pictogram” used the bimodal features, re-sampled the training with DUS, pre-trained the visual encoder on ImageNet, and fine-tuned the network in the architecture of multi-task

learning. Note that the classifier is trained for each checkpoint. Because the classifiers are trained independently, there is no data leakage among different labels. One drawback of this approach is that we need more storage in proportion to the number of classes. With the optimal settings, the proposed neural network achieved an average accuracy of 81.79%.

Qualitative Results. Figure 7 shows the t-SNE [24] projection of the bimodal features learned in the training and validation sets for the checkpoint “(1) insert a pictogram.” We can see that the extracted features are discriminative.

5.6 Use Cases

In order to verify our assessment system, we did a small-scale experiments on five novices, each of whom created two slides and improved the slides based on the feedback of the system. Figure 8 shows the slide examples, which were improved by the novices. Although the advice provided by our assessment system are relatively abstracted than the consultants’ detailed advice, the novices can easily know the shortage of their slide design, and then they added the suggested elements and structures, which made their slides eas-

Table 7 Prediction accuracy (%) of the optimal settings.

Checkpoint	Feature learning	Class-imbalanced learning	Transfer learning	Multi-task learning	Accuracy
(1) Insert a pictogram	Bimodal	DUS	ImageNet	Multi	87.59
(2) Add a subheading	Bimodal	DUS	Mod	Single	79.98
(3) Emphasize words	Visual	DOS	Scratch	Single	88.76
(4) Emphasize areas	Visual	DUS	Scratch	Multi	82.42
(5) Add T1 and T2	Bimodal	DUS	ImageNet	Single	87.00
(6) Use the grid structure	Structural	DOS	Scratch	Multi	79.46
(7) Itemize the text	Bimodal	DUS	Scratch	Multi	79.59
(8) Add a comment	Bimodal	DUS	ImageNet	Multi	86.94
(9) Correct flow	Bimodal	US	ImageNet	Single	71.39
(10) MECE	Bimodal	DUS	ImageNet	Single	74.81
Average accuracy	-	-	-	-	81.79

Fig. 7 t-SNE visualization of learned features.

Fig. 8 Use cases.

ier to understand with a clearer structure.

6. Conclusion

We created a presentation slide dataset containing 1,080 slide pairs, each representing a slide before or after modification. Consultants summarized the top-10 common checkpoints of slide creation for the novices. We proposed a bimodal neural network to recognize whether the design meets the requirement of the checkpoints. By combining visual and structural features, the proposed model improved the average accuracy by 2.25% points compared with using visual features only. The proposed DUS outperformed the uniform sampling by 3.23% points. Transfer and multi-task learning significantly improved the performance of the proposed model by enhancing the capacity of the encoder in feature extraction. Transfer learning from ImageNet improved the performance of the scratch without pre-training by 1.99% points using the visual features only. The multi-task learning with Mod task improved the accuracy of the scratch by 0.82% points using the bimodal features. Finally, the proposed model achieved an average accuracy of 81.79% by combining all learning methods explored in this study in the optimal way for each checkpoint. Based on the advice provided by our assessment system, the novices succeeded in making their slide design easier to understand and well-structured.

Acknowledgments

This work is partially financially supported by JSPS KAKENHI Grant Number JP19K22863.

References

- [1] S. Kim, W. Jung, K. Han, J.-G. Lee, and Y.Y. Mun, "Quality-based automatic classification for presentation slides," ECIR, vol.8416, pp.638–643, 2014.
- [2] S. Oyama and T. Yamasaki, "Visual clarity analysis and improvement support for presentation slides," 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM), pp.421–428, IEEE,

- 2019.
- [3] H. Yokota, T. Kobayashi, T. Muraki, and S. Naoi, "Uprise: Unified presentation slide retrieval by impression search engine," *IEICE Trans. Inf. & Syst.*, vol.87, no.2, pp.397–406, 2004.
 - [4] T. Winters and K.W. Mathewson, "Automatically generating engaging presentation slide decks," *International Conference on Computational Intelligence in Music, Sound, Art and Design (Part of EvoStar)*, vol.11453, pp.127–141, Springer, 2019.
 - [5] S. Yi, X. Wang, and T. Yamasaki, "Impression prediction of oral presentation using lstm and dot-product attention mechanism," *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, pp.242–246, IEEE Computer Society, 2019.
 - [6] S. Yi, H. Yumoto, X. Wang, and T. Yamasaki, "Presentationtrainer: Oral presentation support system for impression-related feedback," *Proc. AAAI Conference on Artificial Intelligence*, vol.34, no.9, pp.13644–13645, 2020.
 - [7] Y. Cui, Y. Song, C. Sun, A. Howard, and S. Belongie, "Large scale fine-grained categorization and domain-specific transfer learning," *Proc. IEEE conference on computer vision and pattern recognition*, pp.4109–4118, 2018.
 - [8] M. Buda, A. Maki, and M.A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural networks*, vol.106, pp.249–259, 2018.
 - [9] H. He and E.A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol.21, no.9, pp.1263–1284, 2009.
 - [10] C. Huang, Y. Li, C.C. Loy, and X. Tang, "Learning deep representation for imbalanced classification," *Proc. IEEE conference on computer vision and pattern recognition*, pp.5375–5384, 2016.
 - [11] Y.X. Wang, D. Ramanan, and M. Hebert, "Learning to model the tail," *Advances in neural information processing systems*, pp.7029–7039, 2017.
 - [12] S.H. Khan, M. Hayat, M. Bennamoun, F.A. Sohel, and R. Togneri, "Cost-sensitive learning of deep feature representations from imbalanced data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol.29, no.8, pp.3573–3587, 2017.
 - [13] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," *Proc. IEEE conference on computer vision and pattern recognition*, pp.9268–9277, 2019.
 - [14] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.
 - [15] J. Pennington, R. Socher, and C.D. Manning, "Glove: Global vectors for word representation," *Proc. 2014 conference on empirical methods in natural language processing (EMNLP)*, pp.1532–1543, 2014.
 - [16] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," *2009 IEEE conference on computer vision and pattern recognition*, pp.248–255, Ieee, 2009.
 - [17] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C.L. Zitnick, "Microsoft coco: Common objects in context," *European conference on computer vision*, vol.8693, pp.740–755, Springer, 2014.
 - [18] R. Caruana, "Multitask learning," *Machine learning*, vol.28, no.1, pp.41–75, 1997.
 - [19] L. Duong, T. Cohn, S. Bird, and P. Cook, "Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser," *Proc. 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing (volume 2: short papers)*, pp.845–850, 2015.
 - [20] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma, "Learning imbalanced datasets with label-distribution-aware margin loss," *Advances in neural information processing systems*, pp.1567–1578, 2019.
 - [21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proc. IEEE conference on computer vision and pattern recognition*, pp.770–778, 2016.
 - [22] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," *Proc. IEEE conference on computer vision and pattern recognition*, pp.2921–2929, 2016.
 - [23] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.
 - [24] L.v.d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol.9, no. Nov, pp.2579–2605, 2008.

Shengzhou Yi is currently a Ph.D. student at Department of Information and Communication Engineering, Graduate School of Information Science Technology, The University of Tokyo. His research topics are oral presentation analysis, slide design assessment, and class-imbalanced learning.

Junichiro Matsugami earned a bachelor's degree from Doshisha University, and master's degrees both from Kobe University and University of East Anglia. After that, he joined the US strategy consulting firm, Monitor group, where he engaged in the marketing and sales projects of pharmaceutical companies. He moved to US-based NGO, Alliance Forum Foundation, and was involved in education and nutrition programs. Currently he is the CEO of Rubato Co., Ltd. He provides business consulting and documentation training to companies.

Toshihiko Yamasaki received the Ph.D. degree from The University of Tokyo. He is currently an Associate Professor at Department of Information and Communication Engineering, Graduate School of Information Science and Technology, The University of Tokyo. He was a JSPS Fellow for Research Abroad and a visiting scientist at Cornell University from Feb. 2011 to Feb. 2013. His current research interests include attractiveness computing based on multimedia big data analysis, pattern recognition, machine learning, and so on.