Shintaro Okada University of Tsukuba Tsukuba, Japan

Kosei Soda University of Tsukuba Tsukuba, Japan Chiharu Hirohana University of Tsukuba Tsukuba, Japan

Takehito Utsuro University of Tsukuba / RIKEN-AIP Tsukuba / Tokyo, Japan

Noriko Kando National Institute of Informatics Tokyo, Japan Kota Kawaguchi University of Tsukuba Tsukuba, Japan

Yasuhide Kawada Logworks Co., Ltd., Tokyo, Japan

## ABSTRACT

In last few years, the Internet and Web contents have become remarkable tools for studying. However, most search engines that can find Web contents applicable for studying are not beginner friendly. Learners must manually compare several pages on the search engine to find beginner friendly Web contents. Visual intelligibility in Web page layout and beginner friendly Web page texts are the requirements of Web contents for beginners. This study focuses on the factors of Web page visual intelligibility for explaining academic concepts. We apply a VGG16 fine-tuning technique, a pre-trained 1000-way image classification to measure Web page visual intelligibility for explaining academic concepts. We also apply a finetuning approach for classifying visually intelligible / unintelligible Web pages. We identify factors influencing visual intelligibility, revealing that the Gradient-weighted Class Activation Mapping technique performs better for this task.

## **KEYWORDS**

Web pages explaining academic concepts, deep learning, CNN, explainable AI, visual intelligibility

#### **ACM Reference Format:**

Shintaro Okada, Chiharu Hirohana, Kota Kawaguchi, Kosei Soda, Takehito Utsuro, Yasuhide Kawada, and Noriko Kando. 2019. Identifying Factors of Visual Intelligibility of Web Pages explaining Academic Concepts. In . ACM, New York, NY, USA, 9 pages.

KDD '19, August 04-08, 2019, Anchorage, Alaska - USA

© 2019 Association for Computing Machinery.

## **1 INTRODUCTION**

A search engine is an important tool for acquiring practical knowledge in an academic community. To find beginner-friendly Web pages through a search engine, many pages should be compared. However, an ineffective manual comparison arises because no systematic criterion can measure a beginner friendly Web page. To overcome this problem, we come up with a technique of measuring beginner-friendly Web pages to explain academic concepts and build a system for promoting the academic study and improve Web learning.

Figure 1 shows an example of a beginner-friendly Web page in explaining an academic concept "probability density function" in statistics. The Web page is beginner-friendly because its title page, formula, explanation texts, and figure are visually intelligible. The explanation text is simple to understand, and has corresponding footnotes below each page. Web pages are not beginner-friendly if they violate at least one of the following requirements: (i) having a visually intelligible layout for a page title, formula, explanation text, and its figure, and (ii) having a simple explanation text and a reference for further studies.

Figure 2 shows evidence of the non-existence of systematic criterion for measuring beginner-friendly Web pages that ranked 10th or higher on the Google search engine based on 105 queries of academic terms in seven academic fields: linear algebra, physics, biology, programming, IT, statistics, and chemistry. The figure plots the beginner-friendly Web pages ranked *N*-th or higher (N = 1, ..., 10) that mostly explain academic concepts from the academic term queries. This evidence supports that no systematic criterion exists for measuring beginner-friendly Web pages and explains the academic concepts in the Google search engine.

To find beginner-friendly Web pages in explaining academic concepts, the two requirements explained above ((i) and (ii)) are required. Thus, we focus on (i) Web page visual intelligibility to explain the academic concepts. We apply the fine-tuning technique of VGG16 [16], a pre-trained 1000-way image classification to explain academic concepts.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

KDD '19, August 04-08, 2019, Anchorage, Alaska - USA



Figure 1: Example of a Beginner-Friendly (i.e., the Web page layout is visually intelligible and its text is beginner-friendly) Web Page in Explaining Academic Concepts (excerpt from https://mathtrain.jp/pmitsudo, in Japanese)



Figure 2: Rate of Beginner Friendly (i.e., Web page layout is visually intelligible and its text is beginner friendly) Web Pages in Explaining Academic Concepts ranked 10th or Higher in the Search Engine

Moreover, we identify factors influencing visual intelligibility of Web pages in explaining academic concepts. The motivation for studying the second task is a personal preference for visual intelligibility. In section 3.2, at least around 10 or more factors constitute visual intelligibility, and learners should consider any of the factors based on their preferences. Aside from measuring visual intelligibility on Web pages, another technique to measure visual intelligibility is crucial. This study applies the Gradient-weighted Class Activation Mapping (Grad-CAM) [14] to this task and demonstrates that Grad-CAM performs better in identifying factors influencing Web page visual intelligibility for explaining academic concepts.

# 2 RELATED WORK

No study has explored beginner-friendly Web pages for explaining academic concepts. However, some studies have estimated skills using slides and audios. Luzard et al. [9] applied machine learning to show words, images, tables, and maximum font sizes in slides; the most significant audio-based features are pitch and filled pauses. Approaches to text readability judgment (e.g., Pitler and Nenkova [11]; González-Garduño and Søgaard [5]) are closely related to measuring a beginner-friendly Web page text.

Explainable artificial intelligence (XAI) also identifies factors influencing Web page visual intelligibility for explaining academic concepts. Bojarski et al. [2] proposed a deep neural network to determine the elements in road images that influence steering decision in autonomous driving. In deep neural network based approaches to visual question answering (VQA) [1, 4, 7, 8, 10], techniques to detect regions that represent elementary objects within an image and formalize their relation to linguistic expressions have been well studied. Ribeiro et al. [12] also proposed another XAI framework that explains the predictions of any classifier and applied it to text classification using support vector machines (SVMs) and image classification by deep neural networks.

This study clearly differs from existing approaches on XAI because in our task focuses on quality judgment on visually intelligible Web pages. In the existing approaches, it is limited to identify specific elementary items, such as an image in skiing, ski plates, and poles. In our case, this is much more complicated because we must identify factors, such as the proportion of characters and figures.

# **3 WEB PAGE REFERENCE DATA TO EXPLAIN ACADEMIC CONCEPTS**

#### 3.1 Academic Fields and Concepts of Study

In this study, we collect science and technology (S&T) academic terms used for queries because they have a similar criterion for judging an entire beginner-friendly Web page, visual intelligibility of a Web page layout, and the beginner-friendliness of Web page text. We select seven S&T fields for this study: linear algebra, physics, biology, programming, IT, statistics, and chemistry. For each field, we select 15 terms as queries of academic concepts at the high school or university levels as listed in Table 1. The queries are selected under the criterion that certain number of Web pages ranked 10th or higher on Google search engine explain academic concepts.

# 3.2 Factors of Visual Intelligibility

Table 2 lists the criteria of Web page visual intelligibility that explain the academic concepts for developing a reference data set. Three absolute criteria and five optional criteria are used. For developing a manual reference data set, positive visually intelligible Web page layout satisfies all of the three absolute criteria and at least two optional criteria.

## 3.3 Reference Data Set

For each academic term in section 3.1, we collect the highest 10 Web pages ranked by the Google search engine using the academic term as the query. In collecting Web pages, we ignore Web pages whose HTML files are inaccessible. We also ignore the following Web pages: Wikipedia<sup>1</sup>, Kotobank<sup>2</sup>, Weblio<sup>3</sup>, and Yahoo!

Chiebukuro<sup>4</sup> (Japanese version of Yahoo! Answers<sup>5</sup>), which are online encyclopedias and a question and answer site. These Web pages are ranked high in most of the seven academic fields. However, they are ineffective in explaining academic concepts and visually intelligible Web page layout. The first author of this study<sup>6</sup> judged the visual intelligibility of each collected Web page and explained the concepts in accordance with the criterion discussed in the previous section. Finally, we consider Web pages that satisfy the positive and negative visual intelligibility samples, as shown in Table 1. Of the seven academic fields, we use the Web pages from four academic fields as training samples, one academic field as development samples for stopping the training procedure, and the remaining two academic fields as test samples.

# 4 MEASURING WEB PAGE VISUAL INTELLIGIBILITY TO EXPLAIN ACADEMIC CONCEPTS

#### 4.1 VGG16

Deep learning techniques have been applied to various research fields and have achieved remarkable improvement in the state of the art.

Pattern recognition domains, such as image recognition, convolutional neural networks (CNN), and a large-scale image data set such as ImageNet [13], achieve high performance in various image recognition. Pre-trained CNN parameters using a large-scale general-purpose data set of images (e.g., natural images) are quite useful for extracting universal features for fine-tuned image recognition tasks of specific domains, such as the medical domain [15, 17].

Following the success derived from this fine-tuning process, this study applies the approach to the visual intelligibility automatic judgment of the Web page layout to explain academic concepts. We employ VGG16 CNN model [16] for extracting universal features.

<sup>1</sup> https://ja.wikipedia.org/

<sup>&</sup>lt;sup>2</sup> https://kotobank.jp/

<sup>3</sup> https://ejje.weblio.jp/

<sup>4</sup>https://chiebukuro.yahoo.co.jp/

<sup>&</sup>lt;sup>5</sup> https://answers.yahoo.com

<sup>&</sup>lt;sup>6</sup> In the preliminary study, where two authors of this study worked to develop reference data set and analyzed their agreement rate, the results of the overall beginner-friendly Web pages in explaining academic concepts and the Web page visual intelligibility layout vary according to the annotators' knowledge level and preferences. In developing reference data set, we prefer consistent reference data with only one annotator.

# Table 1: Query Academic Terms (in Japanese) (positive: visually intelligible, negative: visually unintelligible)

	academic	# of			# of positive/negative Web pages		
	fields	fields queries academic terms (excerpts)				total	
	linear algebra	15	階数 (rank), 共役勾配 (conjugate gradient), 行列式 (determinant), クラ メルの公式 (Cramer's rule), クロネッカーのデルタ (Kronecker delta), 三角行列 (triangular matrix)	62	88	150	
	physics	15	電気力線 (line of electric force), 張力 (tension), 慣性の法則 (Newton's first law), 遠心力 (centrifugal force), 電波 (radio wave), 電流 (electric current)	52	98	150	
train.	IT	15	API, DBMS, HTML, IP アドレス (IP address), JDBC, RDB, SDK, SQL, Unicode, URL, スコープマネジメント (scope management), タ イムマネジメント (time management)	101	49	150	
	biology	15	DNA, ショウジョウバエ (Drosophilidae), 原核生物 (Prokaryote), 減数 分裂 (meiosis), 光合成 (photosynthesis), 細胞 (cell), 葉緑体 (Chloro- plast), ミトコンドリア (mitochondrion)	79	71	150	
dev.	progra- mming	15	C 言語 (C), Java, エスケープシーケンス (escape sequence), コマンド ライン引数 (command line arguments), スコープ (scope), フィールド 値 (field value), ポインタ (pointer)	58	92	150	
	statistics	15	事後分布 (posterior distribution), 事前分布 (prior distribution), 正規分 布, F分布, 信頼区間 (confidence interval), 標準偏差, 主成分分析 (principal component analysis)	94	56	150	
tes.	chemistry	15	イオン結合 (ionic bond), エステル (ester), カルボン酸 (carboxylic acid), ケトン (ketone), 化学反応式 (chemical equation), 化学平衡 (chemical equilibrium)	63	87	150	
total		105	—	509	541	1,050	

# Table 2: Criteria on Visual Intelligibility of Web Pages in Explaining Academic Concepts

	а	The proportion of	A moderate proportion of characters and figures are required. The proportion of			
		characters and fig-	characters should neither be too high nor too low as compared with tables and			
		ures	figures.			
absolute	b	Background color	Background color should neither be black nor primary colors to make Web page			
			characters easy to read.			
criteria	criteria c Contrast		The contrast of the character color and the background color should make the			
			Web page characters easy to read. For example, yellow characters with a white			
			background should be avoided.			
	d	Figures	At least one figure to explain the academic terms should be included in the Web			
			page. A table with characters should be avoided as a figure.			
optional	optional e Representative for- A represe		A representative formula should be included in the Web page.			
		mulas				
criteria	eria f Proportion of text		A moderate proportion of formulas compared with characters is required. The			
		and formulas	proportion of formulas should be moderately high as compared with characters.			
	g	Character size	The size of the characters should not be too small.			
	h	Advertisement	Preferably, the Web page should not show advertisement. Even if the Web page			
			includes an advertisement, the smaller it is, the better the visual intelligibility			
			of the Web page. Preferably, the position of the advertisement should not be			
			centered.			

VGG16 model won second prize in the image classification task and first prize in the single-object localization task in the 2014 ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [13]. Its neural net architecture consists of a stack of 13 convolutional layers and 5 intermediate max-pooling layers. This is followed by three fully-connected layers, a third layer to perform 1000-way

ILSVRC classification with 1000 channels (one for each class). The final layer is a soft-max layer. The VGG16 model is a publicly available pre-trained 1000-way ILSVRC classification with 2014 ImageNet data set. Moreover, the pre-trained VGG16 model is widely transferable to other image recognition tasks through fine-tuning. In this study, as one of the available versions of VGG16 model, we employ an available<sup>7</sup> model within the deep learning library Pytorch<sup>8</sup>, an open source machine learning library for Python.

# 4.2 Training Procedure

First, each Web page is transformed into its Web page layout image<sup>9</sup>, and the fine-tuned VGG16 model is applied to judge the visual intelligibility of the Web page layout image. In fine-tuning the VGG16 model, its three fully-connected layers of 1000-way ILSVRC classification and the output layer are replaced with another two fully-connected layers of binary classification (for judging the visual intelligibility of the Web page layout image) and the output layer. In the fine-tuning process, pre-trained parameters of 13 convolutional layers and 5 intermediate max-pooling layers are used as the initial fine-tuning parameter values. All the parameters and the subsequent two fully-connected layers and the output layer are fine-tuned with the reference training data set (i.e., from the four academic fields of linear algebra, physics, IT, and biology, as shown in Table 1). The training procedure is stopped based on the model application to the development data set (i.e., from the academic field of programming, as shown in Table 1).

# 4.3 Evaluation Results

In the evaluation process, we compare the results when the training samples are 10/10, 5/10, and 1/10 of the training data set, as shown in Table 1. Evaluation results are shown in Figures 3 and  $4^{10}$ .

Figure 3 compares the effect of training with and without finetuning. For the training without fine-tuning, pre-trained parameters of 13 convolutional layers and 5 intermediate max-pooling layers are not used as the initial parameter values. However, their initial parameter values are randomly selected. Then, all of the parameters, the subsequent two fully-connected layers, and the output layer are trained with the reference training data set. Fine-tuning has a remarkable effect on training without fine-tuning based on the results presented in Figure 3(a) that reveals visually intelligible Web pages and Figure 3(b) that reveals visually unintelligible Web pages.

Figure 4 compares the recall-precision curves for detecting visually intelligible / unintelligible Web pages. As shown in Figure 4(b), without fine-tuning, detecting visually intelligible Web pages is easier than detecting visually unintelligible Web pages. However, as shown in Figure 4(a), detecting visually intelligible Web pages when the number of training samples is 5/10 and 1/10. When the number is 10/10, the two recall-precision curves are almost comparable. From these results, we can conclude that, when the two conditions (a) have a sufficient number of training samples and (b) apply finetuning, the difficulties of detecting visually intelligible / unintelligible Web pages are comparable. Otherwise, the task of detecting visually intelligible Web pages.

Finally, for all the cases, recall-precision curves are unsaturated when increasing the number of training samples from 1/10 to 10/10. This finding means that preparing a larger number of training samples improves the performance of detecting visually intelligible / unintelligible Web pages.

# 5 IDENTIFYING FACTORS OF VISUAL INTELLIGIBILITY

# 5.1 Grad-CAM

We employ Grad-CAM [14]<sup>11</sup> to identify factors of visual intelligibility for Web pages in explaining academic concepts, whose overview is shown in Figure 5. The Grad-CAM mechanism computes the gradient for each feature map k (k = 1, ..., 512). Then, the gradients are considered neuron importance weights of feature maps, generating weighted feature maps. At Grad-CAM [14], ReLU function is applied to the weighted feature maps, and negative values are removed before Grad-CAM localization (heatmap) is computed. Grad-CAM localization (heatmap) represents where the model has to look to make particular decision. Unlike the original Grad-CAM [14], we compute the Grad-CAM localization (heatmap) of the class of visual intelligibility and unintelligibility. Instead of removing negative values of gradients, we consider negative values of the gradients to indicate a *negative* influence on the visual intelligibility and interpret it as having a *positive* influence on the class of visual unintelligibility. Finally, the Grad-CAM localization (blue and red heatmaps) is computed as shown in Figure 5. The

<sup>&</sup>lt;sup>7</sup> https://github.com/pytorch/vision

<sup>/</sup>blob/master/torchvision/models/vgg.py

<sup>8</sup> https://pytorch.org/

 $<sup>^9</sup>$  The horizontal to vertical ratio of each image is as 2:3, and each image is compressed into the size of 224  $\times$  224.

<sup>&</sup>lt;sup>10</sup> As for the metric for measuring visual intelligibility of Web pages explaining academic concepts, we use recall and precision of detecting visually intelligible / unintelligible Web pages. Let U be the set of the whole testing Web pages and R<sub>+</sub> be the set of reference Web pages manually judged as visually unintelligible. The set of reference Web pages manually judged as visually unintelligible is is obtained as  $R_- = \overline{R_+} = U - R_+$ . Let  $\theta_{Ibd}$  be the lower bound of the output value returned by the sigmoid function of the output layer and  $S_+(\theta_{Ibd}) (\subseteq U)$  be the set of Web pages for each, of which the fine-tuned VGG16 model returns the output value greater than or equal to the lower bound  $\theta_{Ibd}$ :  $S_+(\theta_{Ibd}) = \left\{x \in U | sigmoid(x) > \theta_{Ibd} \right\}$ . Let  $S_-(\theta_{Ibd}) (\subseteq U)$  be the complement set of  $S_+(\theta_{Ibd})$ :  $S_-(\theta_{Ibd}) = \overline{S_+(\theta_{Ibd})} = U - S_+(\theta_{Ibd})$ . With the set  $S_+(\theta_{Ibd})$  satisfying the lower bound  $\theta_{Ibd}$ , recall\_ $(\theta_{Ibd})$  and precision\_ $(\theta_{Ibd})$  against the reference set  $R_+$  are denoted as: recall\_ $(\theta_{Ibd})$  and precision\_ $(\theta_{Ibd})$  against the reference set  $R_-$  are also denoted as: recall\_ $(\theta_{Ibd}) = \frac{|R_{-\Lambda}S_-(\theta_{Ibd})|}{|R_-|}$ , precision\_ $(\theta_{Ibd}) = \frac{|R_{-\Lambda}S_-(\theta_{Ibd})|}{|S_-(\theta_{Ibd})|}$ . In

the evaluation, the lower bound  $\theta_{lbd}$  of the output value returned by the sigmoid function of the output layer is increased from 0.0 to 1.0 with 0.05 increments in between, and the recall-precision curves are plotted and compared.

<sup>&</sup>lt;sup>11</sup> One alternative approach is to employ saliency map [3, 6], where a deep neural network approach is not required, but it can be applied to any framework. However, it performed worse than the Grad-CAM approach. Thus, this paper reports a detailed approach and evaluation results of Grad-CAM only but not the saliency map approach.

KDD '19, August 04-08, 2019, Anchorage, Alaska - USA



Figure 3: Evaluation Results of Measuring Visual Intelligibility (1) (comparing with/without fine-tuning)



Figure 4: Evaluation Results of Measuring Visual Intelligibility (2) (comparing tasks of detecting Web pages of positive/negative classes: positive: visually intelligible Web pages, negative: visually unintelligible Web pages)

Okada, et al.

KDD '19, August 04-08, 2019, Anchorage, Alaska - USA







academic field: <u>statistics</u>, the manual judgment of visual intelligibility: <u>positive</u>, prob. of measuring visual intelligibility: <u>0.99</u>, (a) rank of evaluation result: <u>4</u>







academic field: <u>statistics</u>, the manual judgment of visual intelligibility: <u>negative</u>, prob. of measuring visual intelligibility: <u>0.01</u>, (b) rank of evaluation result: <u>3</u>



academic field: chemistry, the manual judgment of visual intelligibility: negative, prob. of measuring visual intelligibility: 0.33, (d) rank of evaluation result: 1



KDD '19, August 04-08, 2019, Anchorage, Alaska - USA

 Table 3: Manual Evaluation Criterion of Identifying Factors of

 Visual Intelligibility

rank	description					
4	Within the input image, more than or equal					
	to 80% are overlapped between the region a					
	tomatically identified by the proposed model					
	(Grad-CAM) and the manually identified ref-					
	erence region. Moreover, the polarities of visu-					
	ally intelligible / unintelligible agree. (For the					
	remaining less than 20% overlapped region, it					
	does not matter even if their polarities disagree.)					
3	Within the input image, 60–80% are overlapped					
	between the region automatically identified by					
	the proposed model (Grad-CAM) and the man-					
	ually identified reference region. Moreover, the					
	polarities of visually intelligible / unintelligible					
	agree. (For the remaining less than 40% over-					
	happed region, it does not matter even if their					
	polarities disagree.)					
2	within the input image, more than or equal to 20% are overlapped between the region over					
	to 50% are overlapped between the region au-					
	(Grad-CAM) and the manually identified refer-					
	ence region. It does not matter if their polarities					
	of visually intelligible / unintelligible disagree					
	(The region where their polarities agree should					
	be less than 60%.)					
1	Within the input image, less than 30% are over-					
	lapped between the region automatically iden-					
	tified by the proposed model (Grad-CAM) and					
	the manually identified reference region.					

blue region has a positive influence on the class of visual intelligibility, whereas the red region has a negative influence on the class of visual intelligibility. Thus, a positive influence on the class of visual unintelligibility. The model looks at the blue and red regions to make decisions on visual intelligibility / unintelligibility.

Compared with the saliency map approach [3, 6], the advantages of Grad-CAM are (i) localization heatmap computation is efficient, and (ii) broad region within the input image can be detected, for instance, the background region and blank region have an influence on a particular class. Its disadvantage depends on the architecture of the model, such as the neural network structure. Once the architecture of the model is revised, the Grad-CAM formalization is revised in accordance with the revision of the model architecture.

### 5.2 Evaluation

We manually examine the overlap between the region automatically identified by the proposed model (Grad-CAM) and the manually identified reference region to evaluate whether their polarities of visually intelligible / unintelligible agree / disagree. Then, the manual evaluation result is graded into four ranks in accordance with the evaluation criterion shown in Table 3. Table 4 shows the detailed evaluation result, where, for each of the 10 ranges measuring

 
 Table 4: Evaluation Results of Identifying Factors of Visual Intelligibility

prob. of	ranks of manual evaluation				
measuring visual intelligibility	4 (%)	3 (%)	2 (%)	1 (%)	total
0.0 - 0.1	11 (8.5)	55 (42.6)	56 (43.4)	7 (5.4)	129
0.1 - 0.2	0 (0.0)	6 (60.0)	3 (30.0)	1 (10.0)	10
0.2 - 0.3	0 (0.0)	2 (28.6)	5 (71.4)	0 (0.0)	7
0.3 - 0.4	0 (0.0)	1 (33.3)	1 (33.3)	1 (33.3)	3
0.4 - 0.5	1 (11.1)	3 (33.3)	5 (55.6)	0 (0.0)	9
0.5 - 0.6	0 (0.0)	0 (0.0)	3 (75.0)	1 (25.0)	4
0.6 - 0.7	0 (0.0)	3 (37.5)	5 (62.5)	0 (0.0)	8
0.7 - 0.8	0 (0.0)	2 (25.0)	6 (75.0)	0 (0.0)	8
0.8 - 0.9	0 (0.0)	6 (85.7)	1 (14.3)	0 (0.0)	7
0.9 - 1.0	4 (3.5)	41 (35.7)	61 (53.0)	9 (7.8)	115
total	16 (5.3)	119 (39.7)	146 (48.7)	19 (6.3)	300

visual intelligibility by VGG16, the proportion of the four ranks (4, 3, 2, and 1) is shown. Based on the results, most test samples are ranked 2 or above within the ranges of 0.2–0.9 for measuring visual intelligibility. Overall, the rate of ranks 3 or 4 is approximately 45%, which is higher than that of the saliency map approach [3, 6] (approximately 15% in our implementation and evaluation).

Figure 6 shows the examples of Grad-CAM that identify factors of visual intelligibility, where the identified factors are shown as the blue and red heatmaps. Figures 6(a) - 6(d) show the factors ranked as 4-1, respectively. For each case, along with the manually judged visual intelligibility as positive or negative, the probability of measuring visual intelligibility by VGG16 is shown. As shown in Figures 6(a) and 6(b), when the rank is 3 or higher, text regions whose size is relatively larger than a certain upper bound and advertisement region are identified as visually unintelligible and painted with red in the heatmap. Similarly, regions of figures and menu bars are identified as visually intelligible and painted with blue in the heatmap. In the case of Figure 6(c) where the rank is 2, the regions identified by the Grad-CAM overlap with the reference regions, whereas their polarities are opposite. Specifically, regions of figures and menu bars are identified as visually unintelligible and painted with red in the heatmap, as opposed with the reference polarity. Furthermore, in the case of rank 1 as in Figure 6(d), the regions identified by the Grad-CAM are not overlapped with the reference regions. Specifically, the regions identified as visually intelligible and painted with blue in the heatmap are mostly not overlapped with the reference regions.

## 6 CONCLUDING REMARKS

Based on the motivation to find beginner-friendly Web pages in explaining academic concepts, this study focused on visual intelligibility of Web pages. Future work includes (i) applying a general XAI framework by Ribeiro et al. [12] to the task of identifying the factors of Web page visual intelligibility in explaining academic concepts, (ii) inventing a technique for the automatic induction in analyzing and classifying the identified factors of visual intelligibility and (iii) evaluating the proposed framework through a behavioral study to predict actual learning gains of human students.

#### REFERENCES

- J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. 2016. Learning to Compose Neural Networks for Question Answering. In Proc. NAACL-HLT. 1545–1554.
- [2] M. Bojarski, P. Yeres, A. Choromanska, K. Choromanski, B. Firner, L. D. Jackel, and U. Muller. 2017. Explaining How a Deep Neural Network Trained with Endto-End Learning Steers a Car. *CoRR* (2017).
- [3] R. C. Fong and A. Vedaldi. 2017. Interpretable Explanations of Black Boxes by Meaningful Perturbation. In Proc. ICCV. 3429–3437.
- [4] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. 2016. Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding. In Proc. EMNLP. 457–468.
- [5] A. V. González-Garduño and A. Søgaard. 2017. Using Gaze to Predict Text Readability. In Proc. 12th BEA. 438–443.
- [6] S. Greydanus, A. Koul, J. Dodge, and A. Fern. 2018. Visualizing and Understanding Atari Agents. In Proc. 35th ICML. 1792–1801.
- [7] R. Hu, J. Andreas, T. Darrell, and K. Saenko. 2018. Explainable Neural Computation via Stack Neural Module Networks. In Proc. ECCV.
- [8] R. Hu, M. Rohrbach, J. Andreas, T. Darrell, and K. Saenko. 2017. Modeling Relationships in Referential Expressions With Compositional Modular Networks. In Proc. CVPR. 1115–1124.
- [9] G. Luzard, B. Guamán, K. Chiluiza, G. Castells, and X. Ochoa. 2014. Estimation of Presentations Skills Based on Slides and Audio Features. In *Proc. MLA*. 37– 44.

- [10] D. H. Park, L. A. Hendricks, Z. Akata, A. Rohrbach, B. Schiele, T. Darrell, and M. Rohrbach. 2018. Multimodal Explanations: Justifying Decisions and Pointing to the Evidence. In *Proc. CVPR*, 8779–8788.
- [11] E. Pitler and A. Nenkova. 2004. Revisiting Readability: A Unified Framework for Predicting Text Quality. In *Proc. EMNLP*. 186–195.
  [12] M. T. Ribeiro, S Singh, and C. Guestrin. 2016. "Why Should I Trust You?": Ex-
- [12] M. T. Ribeiro, S Singh, and C. Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proc. 22nd SIGKDD*. 1135–1144.
- [13] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and F.-F. Li. 2014. ImageNet Large Scale Visual Recognition Challenge. *CoRR* abs/1409.0575 (2014).
- [14] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. 2017. Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization. In *Proc. ICCV*. 618–626.
- [15] H.-C. Shin, H. R. Roth, M. Gao, L., Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers. 2016. Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. *IEEE Transactions on Medical Imaging* 35, 5 (2016), 1285– 1298.
- [16] K. Simonyan and A. Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proc. 3rd ICLR.
- [17] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang. 2016. Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning? *IEEE Transactions on Medical Imaging* 35, 5 (2016), 1299–1312.