SlideImages: A Dataset for Educational Image Classification

David Morris[1](✉), Eric Müller-Budack[1], and Ralph Ewerth[1,2]

[1] TIB – Leibniz Information Centre for Science and Technology, Hannover, Germany
{David.Morris,Eric.Mueller,Ralph.Ewerth}@tib.eu
[2] L3S Research Center, Leibniz Universität Hannover, Hannover, Germany

Abstract. In the past few years, convolutional neural networks (CNNs) have achieved impressive results in computer vision tasks, which however mainly focus on photos with natural scene content. Besides, non-sensor derived images such as illustrations, data visualizations, figures, etc. are typically used to convey complex information or to explore large datasets. However, this kind of images has received little attention in computer vision. CNNs and similar techniques use large volumes of training data. Currently, many document analysis systems are trained in part on scene images due to the lack of large datasets of educational image data. In this paper, we address this issue and present SlideImages, a dataset for the task of classifying educational illustrations. SlideImages contains training data collected from various sources, e.g., Wikimedia Commons and the AI2D dataset, and test data collected from educational slides. We have reserved all the actual educational images as a test dataset in order to ensure that the approaches using this dataset generalize well to new educational images, and potentially other domains. Furthermore, we present a baseline system using a standard deep neural architecture and discuss dealing with the challenge of limited training data.

Keywords: Document figure classification · Educational documents · Classification dataset

1 Introduction

Convolutional neural networks (CNNs) are making great strides in computer vision, driven by large datasets of annotated photos, such as ImageNet [1]. Many images relevant for information retrieval, such as charts, tables, and diagrams, are created with software rather than through photography or scanning.

There are several applications in information retrieval for a robust classifier of educational illustrations. Search tools might directly expose filters by predicted label, natural language systems could choose images by type based on what information a user is seeking. Further analysis systems could be used to extract more information from an image to be indexed based on its class. In this case, we have classes such as pie charts and x-y graphs that indicate what type of

J. M. Jose et al. (Eds.): ECIR 2020, LNCS 12036, pp. 289–296, 2020.
https://doi.org/10.1007/978-3-030-45442-5_36

information is in the image (e.g., proportions, or the relationship of two numbers) and how it is symbolized (e.g., angular size, position along axes).

Most educational images are created with software and are qualitatively different from photos and scans. Neural networks designed and trained to make sense of the noise and spatial relationships in photos are sometimes suboptimal for born-digital images and educational images in general.

Educational images and illustrations are under-served in training datasets and challenges. Competitions such as the Contest on Robust Reading for Multi-Type Web Images [2] and ICDAR DeTEXT [3] have shown that these tasks are difficult and unsolved. Research on text extraction such as Morris et al. [4] and Nayef and Ogier [5] has shown that even noiseless born-digital images are sometimes better analyzed with neural nets than with handcrafted features and heuristics. Born-digital and educational images need further benchmarks on challenging information retrieval tasks in order to test generalization.

In this paper, we introduce SlideImages, a dataset which targets images from educational presentations. Most of these educational illustrations are created with diverse software, so the same symbols are drawn in different ways in different parts of the image. As a result, we expect that effective synthetic datasets will be hard to create, and methods effective on SlideImages will generalize well to other tasks with similar symbols. SlideImages contains eight classes of image types (e.g. bar charts and x-y plots) and a class for photos. The labels we have created were made with information extraction for image summarization in mind.

In the rest of this paper, we discuss related work in Sect. 2, details about our dataset and baseline method in Sect. 3, results of our baseline method in Sect. 4, and conclude with a discussion of potential future developments in Sect. 5.

2 Related Work

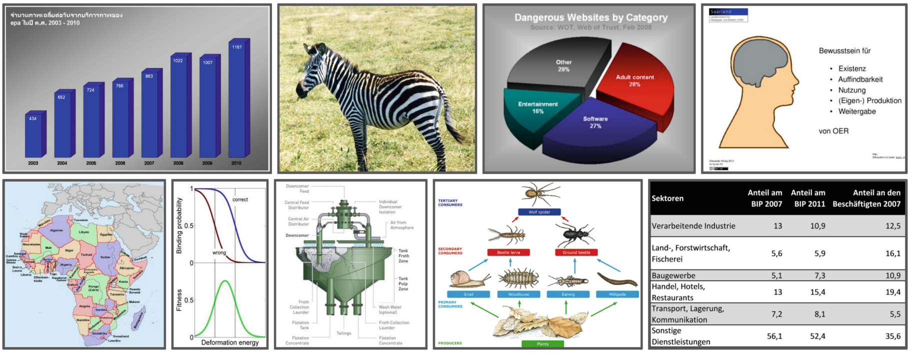
Prior information retrieval publications used or could use document figure classification. Charbonnier et al. [6] built a search engine with image type filters. Aletras and Mittal [7] automatically label topics in photos. Kembhavi et al.'s [8] diagram analysis assumes the input figure is a diagram. Hiippala and Orekhova extended that dataset by annotating it in terms of Relational Structure Theory, which implies that the same visual features communicate the same semantic relationships. De Herrera et al. [9] seek to classify image types to filter their search for medical professionals.

We intend to use document figure classification as a first step in automatic educational image summarization applications. A similar idea is followed by Morash et al. [10], who built one template for each type of image, then manually classified images and filled out the templates, and suggested automating the steps of that process. Moraes et al. [11] mentioned the same idea for their SIGHT (Summarizing Information GrapHics Textually) system.

A number of publications on document image classification such as Afzal et al. [12] and Harley et al. [13] use the RVL-CDIP (Ryerson Vision Lab Complex Document Information Processing) dataset, which covers scanned documents.

Table 1. Comparison of different datasets including the number of classes and images.

	Classes	Images		
		Train	Val	Test
SlideImages	9	2646	292	691
DocFigure	28	19795	0	13172
Head-to-head SlideImages	8	2331	257	575
Head-to-head DocFigure	8	11678	3886	3891



Sektoren	Anteil am BIP 2007	Anteil am BIP 2011	Anteil an den Beschäftigten 2007
Verarbeitende Industrie	13	10,9	12,5
Land-, Forstwirtschaft, Fischerei	5,6	5,9	16,1
Baugewerbe	5,1	7,3	10,9
Handel, Hotels, Restaurants	13	15,4	19,4
Transport, Lagerung, Kommunikation	7,2	8,1	5,5
Sonstige Dienstleistungen	56,1	52,4	35,6

Fig. 1. Train set class examples clockwise from top left: bar charts, photos, pie charts, slide images, tables, structured diagrams, technical drawings, x-y plots, and maps.

While document scans and born-digital educational illustrations have materially different appearance, these papers show that the utility of deep neural networks is not limited to scene image tasks (Fig. 1).

A classification dataset of scientific illustrations was created for the NOA project [14]. However, their dataset is not publicly available, and does not draw as many distinctions between types of educational illustrations. Jobin et al.'s DocFigure [15] consists of 28 different categories of illustrations extracted from scientific publications totaling 33,000 images.

3 Dataset and Baseline System

Techniques that work well on DocFigure [15] do not generalize to the educational illustrations in our use case scenarios (as we also show in Sect. 4.2). Different intended uses or software cause sufficient differences in illustrations that a dataset of specifically educational illustrations is needed.

CNNs and related techniques are heavily data driven. An approach must consist of both an architecture and optimization technique, but also the data used for that optimization. In our case, we consider the dataset our main contribution.

3.1 SlideImages Dataset

When building our taxonomy, we have chosen classes such that one class would have the same types of salient features, and appropriate summaries would also be similar in structure. Our classes are also all common in educational materials. Beyond the requirements of our taxonomy, our datasets needed to be representative of common educational illustrations in order to fit real-world applications, and legally shareable to promote research on educational image classification. Educational illustrations are created by a variety of communities with varying expertise, techniques, and tools, so choosing a dataset from one source may eliminate certain variables in educational illustration. To identify these variables, we kept our training and test data sources separate.

We assembled training and validation datasets from various sources of open access illustrations. Bar charts, x-y plots, maps, photos, pie charts, slide images, table images, and technical drawings were manually selected by a student assistant (supported by the main author) using the Wikimedia Commons image search for related terms. We manually selected graph diagrams, which we also call node-edge diagrams or "structured diagrams," from the Kembhavi et al. [8] AllenAI Diagram Understanding (AI2D) dataset; not all AI2D images contain graph edges [8]. The training dataset of SlideImages consists of 2,938 images and is intended for fine-tuning CNNs, not training from scratch. The SlideImages test set is derived from a snapshot of SlideWiki open educational resource platform (https://slidewiki.org/) datastore obtained in 2018. From that snapshot, two annotators manually selected and labeled 691 images. Our data are available at our code repository: https://github.com/david-morris/SlideImages/.

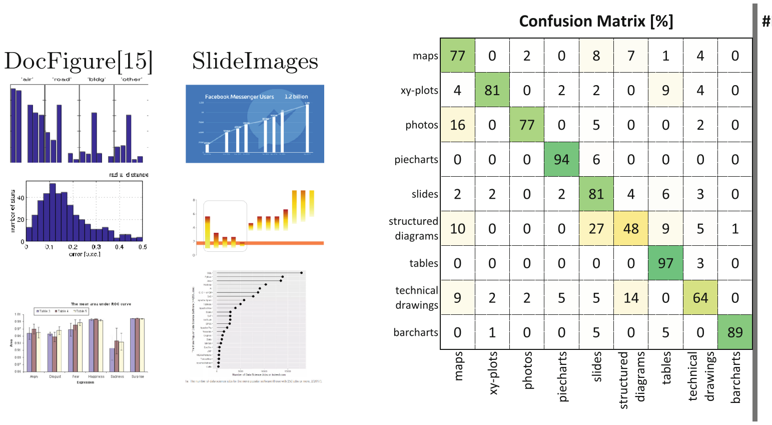
3.2 Baseline Approach

The SlideImages training dataset is small compared to datasets like ImageNet [1], with over 14 million images, RVL-CDIP [13] with 400,000 images, or even DocFigure [15] with 33,000 images. Much of our methodology is shaped by needing to confront the challenges of a small dataset. In particular, we aim to avoid overfitting: the tendency of a classifier to identify individual images and patterns specific to the training set rather than the desired semantic concepts.

For our pre-training dataset, a large, diverse dataset is required that contains a large proportion of educational and scholarly images. We pre-trained on a dataset of almost 60,000 images labeled by Sohmen et al. [6] (NOA dataset), provided by the authors on request. The images are categorized as composite images, diagrams, medical imaging, photos, or visualizations/models.

To mitigate overfitting, we used data augmentation: distorting an image while keeping relevant traits. We used image stretching, brightness scaling, zooming, and color channel shifting as shown in our source code. We also added dropout with a rate of 0.1 on the extracted features before the fully connected and output layers. We used similar image augmentation for pre-training and training.

We use MobileNetV2 [16] as our network architecture. We chose MobileNetV2 as a compromise between a small number of parameters and performance on ImageNet. Intuitively, a smaller parameter space implies a model with more bias and



	maps	xy-plots	photos	piecharts	slides	structured diagrams	tables	technical drawings	barcharts	#Img
maps	77	0	2	0	8	7	1	4	0	96
xy-plots	4	81	0	2	2	0	9	4	0	57
photos	16	0	77	0	5	0	0	2	0	130
piecharts	0	0	0	94	6	0	0	0	0	18
slides	2	2	0	2	81	4	6	3	0	116
structured diagrams	10	0	0	0	27	48	9	5	1	79
tables	0	0	0	0	0	0	97	3	0	66
technical drawings	9	2	2	5	5	14	0	64	0	44
barcharts	0	1	0	0	5	0	5	0	89	85

Fig. 2. Left: examples of bar charts from the DocFigure [15] train set and our own test set. Right: confusion matrix of our baseline system on SlideImages. Entries show percent of true members of the class on the left margin labeled as on the bottom margin. Weighted accuracy average is 80% over all 691 images.

lower variance, which is better for smaller datasets. We initialized our weights from an ImageNet model and pre-trained for a further 40 epochs with early stopping on the NOA dataset using the Adam (adaptive moment estimation) [17] optimizer. This additional pre-training was intended to cause the lower levels of the network to extract more features specific to born-digital images. We then trained for 40 epochs with Adam and a learning rate schedule. Our schedule drops the learning rate by a factor of 10 at the 15th and 30th epoch. Our implementation is available at https://github.com/david-morris/SlideImages/.

4 Preliminary Results

We have performed two experiments, in order to show that this dataset represents a meaningful improvement over existing work, and to establish a baseline. Because our classes are unbalanced, we have reported summary statistics as accuracy averages of each class weighted by number of instances per class.

4.1 Baseline

We set a baseline for our dataset with the classifier described in Sect. 3.2. The confusion matrix in Fig. 2 shows that misclassifications do tend towards a few types of errors, but none of the classes have collapsed. While certain classes are likely to be misclassified as another specific class (such as structured diagrams as slides), those relationships do not happen in reverse, and a correct classification is more likely. Figure 2 shows that our baseline leaves room for improvement, and our test set helps to identify challenges in this task. Viewing individual classification errors highlighted a few problems with our training data. Our training

Table 2. Head-to-head comparison of accuracy (weighted averages).

	SlideImages train	DocFigure train	DocFigure baseline
SlideImages test	80%	78%	75%
DocFigure test	92%	99%	99%

data do not include sufficient structured diagrams with illustrated arrows, or edges which travel only at 90° increments, such as organigrams or some Unified Modeling Language diagrams. Our photos do not include examples with the background removed, but these are common in educational images. These problems should be remedied in future training datasets for this and similar problems.

4.2 Head-to-Head Comparison

The related DocFigure dataset covers similar images and has much more data than SlideImages. To justify SlideImages, we have created a head-to-head comparison of classifiers trained in the same way (as described in Sect. 3.2) on the SlideImages and DocFigure datasets. All the SlideImages classes except *slides* have an equivalent in DocFigure. We have shown the reduction in the data used, and the relative sizes of the datasets, in Table 1. The Head-to-head datasets contain only the matching classes, and in the case of the DocFigure dataset, the original test set has been split into validation and test sets.

After obtaining the two trained networks, we have tested each network on both the matching test set, and the other test set. Although we were unable to reproduce the VGG-V baseline used by Jobin et al., we used a linear SVM with VGG-16 features and achieved comparable results on the full DocFigure dataset (90% macro average compared to their 88.96% with a fully neural feature extractor). The results (Table 2) show that SlideImages is a more challenging and potentially more general task. The net trained on SlideImages did even better on the DocFigure test set than on the SlideImages test set. Despite having a different source and approximately a fifth of the size of the DocFigure dataset, the net trained on SlideImages training set was better on our test set.

5 Conclusions and Future Work

In this paper, we have presented the task of classifying educational illustrations and images in slides and introduced a novel dataset SlideImages. The classification remains an open problem despite our baseline and represents a useful task for information retrieval. We have provided a test set derived from actual educational illustrations, and a training set compiled from open access images. Finally, we have established a baseline system for the classification task. Other potential avenues for future research include experimenting with the DocFigure dataset in the pre-training and training phases, and experimenting with text extraction for multimodal classification.

Acknowledgement. This work is financially supported by the German Federal Ministry of Education and Research (BMBF) and European Social Fund (ESF) (InclusiveOCW project, no. 01PE17004).

References

1. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: CVPR09 (2009)
2. He, M., et al.: ICPR2018 contest on robust reading for multi-type web images. In: 24th International Conference on Pattern Recognition. ICPR 2018, Beijing, China, 20–24 August 2018, pp. 7–12. IEEE Computer Society (2018)
3. Yang, C., Yin, X., Yu, H., Karatzas, D., Cao, Y.: ICDAR2017 robust reading challenge on text extraction from biomedical literature figures (detext). In: 14th IAPR International Conference on Document Analysis and Recognition. ICDAR 2017, Kyoto, Japan, 9–15 November 2017, pp. 1444–1447 (2017)
4. Morris, D.. Tang, P., Ewerth, R.: A neural approach for text extraction from scholarly figures. In: 15th International Conference on Document Analysis and Recognition. ICDAR 2019, Sydney, Australia, 20–25 September 2019, pp. 1438–1443 (2019, to appear)
5. Nayef, N., Ogier, J.: Semantic text detection in born-digital images via fully convolutional networks. In: 14th IAPR International Conference on Document Analysis and Recognition. ICDAR 2017, Kyoto, Japan, 9–15 November 2017, pp. 859–864 (2017)
6. Charbonnier, J., Sohmen, L., Rothman, J., Rohden, B., Wartena, C.: NOA: a search engine for reusable scientific images beyond the life sciences. In: Pasi, G., Piwowarski, B., Azzopardi, L., Hanbury, A. (eds.) ECIR 2018. LNCS, vol. 10772, pp. 797–800. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-76941-7_78
7. Aletras, N., Mittal, A.: Labeling Topics with Images Using a Neural Network. In: Jose, J.M., et al. (eds.) ECIR 2017. LNCS, vol. 10193, pp. 500–505. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-56608-5_40
8. Kembhavi, A., Salvato, M., Kolve, E., Seo, M., Hajishirzi, H., Farhadi, A.: A diagram is worth a dozen images. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9908, pp. 235–251. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_15
9. García Seco de Herrera, A., Markonis, D., Joyseeree, R., Schaer, R., Foncubierta-Rodríguez, A., Müller, H.: Semi–supervised learning for image modality classification. In: Müller, H., Jimenez del Toro, O.A., Hanbury, A., Langs, G., Foncubierta Rodríguez, A. (eds.) Multimodal Retrieval in the Medical Domain. LNCS, vol. 9059, pp. 85–98. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24471-6_8
10. Morash, V.S., Siu, Y., Miele, J.A., Hasty, L., Landau, S.: Guiding novice web workers in making image descriptions using templates. TACCESS **7**(4), 12:1–12:21 (2015)
11. Moraes, P.S., Sina, G., McCoy, K.F., Carberry, S.: Evaluating the accessibility of line graphs through textual summaries for visually impaired users. In: Kurniawan, S., Richards, J. (eds.) Proceedings of the 16th International ACM SIGACCESS Conference on Computers & Accessibility. ASSETS 2014, Rochester, NY, USA, 20–22 October 2014, pp. 83–90. ACM (2014)

12. Afzal, M.Z., Kölsch, A., Ahmed, S., Liwicki, M.: Cutting the error by half: investigation of very deep CNN and advanced training strategies for document image classification. In: 14th IAPR International Conference on Document Analysis and Recognition. ICDAR 2017, Kyoto, Japan, 9–15 November 2017, pp. 883–888 (2017)
13. Harley, A.W., Ufkes, A., Derpanis, K.G.: Evaluation of deep convolutional nets for document image classification and retrieval. In: 13th International Conference on Document Analysis and Recognition. ICDAR 2015, Nancy, France, 23–26 August 2015, pp. 991–995. IEEE Computer Society (2015)
14. Sohmen, L., Charbonnier, J., Blümel, I., Wartena, C., Heller, L.: Figures in scientific open access publications. In: Méndez, E., Crestani, F., Ribeiro, C., David, G., Lopes, J.C. (eds.) TPDL 2018. LNCS, vol. 11057, pp. 220–226. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00066-0_19
15. Jobin, K.V., Mondal, A., Jawahar, C.V.: DocFigure: a dataset for scientific document figure classification. In: 13th IAPR International Workshop on Graphics Recognition. GREC 2019, Sydney, Australia, 20–22 September 2019 (2019, to appear)
16. Sandler, M., Howard, A.G., Zhu, M., Zhmoginov, A., Chen, L.: MobileNetV2: inverted residuals and linear bottlenecks. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018, pp. 4510–4520. IEEE Computer Society (2018)
17. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations. ICLR 2015, San Diego, CA, USA, 7–9 May 2015, Conference Track Proceedings (2015)