LAYOUT-AWARE SUBFIGURE DECOMPOSITION FOR COMPLEX FIGURES IN THE BIOMEDICAL LITERATURE

Xiangyang Shi, Yue Wu, Huaigu Cao, Gully Burns, Prem Natarajan

Information Sciences Institute, Marina del Rey, California, U.S.A.

ABSTRACT

Published scientific figure is a valuable information resource, but often occur as composite images. The ImageCLEF meeting presented a shared evaluation in 2016 to use machine learning to split these composite figures into components automatically. We adapted an existing high-performance object detection method to analyze the substructure of published biomedical figures by developing a novel multi-branch output convolution neural network to predict irregular panel layouts and provide augmented training data to drive learning. Our system has an accuracy of 86.8% on the 2016 ImageCLEF Medical dataset and 83.1% on a new dataset derived from open access papers from the INTACT database of molecular interactions.

Index Terms— subfigure decomposition, convolutional neural network, biomedical data, compound figures

1. INTRODUCTION

Influential search and recommendation engines such as PubMed/PMC, CiteSeer, Semantic Scholar, ScienceDirect, and Meta provide search and recommendation services to scientists by helping them find papers of interest through their online platforms. These systems run information extraction processes over their indexed content, providing access to the figures of scientific papers as part of this functionality. It has been estimated that over 30% of published scientific figures consist of several sub-figures [1], suggesting that computational methods of separating subfigures could immediately improve functionality of these important tools. Subfigures also act as elements in the discourse structure of scientific papers, being cited as evidence to support the argument being made [2]. These subfigures are typically denoted by letters. Crucially, named subfigures may often have multiple subfigures themselves, and existing separation methods will attempt to split figures to the most fine-grained resolution, even though the most semantically relevant delineation may occur at an intermediate-level grouping. In Figure 1, we illustrate how the output of a state-of-the-art figure delineation system [3] misses this intermediate-level substructure.

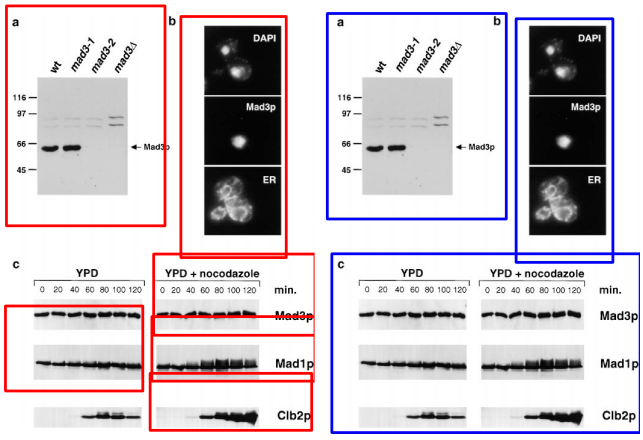

Fig. 1. An example taken from figure 3 of [4] of subfigure decomposition at fine and intermediate granularity. Outlines are generated by the baseline (red) and our experimental systems (blue) presented in this paper.

Existing state-of-art machine vision systems apply deep learning to natural images in order to recognize objects [5, 6]. Rather than only apply these methods to the subfigure delineation challenge, we have extended existing architectures to include a prediction step for the grid-like structure of subfigure panels. This permits our system to learn layouts and predict how these composite images will be constructed.

Biomedical papers use a multitude of charts, images, and diagrams in complex figures that overlap and intermingle without any strictly enforced formatting guidelines. The domain is large, and semantically complex, providing a rich variety of different types of images [7] and has been used to drive subfigure separation shared challenge tasks. Notably, we have analyzed and processed images from open access papers indexed within the European Bioinformatics Institute's (EBI) INTACT project [8] to supplement this data.

This work was supported by grant LM012592 from the NIH's National Library of Medicine. This work is based on research sponsored by the Defense Advanced Research Projects Agency under agreement number FA8750-16-2-0204. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Defense Advanced Research Projects Agency or the U.S. Government.

The contributions of this paper is as follows: (A) We developed a layout-aware neural network for compositing subpanels. To create this network, we made a novel generalization to the You Only Look Once (YOLO) object detector and obtained significant improvement in detection accuracy. (B) We cleaned and annotated 15,436 compound figures from the INTACT and ImageCLEF 2016 dataset.

2. PRIOR WORK

Biomedical images are already the subject of extensive analysis with machine vision methods. The Yale Imagefinder application was an early figure-based search engine built within the bioinformatics community [9], providing valuable statistics about the distribution of different types of figures over the literature [10]. ImageCLEF is the image retrieval competition task of the 'Conference and Labs of the Evaluation Forum' (CLEF). Our chosen specific task, subfigure decomposition within the biomedical literature was the subject of a shared challenge at the ImageCLEF meeting in 2016 [1], providing training data that we use in this paper. This dataset is the first biggest compound image collection from publications with annotations of bounding boxes of subfigures, comprising over 7,000 applicable cases.

Lee. et al. (2015) [11] and Taschwer et al. [12] identified sub-elements of composite scientific figures by identifying large regions of background color or detecting lines as boundaries of subfigures. Li. et al 2016 [13] developed methods that clustered connecting component to attempt to accomplish the same goal. Other work [14] also split the compound figures according to the capital index of each subfigures. These 'traditional' computer vision methods often assume that subfigures must be separated by wide boundaries or be denoted by capital indices. Frequently, such assumptions fail: subfigures may be arranged in reverse order; subfigures may not be clearly separable by an x-y cut; subfigures may overlap. More recently, Convolution Neural Network (CNN) systems are emerging as the preferred methodology to analyze scientific figures.

You Only Look Once (YOLO) is a CNN based objective detection network[6]. Before YOLO, visual object recognitions systems would process a large number of candidate bounding boxes and then learn whether they corresponded to objects using a CNN [15, 16]. This methodology was slow, based on repeatedly performing the CNN forward computation. YOLO processes the image once and then treats bounding box prediction as a regression problem over the image's feature map. YOLO v2 [17] is an advanced version of YOLO. It removes fully connection layer and adds a passthrough layer for the network to be sensitive to small objects. Batch normalization layers are introduced for better convergence. YOLO v2 is also fed with inputs of different sizes after a few epochs to make the network robust to work on images of different resolution. Tsutusi et al. (2017) used YOLO v2 to split scientific figures [3] with transfer learning [18] trained on ImageNet data [19]. After that, they synthesized data by pasting subfigures with random widths into an empty image frame either at random or onto a grid-based template. We adopted a similar grid-based method for generating training data.

3. MATERIALS AND METHODS

The experimental work in this study is based on comparing our 'Layout Aware Decomposition Network' (LADN) methodology with a state-of-the-art baseline system [3].

3.1. Baseline Methodology: Tsutusi & Crandall 2017

Tsutusi et al. 2017 [3] used YOLO v2 by pretraining on ImageNet and training of artificially-constructed datasets [3]. The authors provided access to source code and their trained model at https://bit.ly/2D8bDkC permitting us to reuse their trained model as a baseline. The loss function they optimize is as follows:

$$
\begin{aligned}
L_i = \lambda_{coord} \sum_{j=1}^{B} \mathbb{1}_{ij}^{\text{obj}} \left[(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 \right] \\
+ \lambda_{coord} \sum_{j=1}^{B} \mathbb{1}_{ij}^{\text{obj}} \left[\left(\sqrt{w_i} - \sqrt{\hat{w}_i} \right)^2 + \left(\sqrt{h_i} - \sqrt{\hat{h}_i} \right)^2 \right] \quad (1) \\
+ \sum_{j=1}^{B} \mathbb{1}_{ij}^{\text{obj}} \left(C_i - \hat{C}_i \right)^2 + \lambda_{\text{noobj}} \sum_{j=1}^{B} \mathbb{1}_{ij}^{\text{noobj}} \left(C_i - \hat{C}_i \right)^2
\end{aligned}
$$

Where x_i, y_i are predicted coordinates, w_i, h_i are the predicted width and height of bounding box, C_i is the confidence score of how sure a cell is about finding an object. Since the task does not involve classification, we ignore any loss penalty for misclassification in traditional YOLO loss function. $\hat{x}_i, \hat{y}_i, \hat{w}_i, \hat{h}_i, \hat{C}_i$ comprise the overall target of the regression task. $\mathbb{1}_i^{\text{noobj}}$ denotes whether the object appears in cell i, and $\mathbb{1}_{ij}^{\text{noobj}}$ denotes that the j^{th} bounding box predictor in cell i is 'responsible' for the prediction. $\lambda_{coord}, \lambda_{noobj}$ are tunable parameters. λ_{coord} is set to penalize coordinate errors. Since most cells do not contain objects, there is a tendency to push all confidence scores to 0, which may lead to zero foundings in the network. This problem can be addressed by decreasing loss from cells that do not contain objects. The total loss is the sum of all losses over all cells:

$$
Loss = \sum_{i=0}^{S^2} L_i \quad (2)
$$

3.2. Layout Aware Decomposition Network (LADN)

In comparison to natural images, scientific figures usually have fewer, larger components. Since YOLO is designed

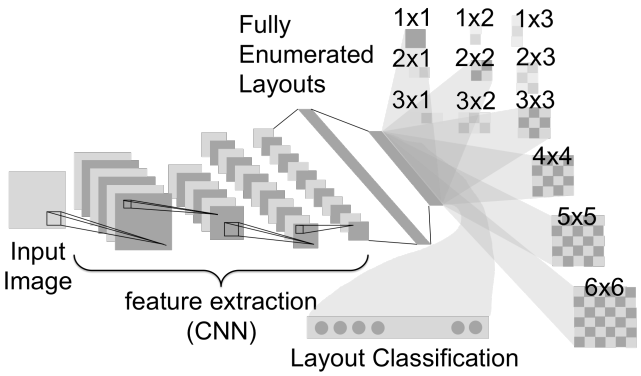

Fig. 2. Neural network overview of our 'Layout Aware Decomposition Network'.

for natural images, it usually delineates the smallest possible atomic subelements. Our solution is to design a network that explicitly selects a scale for its output delineation.

We define the concept of 'rough layout' as a two dimensional measure, where the first element refers to maximum number of subfigures a vertical line can go through in the figure and the second element is the maximum based on a horizontal line. Rough layout can be used for estimating the maximum number of target subfigures directly.

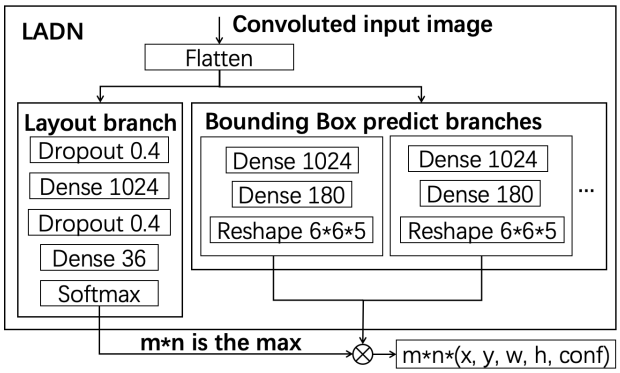

Fig. 3. Neural network construction of our 'Layout Aware Decomposition Network'.

After we convoluted the input map into a $7 \times 7 \times 1024$ output, we flatten the feature map for further processing. We predicted the rough layout using softmax on each of the 36 branches. Two dense layers and two dropout layers with drop rate of 0.4 were introduced to avoid overfitting. We pretrained 36 different configurations of layout, and used the best bounding box prediction branch from the softmax as shown. In each branch, the output will be in size of $m \times n \times 5$, where m and n refer to number of lines and rows in the feature map and each element in the feature is a vector of $(x, y, width, height, confidence)$. These object detection branches worked best with images with the same rough layout as the output size. Each bounding box prediction branch consisted of two dense layers and a reshape layer. Dense layers were introduced to get flexible shapes of following layers.

The combined loss function for LADN is:

$$loss = \sum_{m=1}^{M}\sum_{n=1}^{N}(\frac{1}{mn}\sum_{i=0}^{mn} L_i) - \lambda_{layout}\sum_{k=1}^{MN}\mathbb{1}_k^{\text{layout}}\log(l_k) \quad (3)$$

In eqn. 3, M and N refer to the maximum number of lines and rows in branches. L_i shows the loss of cell i. λ_{layout} is a hyper-parameter to balance bounding box prediction branches and the layout prediction branch. In most cases, losses from the bounding boxes were larger but losses from rough layout prediction were more important. We tune λ_{layout} for better converge of our network. $\mathbb{1}_{ij}^{\text{noobj}}$ denotes whether the compound figure fits the k^{th} layout and l_k denotes the predicted probability of given layout. Here we used a cross-entropy for rough layout prediction instead of sum of square error.

4. RESULTS

We used ImageCLEF 2016 and data extracted from open access papers curated into the INTACT database of molecular interactions (https://www.ebi.ac.uk/intact/) as training/testing data. ImageCLEF has pre-annotated subfigures with bounding boxes of few errors. It contains 6,783 training images and 1,614 test images. In order to evaluate our model on ImageCLEF, we annotated the rough layout of each image manually. We extracted high quality compound figures from INTACT PDF files, where we manually annotated 9,112 images and set 2,000 of them as a test set.

4.1. Training Process

The neural network model involves multiple branches and a large fully connected layer. As an initialization step, we pretrained our model on the ImageNet dataset for 160 epochs with its default resolution of 224×448 and 10 more epochs with our target resolution of 448×448.

We applied 5-fold cross-validation to make full use of all training samples. We fine-tuned the pretrained model in a series of steps to prevent the gradient from multiple branches interfering with each other. First we trained 26 more epochs only on the bounding box prediction branch with a layout of 6×6 solely. We then froze the parameters of all layers involved and trained the other parts together.

4.2. Results and Explanation

The performance of the LADN system is shown in Table 1. We compare our work with Lee et al. 2015 [11] (the most widely used public tool for this task), Li et al. 2016 [13] (which won the ImageCLEF 2016 competition), and Tsutsui

Dataset	Method	Accuracy	Precision	Recall
Image CLEF 2016	Lee et al.[11]	0.57	0.82	0.38
	Li et al. [13]	0.84	NA	NA
	YOLO	0.859	0.880	0.775
	LADN64	0.852	0.878	0.818
	LADN36	**0.868**	**0.896**	**0.824**
	LADN+36	0.816	0.832	0.787
INTACT 2017	Greedily Segment	0.524	0.834	0.453
	YOLO	0.765	0.792	0.727
	LADN64	0.804	0.831	0.797
	LADN36	**0.831**	**0.849**	**0.835**
	LADN+36	0.723	0.749	0.711

Table 1. Quantitative classification data for figure separation.

& Crandall 2017 [3]. We also implemented a simple heuristic method based on greedily segmenting images around the letters used to label the subfigures. We would iterate over the labeling letters in reverse alphabetical order, drawing a bounding box between coordinates of the labeling letter to the right-bottom corner of the available space. We would extract that bounding box and replace that area with background color before iterating to the next letter. This method had higher precision than several of the other non-CNN-based methods.

We also implemented various versions of our network. LADN64 refers to 'Layout Aware Decompostion Network' with 64 (from 1×1 to 8×8) different possible layouts implemented as different output branches for the neural net (see Fig. 3). LADN+36 refers to a similar network configuration where we replaced the fully connected layers with a convolutional layer of 4 sizes of convolutional kernel. (2×2, 2×3, 3×2, and 3×3).

Accuracy scores were calculated by tools provided by ImageCLEF 2016, and were defined as the number of correctly predicted images (with $> 66\%$ overlap with the ground truth) divided by the maximum number of ground truth subfigures detected. The total accuracy of the dataset is the average of each image. Precision refers to the number of correctly detected images divided by the number of images recognized. Recall refers to proportion of subfigures correctly detected over the set of all subfigures.

Although these metrics are widely accepted by the community, we highlight some issues here. For Li et al. utilized connecting component and composed several components to a subfigure. Some parts may be missing such as the legends of charts, but the metric only focus on the area or number. The metric will never report an incorrect delineation that misses these small parts. The basic YOLO approach may produce delineations much smaller than an intermediate ground truth but as long as the smaller delineation occurs within the space provided by the larger rectangle, the scoring function will measure it as 'correct'.

LADN36 outperforms all other methods (Table 1). LADN64 performs less well due to incorrect layout predictions, since it is rare to encounter scientific figures made up 8×8 subpanels. LADN+36 removes the fully connected layer (which is similar to YOLO v2 improvements over YOLO v1), but failed to get higher accuracy. We also note that the training process is sometimes unstable (see Discussion).

4.3. LADN Rough Layout Prediction Accuracy

As shown in Table 2, the LADN system predicts the correct layout as it's first prediction in only a small proportion of cases. Most errors of this nature occur because the wrong branch was predicted. INTACT data has more variety so the performance on ImageCLEF is higher.

Model	Data	top-1	top-5	top-10
64 branches	ImageCLEF 2016	0.30	0.63	0.78
64 branches	INTACT 2017	0.45	0.72	0.84
36 branches	ImageCLEF 2016	0.42	0.81	0.90
36 branches	INTACT 2017	0.56	0.86	0.96

Table 2. Rough layout prediction accuracy

5. DISCUSSION

We developed a CNN model that decomposes the substructure of published scientific figures into an intuitive format that closely corresponds to the mid-level delineation used by authors to designate subfigures. By using one output branch to roughly decide the scale of the layouts and multiple branches which have different kernel sizes suiting different scales of input, the network can predict the structure of compound figures of different sizes with an accuracy of 86.8%.

We used a fully connection layer to link the features and different branches similarly as YOLO v1 did instead apply convolutional kernels of different size. YOLO v2 elminated the fully connection layer for high performance. Fully-connected layers lose some position information, which could act as a bottleneck for training. LADN+ replaced them with different sizes of convolutional kernels to adjust the output size but failed to perform. In fact, deeper network may cause gradient vanishment and instability for training, which can be addressed by adding residual block as YOLO v3[20] does. It's reported that residual block can effectively increase the accuracy of object detection networks.

We synthesized the training data by randomly generating composite images sampled from the original dataset. We anticipate that Generative Adversarial Networks[21] may provide a powerful approach to solving issues of data augmentation. By adding a classification softmax after each element of bounding box branch, we can embed classification task to the network. Since we focus on the compound figure decomposition problem and ImageCLEF dataset has not classification annotation within the subfigure bounding box, we did not test the performance of LADN to classify. Future works can be focused on additional feature and application of our network.

6. REFERENCES

[1] Alba Garcia Seco de Herrera, Roger Schaer, Stefano Bromuri, and Henning Müller, "Overview of the imageclef 2016 medical task," in *Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum*, 2016, pp. 219–232.

[2] Gully Burns, Pradeep Dasigi, and Eduard H. Hovy, "Extracting evidence fragments for distant supervision of molecular interactions," in *Proceedings of the First Workshop on Enabling Open Semantic Science co-located with (ISWC 2017)*, 2017, pp. 7–14.

[3] Satoshi Tsutsui and David J. Crandall, "A data driven approach for compound figure separation using convolutional neural networks," in *14th IAPR International Conference on Document Analysis and Recognition, ICDAR 2017*, 2017, pp. 533–540.

[4] K. G. Hardwick, R. C. Johnston, D. L. Smith, and A. W. Murray, "MAD3 encodes a novel component of the spindle checkpoint which interacts with Bub3p, Cdc20p, and Mad2p.," *The Journal of cell biology*, vol. 148, no. 5, pp. 871–882, 2000.

[5] Yann LeCun, Yoshua Bengio, and Geoffrey E. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[6] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi, "You only look once: Unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, 2016, pp. 779–788.

[7] Tobias Kuhn, Thaibinh Luong, and Michael Krauthammer, "Finding and accessing diagrams in biomedical publications," in *AMIA 2012, American Medical Informatics Association Annual Symposium*, 2012.

[8] Samuel Kerrien, Bruno Aranda, Lionel Breuza, Alan Bridge, and et al., "The intact molecular interaction database in 2012," *Nucleic Acids Research*, vol. 40, no. Database-Issue, pp. 841–846, 2012.

[9] Songhua Xu, James P. McCusker, and Michael Krauthammer, "Yale image finder (YIF): a new search engine for retrieving biomedical images," *Bioinformatics*, vol. 24, no. 17, pp. 1968–1970, 2008.

[10] Tobias Kuhn, Mate Levente Nagy, Thaibinh Luong, and Michael Krauthammer, "Mining images in biomedical publications: Detection and analysis of gel diagrams," *J. Biomedical Semantics*, vol. 5, pp. 10, 2014.

[11] Po-Shen Lee and Bill Howe, "Detecting and dismantling composite visualizations in the scientific literature," in *Pattern Recognition Applications and Methods - 4th International Conference, ICPRAM*, 2015, pp. 247–266.

[12] Mario Taschwer and Oge Marques, "Automatic separation of compound figures in scientific articles," *Multimedia Tools Appl.*, vol. 77, no. 1, pp. 519–548, 2018.

[13] Pengyuan Li, Scott Sorensen, Abhishek Kolagunda, Xiangying Jiang, Xiaolong Wang, Chandra Kambhamettu, and Hagit Shatkay, "Udel cis working notes in imageclef 2016," in *CLEF*, 2016.

[14] Emilia Apostolova, Daekeun You, Zhiyun Xue, Sameer K. Antani, Dina Demner-Fushman, and George R. Thoma, "Image retrieval from scientific publications: Text and image content processing to separate multipanel figures," *JASIST*, vol. 64, no. 5, pp. 893–908, 2013.

[15] Ross B. Girshick, "Fast R-CNN," in *2015 IEEE International Conference on Computer Vision*, 2015, pp. 1440–1448.

[16] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 91–99.

[17] Joseph Redmon and Ali Farhadi, "YOLO9000: better, faster, stronger," in *Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6517–6525.

[18] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson, "How transferable are features in deep neural networks?," in *Advances in Neural Information Processing Systems*, 2014, pp. 3320–3328.

[19] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, and et al., "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[20] Joseph Redmon and Ali Farhadi, "Yolov3: An incremental improvement," *CoRR*, vol. abs/1804.02767, 2018.

[21] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, 2014, pp. 2672–2680.