Supplementary Material : Parsing and Summarizing Infographics with Synthetically Trained Icon Detection

Spandan Madan* Harvard University
Zoya Bylinskii† Adobe Research
Carolina Nobre‡ Harvard University
Matthew Tancik UC Berkeley
Adria Recasens MIT
Kimberli Zhong MIT
Sami Alsheikh MIT
Aude Oliva MIT
Fredo Durand MIT
Hanspeter Pfister Harvard University

1 Supplemental Material

1.1 Training an icon detection mechanism

Synthetic training data: To generate our synthetic data, we randomly sampled 600×600px windows from the *Visually29K* infographics and pasted icons onto sufficiently empty patches. Specifically, from each window, we sampled random patches with varying location and size, and measured the amount of entropy per patch to locate valid candidates. We deemed a patch valid if it had a low enough entropy (below a predefined threshold), because that meant the patch was "emptier" of texture, more likely to belong to the infographic's background, and not overlap with another visual or textual element. To measure entropy, Canny edge detection was applied to the patch. The resulting edge values were weighted by a Gaussian window centered on the patch (to give more weight to edges in the center of the patch), and summed to quantify the local entropy, with value ranging from 0 to 1. If the entropy value was below a predefined threshold, the patch was kept, otherwise it was discarded and a new patch candidate was sampled from the window. A randomly selected icon from our scraped icon collection was then pasted onto each valid patch in a window. An additional constraint required the icon to meet a predefined contrast threshold with the patch to ensure it would be visually detectable, or else a new icon would be selected. The process of first selecting a patch and then pasting an icon into it was repeated until a desired number of icons were pasted per window.

Training details: We used Faster R-CNN with parameters pre-trained on ImageNet. We then fine-tuned the model for 30K iterations on our 10K synthetic training instances using a stochastic gradient descent solver with a momentum of 0.9 and a fixed learning rate of 10^{-3}. Early termination was used during training because the network was found to converge in significantly fewer epochs than the original paper [6]. Other Faster R-CNN parameters were kept at their default values (e.g., anchor scales $[8, 16, 32]$). Each iteration used a single window with pasted icons, and Faster R-CNN used the window to generate a mini-batch of 300 region proposals. Training took 3.5 hours on an NVIDIA Titan Xp.

Evaluating different synthetic training strategies: We analyzed how different synthetic data generation parameters affect the icon proposals produced by the final trained model. We ran tests on the set of 400 validation infographics containing human annotations of 7,020 bounding boxes. We performed a grid search on 4 parameters, varying them one at a time: (a) number of icons pasted per window, (b) variation in the size of augmented icons, (c) contrast threshold between the icon and the patch, calculated as a difference in color variance between the patch and the icon, and (d) entropy threshold for a patch candidate to be deemed valid (i.e., icons are pasted into patches with entropy lower than the threshold).

We tried 5 settings for the number of icons pasted per window, from 1 to 16, doubling the number of icons for each experiment. We found no statistically significant differences in the mAP scores of the models trained with these settings. However, increasing the number of icons incurs a time cost for generating the synthetic data, since finding enough valid image patches to paste icons into becomes challenging. We found that higher scale variation during training helps the model detect icons in infographics, which often occur at different scales. By using icons with sizes ranging between 30 to 480 pixels per side (within a window sized 600×600px), we achieved the highest mAP scores. Other settings we tried included limiting the maximum icon size to 30, 60, 120, and 240 pixels per side.

We found no significant effects of varying the contrast and entropy thresholds independently, while keeping the other parameters fixed. However, when we disregard both thresholds and place icons entirely at random in the image windows, the performance of the trained model degrades significantly (see the "ablation experiments" in the paper). For generating icon proposals on test images, we finally chose the model with the highest mAP score on the 400 validation images, containing 4 icons per window, with icons varying in size from 30 to 240 pixels per side. We note that the ground truth human annotations also include an average of 4 icons per window, so our synthetic data generation roughly mimics the distribution of actual icons in in-the-wild infographics.

1.2 Topic prediction

Predicting topics from text: On average, we extracted 236 words per infographic, of which 170 had *word2vec* representations [3, 5]. The 300-dimensional mean *word2vec* of the bag of extracted words was used as the global feature vector of the text for the infographic. This feature vector was fed through a shallow network: a 300-dimensional fully-connected linear layer, followed by a ReLu, a 391-dimensional fully-connected output layer, and a sigmoid. Since each infographic could have multiple tags, we set this up as a multi-label problem with 391-dimensional one-hot encoded target vectors and the binary cross-entropy (BCE) loss. We used 26K infographics from our *Visually29K* dataset for training. Since the 300-dimensional feature vectors of all these infographics fit in memory, we trained on all of them in a single batch for 20K iterations with a learning rate of 10^{-3}.

Tagging icons: We trained an icon classifier using icon images from Google along with their associated tags. We had some variation in the number of icons scraped per tag, but for 90% of the 391 tags, we collected at least 380 icons per tag, for a total of 250K icons. We used 80% of these images for training and 20% for validation. We found that including icons both with and without transparent backgrounds during training improved the generalization of the classifier to automatically-detected icons at test-time, over just training on icons with transparent backgrounds. Training was set up as a multi-class problem with 391 tag classes. We used the ResNet18 architecture [4] pre-trained on ImageNet, and fine-tuned

*e-mail: spandan_madan@g.harvard.edu
†e-mail: zoya@adobe.com
‡e-mail: cnobre@g.harvard.edu

all the layers on 200K icon images, for a total of 4 epochs. We used cross entropy loss with a learning rate of 10^{-4} using RMSProp.

For evaluation, we used the 544 infographics for which we have human annotations (see the description of the user study labeled Task 2). We ran our automatic icon detector on all these infographics, and for each of 391 tags, we used the icon classifier's confidence to re-rank all the detected icons. Fig.8 in the paper contains the highest ranked icon proposals for a few different tags. For each icon proposal, we measure overlap with human annotations: if an icon proposal sufficiently overlaps with a ground truth bounding box (IOU$>$ 0.5), that proposal is considered successful. We obtained an mAP of 25.1% by averaging the precision of all the retrieved icon proposals, across all tags.

1.3 Memorability of visual hashtags

To test the utility of our multi-modal summaries, we ran a pilot study where users were asked to browse through a collection of 138 thumbnails of infographics and select ones which they find interesting as shown in Fig 2. For half the infographics, when a user hovered their mouse over them the extracted multi-modal summary was displayed as an overlay along with the title of the infographic (ex: second row, fourth column in Fig 2). For the other half, only the title was displayed as an overlay. The users could also click the thumbnails to be redirected to the full resolution infographic.

Our preliminary results across 15 participants show that multi-modal summaries help increase recall of previously seen infographics by 19.65% on average (median), as compared to seeing the infographic thumbnails with only the titles. This supports prior work that has shown that icons/pictograms can increase the memorability of content [1, 2]. For infographics shown with only titles, users were able to recover them with a very high precision at 95.84%, but had a low recall rate of 61%. To condense these numbers into a single metric of accuracy, we calculate the balanced F-score as shown in equation **??** below. We find that users remember infographics shown with multi-modal summaries with a median accuracy of 78.6%, as compared to 63.3% for those shown only with titles. Thus, our preliminary results suggest that including multi-modal summaries along with titles can help with tasks like browsing through large collections of infographics by increasing their recall.

References

[1] Michelle A Borkin, Zoya Bylinskii, Nam Wook Kim, Constance May Bainbridge, Chelsea S Yeh, Daniel Borkin, Hanspeter Pfister, and Aude Oliva. 2015. Beyond memorability: Visualization recognition and recall. *IEEE transactions on visualization and computer graphics* 22, 1 (2015), 519–528.

[2] Michelle A Borkin, Azalea A Vo, Zoya Bylinskii, Phillip Isola, Shashank Sunkavalli, Aude Oliva, and Hanspeter Pfister. 2013. What makes a visualization memorable? *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (2013), 2306–2315.

[3] Google. accessed in October 2017. Word2Vec Model. https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit?usp=sharing. (accessed in October 2017).

[4] K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep Residual Learning for Image Recognition. In *CVPR*.

[5] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *CoRR* abs/1301.3781 (2013). http://arxiv.org/abs/1301.3781

[6] S. Ren, K. He, R. Girshick, and J. Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *ICCV*.

Figure 1: Sample successes and failures of our automated multi-modal summarization pipeline. In both (a) and (b), the predicted text tags for the infographics are correct, and the predicted visual hashtags (solid blue boxes) overlap with human annotations (red boxes). Because a single tag might not be sufficient to summarize an infographic, we also provide an additional predicted text tag (second most likely) and corresponding visual hashtag for (a) and (b). In (c)-(e) the text model predicts the wrong tag. In (c), the semantic meaning of the predicted tag is preserved, so the visual hashtag is still correct. In (d) and (e), the wrong visual hashtags are returned as a result of the text predictions. However, we show that if the correct text tag would have been used (bottom, red), correct visual hashtags would have been returned. In dashed blue are all our icon proposals for each infographic. The underlying infographics have been faded to facilitate visualization. ©*Evanmade Graphic Design, Richard Leeds, Blue Stacks, FuelFreedom.org, CreditLoan.com*

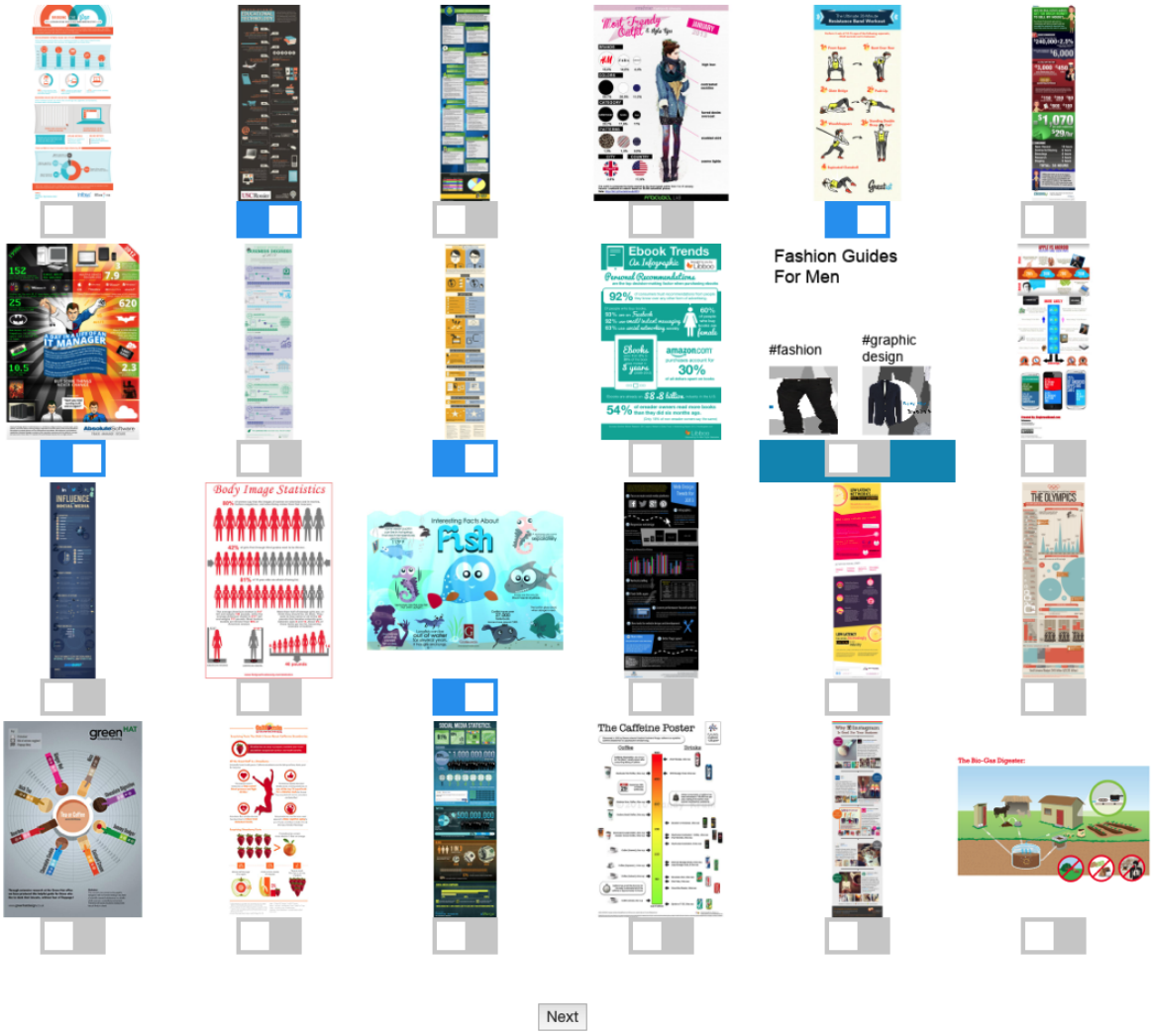

Figure 2: Screenshot of the pilot study to understand the utility of multi-modal summaries in browsing through a large collection of infographic thumbnails. Hovering over the thumbnails, users were shown multi-modal summaries along with titles for half the infographics (ex: second row, fourth column in image grid above) and asked to select the ones which they find interesting. For the other half they saw only the titles. Our results show that the multi-modal summaries lead to a 19.6% increase in the recall of previously seen infographics, as compared to the titles-only setting.