GRE: Evaluating computer vision models on Generalizability, Robustness, and Extensibility

Junwon Park
Stanford University
junwon@cs.stanford.edu

Abstract

Computer vision models are commonly evaluated on a test dataset that is sampled from the same data distribution as the data used for training the models. This methods leads to an inaccurate overstatement of model performance by failing to distinguish models that pays attention to wrong features or by rewarding models that have built a bias against long tail patterns. This work proposes a method of evaluating the generalizability, robustness, and extensibility of computer vision models on the Visual Question Answering task using datasets of unseen compositions of seen objects and scenes from the training dataset. We compare the performance of Stacked Attention Visual Question Answering model and human subjects on both the unaltered VQA dataset and the altered GRE datasets, and discover that gap between model performance and human performance on the VQA task is 6.02 percent points on the unaltered VQA dataset and 37.56 percent points on the altered GRE dataset. The result shows that the GRE method reveals the margin between the model performance and human performance more accurately on the VQA task than does the traditional evaluation metric using the accuracy score on a test dataset that comes from the same data distribution as the train dataset.

Introduction

As a standard practice, machine learning models are evaluated using a test dataset that is sampled from the same data distribution as the train dataset. Such evaluation practice allows models to unfairly leverage on its a priori knowledge of data distributions to achieve exaggerated accuracy scores. (Jabri, Joulin, and Van Der Maaten 2016) As Ribeiro et al. demonstrates, these inaccurate inflations in performance measurement undermines the trust of people in machine learning models they interact with. (Ribeiro, Singh, and Guestrin 2016) In real world deployment, computer vision models often encounter different distributions and a higher variance in data compared to the dataset used in training. As a consequence, machine learning models are over-promising in performance evaluations and under-delivering in real world deployment.

Figure 1: We introduce an evaluation method that tests for the unseen compositions of seen objects and scenes by recombining objects and scenes from the train dataset.

One instance of this problem can be found in the Visual Question Answering (VQA) task for computer vision models. (Agrawal et al. 2017) In the multiple choice VQA task, a computer vision model is asked to choose an answer among a list of answers given an image and a question. Multiple papers in the past have pointed out that the distribution of data in the VQA dataset leads to exaggerated evaluation of bad models, and called for augmenting the evaluation process with new baselines. (Jabri, Joulin, and Van Der Maaten 2016) For example, computer vision models can perform strongly on questions that ask for the color of banana even without looking at the image, because all instances of banana in the dataset are yellow (Goyal et al. 2017). While assuming the color of bananas to be yellow is useful in achieving a high precision score during evaluation, the inability of computer vision models to correctly identify colors of the rare green, red, blue bananas should be exposed to reveal the gap between the model intelligence and human intelligence, since humans can quickly accommodate the concept of bananas with colors they have never encountered before,

CS 229: Machine Learning - Fall 2019

and classify blue bananas accurately as bananas.

A test dataset that does not capture the high variance of real world data fails to catch biased models that err on patterns with few occurrences in the dataset. For example, state-of-the-art computer models trained on the VQA dataset have shown to err on such patterns by wrongly answering woman on skateboard or in front of a computer to be man. (Hendricks et al. 2018) Several work in computer vision examined attention heat maps to assist the error analysis of computer vision models (Shih, Singh, and Hoiem 2016) (Selvaraju et al. 2017), but these methods are only useful for debugging errors after errors have been made known to the developers. A good evaluation of computer vision models should expose such errors before deployment by emulating the high variance of data distribution that were not expected by dataset developers.

This work posits that computer vision models should not be evaluated using an aggregate accuracy score on test datasets lack the high variance of data points of the real world. Instead, the models should be evaluated with three augmented test datasets that each measures generalizability, robustness, and extensibility, and report the composite score that better captures the models' ability, or lack thereof, to respond to patterns they have not been exposed to during train and validation.

Concretely, we define generalizability, robustness, and extensibility as the following:

- Generalizability = If we replace the object with another object from the same category, and the model correctly predicts the same answer, then the model has a high generalizability.
- Robustness = If we replace the scene, and the model correctly predicts the same answer, then the model has a high robustness.
- Extensibility = If we replace the object with an object from a different category, and the model correctly predicts the new value, then the model has a high extensibility.

We build a system that receives a dataset of test images, then generates variations of test images with different foreground objects and background scenes. Next, we run experiments to compare the performance of a state-of-the-art VQA model and humans using both the traditional accuracy score on the original test dataset and the new composite GRE score based on the transformed test datasets. Through the result, we demonstrate that the margin between human performance and model performance is better revealed in when computer vision models are evaluated on GRE datasets rather than unaltered datasets..

Related Work

The machine learning community and computer vision community have recognized the problems in traditional methods of evaluating the models.

Kafle et al. suggested augmenting the VQA dataset using automated data annotation techniques to build a more complete dataset featuring a wide variety of question-answer pairs. (Kafle, Yousefhussien, and Kanan 2017) For example, they used object recognition models to add ("is the *object* present?", "yes") question-answer pair to the image for every object where the detected object's area was greater than 2000 pixels. Using this method, they introduced 81.5% augmentation to YES/NO questions, 4.8% to Number questions and 13.6% to Other questions. While their work contributes to a more complete annotation of objects that are already present in the dataset, their work does not augment the dataset with scenarios that are not present in the dataset, such as blue bananas. Unlike their work, this work augments the VQA dataset with transformed images, thereby introducing new visual scenarios that were not present in the original dataset.

ImageNet-C and ImageNet-P introduced corruptions to the images in the ImageNet dataset by adding static noise and applying filters to the testset, such as overlaying a translucent fog image on top of the images. (Hendrycks and Dietterich 2019) However, their syntactic transformations, such as adding white pixels to the image to emulate snow without regarding the actual image content, does not change the data distribution of semantic information. For example, in all 15 types of algorithmically generated corruptions to an image, a bird is always sitting on a tree branch. In contrast, our method tests robustness in unseen data distributions by replacing objects like tree branches with new objects like a rock, or a skateboard.

Anchors removed regions of interest from original images and overlaid them on unrelated images to understand the visual anchors in the images that are minimally sufficient to leading computer vision models to the correct answers. (Ribeiro, Singh, and Guestrin 2018) A problem with this approach is that Anchors overlays incoherent parts of an image, for example the face, the back, and the tail of a dog, but not its legs, onto an unrelated scene, such as underwater, making the resulting image syntactically unrealistic. In contrast, this work surgically replaces objects or scenes in complete parts, thereby maintaining visual syntactic coherence while introducing new semantic distributions.

Errudite is an interactive interface for task-agnostic and model-agnostic error analysis of natural language processing models for generating new example data points to reproduce and better understand the errors that models make. (Wu et al. 2019) Unlike Errudite which assigns the duty of forming hypotheses about error conditions to the human user, this work automatically tests systematic alterations to data points, helping human users discover patterns of alterations that lead to a reduced performance in their models.

Figure 2: Object masking receives an image, and outputs a segmentation mask over an object to remove.

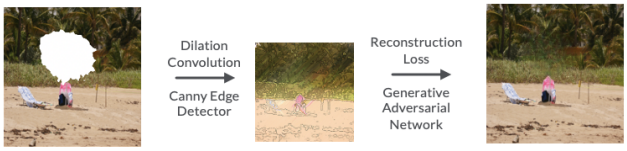

Figure 3: Object removal receives an image with segmentation mask, and paints the region behind the segmentation mask without the object in the original image.

Method

We build a system that receives a VQA dataset consisting of images, questions, and answers, then generates three new datasets, respectively for testing generalizability, robustness, and extensibility. To achieve this, the system has three serial processes: object masking, object removal, image overlay.

Object Masking

Object masking receives an image, and outputs a segmentation mask over an object to remove. For object masking stage, we extend the Visual Question Segmentation model, which produces segmentation masks over objects conditioned on the image and question as input, and build a model that produces segmentation masks over objects conditioned on the answer as well as the image and question. (Gan et al. 2017)

Jabri et al. proposes a method of transforming the multi-class classification problem of the multiple choice VQA task into a binary classification problem and solve them with the multilayer perceptrons (MLP) model. (Jabri, Joulin, and Van Der Maaten 2016) The multilayer perceptrons model is a neural network with the following activation function:

$$y = \sigma(W_2 \max(0, W_1 x_{iqa}) + b) \tag{1}$$

where the x_{iqa} term is the concatenation of feature representation of image, question, and candidate answer about the image. We extract the image features using the modification of the output layer proposed in the VQS model. (Gan et al. 2017) VQS imposes a sigmoid function for each of C = 256 attributes in ResNet output layer, and train the network using the binary cross-entropy loss. Image features are ResNet pool5 activation and attribute features extracted as described above. Question and candidate answer are encoded by averaging the 300D word2vec vectors of the words and applying the l_2 normalization. The resulting trained model outputs a semantically segmentation the right visual entities out of image given image, question, and answer, which the GRE system uses as the segmentation mask.

Object Removal

Object removal receives an image with segmentation mask, and paints the region behind the segmentation mask without the object in the original image. For object removal stage, we train a model proposed by Nazeri et al. which performs inpainting by hallucinating edges in the masked pixel regions, then uses a generative adversarial network to produce a realistic image in the regions of interest. (Nazeri et al. 2019)

The object removal stage itself is a system of two networks: the edge generator and the image completion network. The edge generator uses an implantation of Canny Edge Detector, which detects edges from images by taking pixels with high optical flow gradients with respect to neighboring pixels, for extracting edges from the input image with segmentation mask. Then a network is trained on the objective that takes into account both the adversarial loss and the feature-matching loss:

$$\min_{G_1} \max_{D_1} \mathcal{L}_{G_1} = \min_{G_1} (\lambda_{adv,1} \max_{D_1} (\mathcal{L}_{adv,1}) + \lambda_{FM} \mathcal{L}_{FM}) \tag{2}$$

$\lambda_{adv,1}$ and λ_{FM} are regularization parameters. The adversarial loss is defined as following:

$$\mathcal{L}_{adv,1} = \mathbb{E}_{(C_{gt}, I_{gray})}[\log D_1(C_{gt}, I_{gray})] \tag{3}$$

$$+\mathbb{E}_{(I_{gray})} \log[1 - D_1(C_{pred}, I_{gray})] \tag{4}$$

C_{gt} is the edge map. I_{gray} is the grayscale image. C_{pred} is the predicted edge map over the masked region.

The feature-matching loss is defined as following:

$$\mathcal{L}_{FM} = \mathbb{E}[\sum_{i=1}^{L} \frac{1}{N_i} ||D_1^{(i)}(C_{gt}) - D_1^{(i)}(C_{pred})||_1] \tag{5}$$

N_i is the number of elements in the i'th activation layer. D_1^i is the activation in the i'th layer of discriminator.

The image completion network is trained with the incomplete color image I_{gt} conditioned on a composite edge map C_{comp}, and produces a predicted colorization, I_{pred}. The objective function for image completion network depends on perceptual loss and style loss. The perceptual loss is defined as following:

$$\mathcal{L}_{perc} = \mathbb{E}[\sum_{i=1}^{L} \frac{1}{N_i} ||\phi_i(I_{gt}) - \phi_i(I_{pred})||_1] \tag{6}$$

Figure 4: Object overlay receives a scene image, an object image, and the bounding box of interest, then places the object image over the scene image.

The style loss is defined as following:

$$\mathcal{L}_{style} = \mathbb{E}_j[||G_j^{\phi}(I_{pred}) - G_j^{\phi}(I_{gt})||_1] \quad (7)$$

The resulting system of edge generator and image completion network performs an image inpainting by colorizing the region behind the segmentation mask.

Object Overlay

Object overlay receives a scene image, an object image, and the bounding box of interest, then places the object image over the scene image. The system draws a bounding box that has the same aspect ratio as the input object image and fits inside the segmentation mask, scales the object image to match the size of the bounding box, then copies pixels from the object image into the bounding box of the scene image.

An implementation of the GRE system is available on GitHub (https://github.com/productceo/gresco) for replication studies and future work.

Experiments

To explore the capability of GRE evaluation method in revealing the gap between model performance and human performance on computer vision tasks, we compare the gap between model performance and human performance after evaluating both on the unaltered test set that is sampled from the same distribution as the dataset used for training the models, and the three datasets that result from applying the GRE method on the original dataset.

Concretely, we use the GRE method to generate 3 datasets based the VQA V2 validation dataset, which has 40,504 MS-COCO images and 214,354 question and answer annotations. For the objects and scenes to introduce, we use Open Images V2 dataset, 9 million images across 93 classes, and MIT Places dataset, with 10 million images across 365 classes, respectively. We implement Stacked Attention Visual Question Answering (SAVQA) model (Yang et al. 2016), Bottom Up Visual Question Answering (BUVQA) model (Anderson et al. 2018), and a CNN-LSTM model.

Table 1: Accuracy score measured for SAVQA model and human subjects for each of the datasets: VQA, VQA-Generalizability, VQA-Robustness, VQA-Extensibility, and the GRE score which is an average of the accuracy scores measured in the three altered datasets

	SAVQA (%)	Human (%)
VQA Dataset	84.39	90.40
VQA-G Dataset	36.38	87.00
VQA-R Dataset	31.69	46.75
VQA-E Dataset	30.98	78.00
GRE Score	33.02	70.58

For every image from the VQA dataset, VQA-G dataset introduces 3 variants featuring 3 randomly selected objects from the same object class (e.g. replacing "cup" with another "cup"), VQA-R dataset introduces variants featuring 3 randomly selected scenes from each of 3 randomly selected scene classes (e.g. replacing the background of a car image from "city" to "beach"), and the VQA-E dataset introduces variants featuring 3 randomly selected objects from each of 3 randomly selected object classes (e.g. replacing "banana" with "dog").

We hypothesize that the absolute difference between the model performance and the human performance will be larger when the evaluation uses GRE method as opposed to the traditional method of using an unaltered dataset.

Results

Note: We implemented, trained, and evaluated three VQA models (SAVQA, BUVQA, CNN-LSTM) on the VQA dataset, but did not have resources to run BUVQA and CNN-LSTM on GRE datasets in time. We worked with SAVQA which performed stronger than the other two on the original VQA dataset. The evaluation of all thre computer vision models on the original VQA dataset is recorded in this Google Spreadsheet: https://bit.ly/2CMsF5D.

GRE system detected objects successfully in 82,577 images out of the x images in the input VQA validation dataset, and introduced 186,497 new image, question, answer tuples to the dataset for evaluating generalizability and 388,158 each for robustness and extensibility.

Result shows that the margin between human performance and model performance is 6.02 percent points when measured with original VQA dataset and 37.56 percent points when measured with GRE datasets. The difference between humans and models is greater when evaluated on the GRE datasets than on the original VQA dataset. Thus, our hypothesis holds.

Discussion

The result demonstrates that the traditional evaluation method using an unaltered dataset understates the gap between model performance and human performance which is revealed in the GRE evaluation method.

It is noteworthy that human subjects also saw a significant drop in performance on robustness and extensibility. This, however, is due to the system failing to generate semantically meaningful images that confused the human subjects. A common source of error was the object masking stage. An incorrect masking of the segmentation model led to new objects being added over existing objects without removing them, or pasting only parts of the objects to new scenes. Both forms of failure led to semantically invalid visual scenes that human subjects could not interpret. Out of 90 randomly selected samples of resulting images across the three transformations, 34 were flagged by human subjects to be semantically invalid, exposing that the system currently fails on a significant portion of the dataset. As we continue this work, we see a need to improve the system which we can indirectly measure by closing the gap between human performance on GRE datasets and the original VQA dataset.

Future Work

This work inspires three future work directions.

Data Augmentation with GRE Datasets

GRE method introduces complex and uncommon combinations of objects and scenes into a dataset which is important to making the dataset more complete by emulating the variance of data present in the real world. The same method could be applied to the train dataset, rather than the test dataset, as a data augmentation method to introduce a larger and a more complete training dataset for computer vision models. We predict that computer vision models trained with datasets that are augmented using the GRE method will exhibit a greater performance than the models trained on unaugmented datasets on generalizability, robustness, and extensibility measures.

Other Computer Vision Tasks

This work explored applying the GRE method to evaluate computer vision models that perform the visual question answering task. In theory, the GRE method is applicable to all computer vision tasks, since the GRE method performs transformations conditioned only on the images. Thus, future work could apply GRE method on other computer vision tasks ranging from object classification to scene graph prediction to demonstrate whether the GRE method generalizes to other computer vision tasks.

Single-Source-Multi-Target Transformation GAN

We introduced a system of multiple computer vision models and algorithms that work together to transform images, because we could not find literature on a generative adversarial network architecture that can transform images from a single source domain to multiple target domains. This remains as an exciting and unexplored area of research.

Conclusion

GRE is an evaluation metric that introduces new combinations of objects and scenes into any image dataset to measure the generalizability, robustness, and extensibility of computer vision models. An experiment on the VQA task demonstrates that the GRE method outperforms the traditional evaluation metric using an unaltered dataset sampled from the same distribution as the train dataset in revealing the gap between the humans and models. GRE envisions a future in which evaluation metrics can reveal the underperformance of models that are observing wrong features or are biased against long tail patterns before they are deployed into the real world.

Contributions

Professor Michael Bernstein and Professor Fei-Fei Li advised me, and Ph.D Candidate Ranjay Krishna actively provided mentorship through the project. Khaled Jedoui built the object masking stage which extends the VQS model. As acknowledged in the body, this work is made possible thanks to the datasets provided by VQA team at Virginia Tech and Georgia Tech, Places team at MIT CSAIL, Open Images team at Google.

References

[Agrawal et al. 2017] Agrawal, A.; Lu, J.; Antol, S.; Mitchell, M.; Zitnick, C. L.; Parikh, D.; and Batra, D. 2017. Vqa: Visual question answering. *International Journal of Computer Vision* 123(1):4–31.

[Anderson et al. 2018] Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6077–6086.

[Gan et al. 2017] Gan, C.; Li, Y.; Li, H.; Sun, C.; and Gong, B. 2017. Vqs: Linking segmentations to questions and answers for supervised attention in vqa and question-focused semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, 1811–1820.

[Goyal et al. 2017] Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6904–6913.

[Hendricks et al. 2018] Hendricks, L. A.; Burns, K.; Saenko, K.; Darrell, T.; and Rohrbach, A. 2018. Women also snow-

board: Overcoming bias in captioning models. In *European Conference on Computer Vision*, 793–811. Springer.

[Hendrycks and Dietterich 2019] Hendrycks, D., and Dietterich, T. 2019. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*.

[Jabri, Joulin, and Van Der Maaten 2016] Jabri, A.; Joulin, A.; and Van Der Maaten, L. 2016. Revisiting visual question answering baselines. In *European conference on computer vision*, 727–739. Springer.

[Kafle, Yousefhussien, and Kanan 2017] Kafle, K.; Yousefhussien, M.; and Kanan, C. 2017. Data augmentation for visual question answering. In *Proceedings of the 10th International Conference on Natural Language Generation*, 198–202.

[Nazeri et al. 2019] Nazeri, K.; Ng, E.; Joseph, T.; Qureshi, F.; and Ebrahimi, M. 2019. Edgeconnect: Generative image inpainting with adversarial edge learning. *arXiv preprint arXiv:1901.00212*.

[Ribeiro, Singh, and Guestrin 2016] Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144. ACM.

[Ribeiro, Singh, and Guestrin 2018] Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2018. Anchors: High-precision model-agnostic explanations. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[Selvaraju et al. 2017] Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, 618–626.

[Shih, Singh, and Hoiem 2016] Shih, K. J.; Singh, S.; and Hoiem, D. 2016. Where to look: Focus regions for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4613–4621.

[Wu et al. 2019] Wu, T.; Ribeiro, M. T.; Heer, J.; and Weld, D. S. 2019. Errudite: Scalable, reproducible, and testable error analysis. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 747–763.

[Yang et al. 2016] Yang, Z.; He, X.; Gao, J.; Deng, L.; and Smola, A. 2016. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 21–29.