Supplementary Material: A Benchmark for Compositional Visual Reasoning

1 Social Impact

Our work presents a novel visual reasoning benchmark that evaluates sample efficiency and compositionality. Benchmarks have played an important role in guiding the development of Machine Learning and AI research. Since our benchmark exposes shortcomings of modern architectures in terms of sample efficiency and compositionality, it promotes the development of efficient and human-inspired solutions for visual reasoning and AI research in general. The byproduct of sample-efficient solutions is a lower training carbon footprint. Furthermore, our work encourages researchers to propose general solutions that can be applied to other task setups and extend beyond the simple odd-one-out task. Finally, this work has implications for cognitive science research since the benchmark can be utilized to test various hypotheses on non-verbal learning and abstraction in humans. We could not identify our research’s negative ethical or social impact.

2 Dataset Generation

Prior Assumptions

- Object are defined as closed contours sampled randomly with 2 parameters: the radius of the largest circle that can fit inside the object and the distance between two consecutive points in the closed contour.

- The object size scales the width and height of the object linearly.

- The background color is always white and object colors are sampled in the HSV color space.

- The center of an object is defined as the mid point between its horizontal and vertical extents. It is used for placing the object based on its position. By default, objects are not in contact and do not overlap unless they have an insideness or a contact relation. To ensure this, their positions are sampled with rejection sampling.

All parameter values are sampled from a fixed range for each task. For example, the object size is sampled such that all objects can fit inside the image. Within the specified ranges, value combinations might violate priors. Rejection sampling is employed to avoid choosing such value combinations.

Dataset Statistics Among the 103 tasks, 9 are based on a single elementary relation, 20 are compositions of one elementary relation, 65 are compositions of a pair of relations and 9 are compositions of more than 2 elementary relations. Figure 1 details the number of unique rules for each pair of elementary relations. The dataset was designed to include at least one unique rule for each pair. However, one of its limitations is the uneven distribution of unique rules across the pairs. This limitation could bias results in favor of models that perform better on pairs with a larger number of unique rules.

Genralization Sets  Among the limitations of our dataset is that tasks do not account for shortcuts in the test split. Shortcuts are biases in the tasks that neural networks exploit to solve them. An explanatory example is the counting task. If objects have the same size, the neural network can easily solve the task by summing the pixels of the image without analyzing the scene. If a model exploits this shortcut, it does not need to learn the concept of an object or counting. To account for this limitation and evaluate out-of-distribution generalization, we develop a generalization test set that differs from the in-distribution test set in several ways.

• **Parameter value ranges**: the ranges used for sampling parameters, especially target-relation parameters, are changed.

• **Object contour specification**: the distance between dots in contours is randomized locally, resulting in fuzzier contours.

• **Sets of random and fixed parameters**: the generation process samples 1 value that is fixed across the 4 images for fixed parameters and samples 4 values for random parameters. In the generalization test set, the sets of random and fixed parameters are changed without affecting the rules.

3 Dataset Documentation

We provide the full dataset generation and model training code [here](#) We bear all responsibility in case of violation of rights. The dataset and the generation methods are released with the Apache License, Version 2.0. The Github repository will reflect all updates to the dataset.

4 Experiment Details

**Learning to spot the Odd-One-Out**  The training setup for standard vision models is straightforward; models are trained to represent the odd-one-out differently from the three other images. The four images of the problem are fed separately to the model. Their representations are transformed into a low-dimensional space where the distances between the four representations are computed. The cosine similarity of the odd-one-out to the group is minimized with a cross-entropy-based loss. Given the 4 image representations \( x_i \), the logits \( y_i \) used for computing the softmax cross-entropy loss are the negative sum of the similarity scores.

\[
y_i = -\sum_{j \neq i} \frac{x_i \cdot x_j}{||x_i||_2 \cdot ||x_j||_2}
\]

Although we follow a specific training setting for the baseline architectures, we do not impose this methodology for future work as more sophisticated methods for comparing the four images of a problem can be designed.

**RPM Baselines**  In order to provide a fair comparison for models designed for solving RPMs, we adapt the odd-one-out task to the matrix and choice selection task setup. In the RPM setting, models are fed the nine panels with tags that indicate the position on the matrix. Each of the eight choice
panels is concatenated individually to the eight context panels. The model outputs a logit for each of the eight matrices used to compute the cross-entropy loss. The training process is explained in detail here [1]. We discard the position tags in our setting since the four images have no sense of progression. We replace context panels with the four problem images and use the same four images as choice panels, the correct choice being the outlier.

**Self-Supervised Pretraining** The SSL pretraining objective function maximizes the similarity between transformations of the same image. When SSL is performed on datasets such as ImageNet [3], where the downstream task is classification, the images are transformed in ways that maintain the class information. These transformations are augmentations, including spatial transformations such as random cropping and flipping, noise (e.g., additive Gaussian noise or blurring), and color transformations such as grey-scale and color jittering. In our setup, we use the augmentations: random resize, random Gaussian blurring, random horizontal flips, and random rotations in multiples of 90 degrees. To ensure the variety of images covers all structures used in the dataset, we select one image from each problem in the dataset, in total 1M images. This number of images is equivalent to the number of images in the ImageNet dataset, which is customarily used in SSL pretraining.

![Figure 3](image-url) **Figure 3: RPM training setup**: (a) A sample RPM problem adapted from [8], the matrix contains context panels, and a choice is taken from the answer set. (b) Inference in RPM models, SCL and WReN. The model takes all context panels with one of the choices and outputs a score. These scores from 8 choices are used for computing the cross-entropy loss. (c) An odd-one-out problem based on size and color. (d) The problem is adapted to RPM by placing all images as context and choice. The odd-one-out has the highest score among the choices.
5 Architectures and Hyperparameters

We adapt model architectures from reference implementations: ResNet-50\(^1\), ViT\(^2\), SCL\(^3\) and WReN\(^4\). We also endow SCL with ResNet18, a strong vision backbone. We name this architecture SCL-ResNet-18. All models are trained using images with size 128 × 128 pixels.

In preliminary experiments, we hand-tuned the learning rate and weight decay for all models and selected the hyperparameters that achieved the highest performance in the joint training setting for each model. We also analysed the random seed effect and observed that training results are robust with respect to the seed.

We equipe standard vision models, ViT-small and Resnet 50, with an MLP that extracts task-specific information. It takes as input image features and the task embedding and outputs a lower dimensional vector used for computing the pairwise distances and the loss. The MLP contains 2 layers, the hidden layer’s size is 2048 and the output size is 128. The task embedding space has 64 dimensions. All models were trained using Adam optimizer\(^6\) for 100 epochs with early stopping after 30 epochs. The mini-batch size used for training all models is 64, except when the training set is smaller. The learning rates are scaled linearly with the batch size. The learning rate and weight decay values are provided in\(^1\) All experiments were conducted on an internal cluster. We used 1500 GPU hours on NVIDIA V100, TitanRTX and QuadroRTX.

<table>
<thead>
<tr>
<th>Backbone Params</th>
<th>Total Params</th>
<th>learning rate</th>
<th>weight decay</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet 50</td>
<td>23.5 M</td>
<td>28.1 M</td>
<td>0.0001</td>
</tr>
<tr>
<td>ViT-small</td>
<td>21.6 M</td>
<td>21.8 M</td>
<td>0.000001</td>
</tr>
<tr>
<td>SCL</td>
<td>176 k</td>
<td>176 k</td>
<td>0.001</td>
</tr>
<tr>
<td>WReN</td>
<td>1.5 M</td>
<td>1.5 M</td>
<td>0.0001</td>
</tr>
<tr>
<td>SCL-ResNet 18</td>
<td>11.2 M</td>
<td>11.6</td>
<td>0.0005</td>
</tr>
</tbody>
</table>

Table 1: Model sizes and training hyperparameters.

---

\(^1\)https://pytorch.org/vision/stable/models.html
\(^2\)https://github.com/facebookresearch/moco-v3
\(^3\)https://github.com/dhh1995/SCL
\(^4\)https://github.com/Fen9/WReN
We provide more results on compositionality evaluation in Figure 4 and Table 3. Results on the joint training setting are consistent with the individual training setting results. We also provide results on the generalization test set in Table 2 and Figure 5. Finally, we expand on task difficulty results in Figure 6.

7 Behavioral Experiments

We used 20 problem samples for each rule, which corresponds to the lowest number of samples used for training baseline models. We recruited 21 participants from Prolific. All participants signed a consent form prior to participation and received $10.50 US per hour for participation. The study was approved by the Institutional Review Board of Brown University. Participants were instructed to identify the odd stimulus violating the rule they had to infer over a series of trials. Prior to the practice phase, they were quizzed on their understanding of the task. Participants practiced the task on a separate set of visual stimuli different from the benchmark. During the experiment, participants were informed about the start of each block as well as the concomitant rule switch. For each trial, they were presented with 4 choices on the screen and instructed to choose the image which seemed to be different according to the rule that they had to learn. They rated confidence in their choice and received feedback after each trial. In addition, they were asked to describe the rule at the end of each block.

### Table 2: Out-Of-Distribution Generalization Results

<table>
<thead>
<tr>
<th>N train samples</th>
<th>20</th>
<th>50</th>
<th>100</th>
<th>200</th>
<th>500</th>
<th>1000</th>
<th>SES</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>rand-init</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ResNet-50</td>
<td>26.3</td>
<td>0</td>
<td>28.1</td>
<td>0</td>
<td>29.1</td>
<td>1</td>
<td>30.3</td>
<td>2</td>
</tr>
<tr>
<td>ViT-small</td>
<td>26.9</td>
<td>0</td>
<td>28.1</td>
<td>0</td>
<td>28.8</td>
<td>1</td>
<td>29.4</td>
<td>2</td>
</tr>
<tr>
<td>SCL</td>
<td>26.0</td>
<td>0</td>
<td>27.8</td>
<td>0</td>
<td>28.0</td>
<td>0</td>
<td>27.9</td>
<td>0</td>
</tr>
<tr>
<td>WReN</td>
<td>27.2</td>
<td>0</td>
<td>28.8</td>
<td>1</td>
<td>29.4</td>
<td>1</td>
<td>30.1</td>
<td>2</td>
</tr>
<tr>
<td>SCL-ResNet-18</td>
<td>28.8</td>
<td>1</td>
<td>31.1</td>
<td>1</td>
<td>31.7</td>
<td>3</td>
<td>32.4</td>
<td>4</td>
</tr>
<tr>
<td>joint</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ResNet-50</td>
<td>26.0</td>
<td>0</td>
<td>26.6</td>
<td>0</td>
<td>27.8</td>
<td>1</td>
<td>30.0</td>
<td>1</td>
</tr>
<tr>
<td>ViT-small</td>
<td>26.2</td>
<td>0</td>
<td>26.4</td>
<td>0</td>
<td>26.6</td>
<td>1</td>
<td>26.9</td>
<td>1</td>
</tr>
<tr>
<td>SCL</td>
<td>25.4</td>
<td>0</td>
<td>25.6</td>
<td>0</td>
<td>27.5</td>
<td>0</td>
<td>30.3</td>
<td>0</td>
</tr>
<tr>
<td>WReN</td>
<td>26.1</td>
<td>0</td>
<td>25.9</td>
<td>0</td>
<td>26.8</td>
<td>0</td>
<td>27.8</td>
<td>0</td>
</tr>
<tr>
<td>SCL-ResNet-18</td>
<td>26.0</td>
<td>0</td>
<td>27.0</td>
<td>0</td>
<td>29.9</td>
<td>3</td>
<td>32.1</td>
<td>4</td>
</tr>
<tr>
<td>SSL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rand-init</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ResNet-50</td>
<td>32.0</td>
<td>5</td>
<td>37.0</td>
<td>6</td>
<td>38.8</td>
<td>10</td>
<td>40.9</td>
<td>10</td>
</tr>
<tr>
<td>ViT-small</td>
<td>36.2</td>
<td>5</td>
<td>39.7</td>
<td>7</td>
<td>40.6</td>
<td>7</td>
<td>41.7</td>
<td>7</td>
</tr>
<tr>
<td>joint</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ResNet-50</td>
<td>34.0</td>
<td>4</td>
<td>34.3</td>
<td>4</td>
<td>37.9</td>
<td>8</td>
<td>38.4</td>
<td>6</td>
</tr>
<tr>
<td>ViT-small</td>
<td>34.1</td>
<td>6</td>
<td>33.0</td>
<td>5</td>
<td>32.5</td>
<td>4</td>
<td>33.2</td>
<td>6</td>
</tr>
</tbody>
</table>

The table shows the Out-Of-Distribution Generalization Results. Most models perform significantly worse on the generalization test set.
<table>
<thead>
<tr>
<th>N train samples</th>
<th>rand init</th>
<th>transfer</th>
<th>difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>ResNet-50</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>25.9</td>
<td>27.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>30.0</td>
<td>30.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4.10</td>
<td>3.06</td>
</tr>
<tr>
<td><strong>ViT-Small</strong></td>
<td></td>
<td>25.9</td>
<td>27.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>28.3</td>
<td>27.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.41</td>
<td>0.89</td>
</tr>
<tr>
<td><strong>SCL</strong></td>
<td></td>
<td>26.2</td>
<td>29.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>30.3</td>
<td>32.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4.11</td>
<td>3.43</td>
</tr>
<tr>
<td><strong>WReN</strong></td>
<td></td>
<td>28.8</td>
<td>30.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>29.8</td>
<td>32.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.04</td>
<td>2.03</td>
</tr>
<tr>
<td><strong>SCL-ResNet-18</strong></td>
<td></td>
<td>28.7</td>
<td>33.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>36.3</td>
<td>39.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7.60</td>
<td>5.80</td>
</tr>
<tr>
<td><strong>ResNet-50</strong></td>
<td></td>
<td>25.5</td>
<td>26.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>29.8</td>
<td>30.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4.31</td>
<td>4.30</td>
</tr>
<tr>
<td><strong>ViT-Small</strong></td>
<td></td>
<td>25.6</td>
<td>26.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>27.6</td>
<td>27.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.00</td>
<td>1.97</td>
</tr>
<tr>
<td><strong>SCL</strong></td>
<td></td>
<td>25.3</td>
<td>26.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>27.2</td>
<td>28.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.87</td>
<td>1.90</td>
</tr>
<tr>
<td><strong>WReN</strong></td>
<td></td>
<td>27.0</td>
<td>26.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>28.0</td>
<td>29.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.06</td>
<td>2.56</td>
</tr>
<tr>
<td><strong>SCL-ResNet-18</strong></td>
<td></td>
<td>26.2</td>
<td>27.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>32.5</td>
<td>34.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6.33</td>
<td>6.76</td>
</tr>
</tbody>
</table>

Table 3: **Curriculum Condition**: Models are pretrained on the elementary tasks before finetuning on the complex tasks (transfer). They are compared to models trained from a random initialization (rand init).
Figure 5: **Performance across settings for all the models.** The accuracy is aggregated over all rules. Random choice accuracy is 0.25.
Figure 6: **Task difficulty**: Average accuracy on the elementary rules and their pair-wise compositions. **Individual vs. Joint**: Models are trained on each rule separately or trained jointly on all rules. **Rand-Init vs. SSL**: models are randomly initialized or pretrained with self-supervision.

Figure 7: **Behavioral experiment instructions.**
8 Rule Examples

Figure 8: Elementary rules
Figure 9: Composition rules
Figure 10: Composition rules
9 Datasheet

9.1 Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

CVR was created for evaluating the sample efficiency and compositionality of visual reasoning models. An important difference between current non-verbal reasoning benchmarks is the number of unique task rules they compose out of their priors and the number of samples available for training architectures on individual rules. On one end, ARC has a large number of unique tasks (1000) and very low number of training samples (3). On the other end, SVRT, RAVEN and PGM have a low number of unique rules and a large number of training samples. These fundamental differences explain the success of neural architectures in one case; few unique tasks and many training samples, and their failure in the other; many unique tasks and few training samples. CVR fills a gap between the current benchmarks by proposing a large number of unique tasks and a reasonable number of training samples. Furthermore, our benchmark encourages the development of more sample-efficient and versatile neural architectures by evaluating both aspects.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

This dataset was created by Aimen Zerroug, Mohit Vaishnav, Julien Colin, Sebastian Musslick and Thomas Serre, in affiliation with ANITI, Serrelab, Brown University and CerCo.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

This work is funded by NSF (IIS-1912280) and ONR (N00014-19-1-2029) to TS. Additional support was provided by the ANR-3IA Artificial and Natural Intelligence Toulouse Institute (ANR-19-P3A-0004). Computing resources used supported by the Center for Computation and Visualization (NIH S10OD025181) at Brown and CALMIP supercomputing center (Grant 2016-p20019, 2016-p22041) at Federal University of Toulouse Midi-Pyrénées.

9.2 Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The dataset is generated synthetically. Each sample consists of an odd-one-out problem that contains 4 images. The goal is to spot outlier among the 4 images.

How many instances are there in total (of each type, if appropriate)? The dataset contains 103 unique tasks, that we call rules. For each task, it contains 10000 training samples, 500 validation samples, 1000 test samples and 1000 generalization test samples.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

Since the dataset is synthetic, an arbitrary number of samples can be generated using the provided code. We limit the number of generated samples to 10000 and the random seed for benchmarking consistency.
What data does each instance consist of? “Raw” data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each odd-one-out problem consist of 4 images. Each image contains objects defined as closed contours. The outlier can be detected based on relationships between objects.

Is there a label or target associated with each instance? If so, please provide a description.

The label in each problem is the index of the outlier.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

No information is missing, data is complete.

Are relationships between individual instances made explicit (e.g., users’ movie ratings, social network links)? If so, please describe how these relationships are made explicit.

A task token is provided with each sample.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

We provide training, validation, test and generalization test splits for each task. They are generated using different random seeds.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

Since the generation process relies on random sampling from uniform distributions, it is possible to generate the same problem twice. However, this possibility is minimal because all generation processes use several continuous and discrete variables.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self-contained.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.

No, the dataset is synthetic.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

No, the dataset is synthetic.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No, the dataset contains no human-related data.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

N/A
Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

N/A

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

N/A

Any other comments?

9.3 Collection Process

N/A This data is synthetic, no data collection process was involved.

9.4 Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

No processing is applied.

Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the “raw” data.

N/A

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

N/A

9.5 Uses

Has the dataset been used for any tasks already? If so, please provide a description.

This is the first use for the dataset.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

N/A

What (other) tasks could the dataset be used for?

The dataset is created for non-verbal visual reasoning. However, it can be modified to incorporate language and be used for tasks such as visual question answering (VQA).

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks)? If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?
No, the dataset contains no human-related data. There is no risk for harm.

Are there tasks for which the dataset should not be used? If so, please provide a description.

The dataset is intended for neural network architecture evaluation only.

9.6 Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

Yes, the dataset is publicly available.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub) Does the dataset have a digital object identifier (DOI)?

The dataset is available in a Github repository.

When will the dataset be distributed?

The dataset is distributed in 2022.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The dataset is released under the Apache License, Version 2.0.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

N/A

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

N/A

9.7 Maintenance

Who will be supporting/hosting/maintaining the dataset?

Aimen Zerroug is supporting and maintaining the dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

aimen_zerroug@brown.edu

Is there an erratum? If so, please provide a link or other access point.

All fixes and updates to the dataset will be visible in reflected in the Github repository.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

This information can be found on the github repository.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a
fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

N/A

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

N/A

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

Contributors are allowed and welcome to do so and should contact the authors about incorporating changes.
References


