Testing Deep Neural Networks on the Same-Different Task

Nicola Messina, Giuseppe Amato, Fabio Carrara, Fabrizio Falchi, Claudio Gennaro
Institute of Information Science and Technologies
National Research Council
Pisa, Italy
{name.surname}@isti.cnr.it

***Abstract*—Developing abstract reasoning abilities in neural networks is an important goal towards the achievement of human-like performances on many tasks. As of now, some works have tackled this problem, developing ad-hoc architectures and reaching overall good generalization performances. In this work we try to understand to what extent state-of-the-art convolutional neural networks for image classification are able to deal with a challenging abstract problem, the so-called *same-different* task. This problem consists in understanding if two random shapes inside the same image are the same or not. A recent work demonstrated that simple convolutional neural networks are almost unable to solve this problem. We extend their work, showing that ResNet-inspired architectures are able to learn, while VGG cannot converge. In light of this, we suppose that residual connections have some important role in the learning process, while the depth of the network seems not so relevant. In addition, we carry out some targeted tests on the converged architectures to figure out to what extent they are able to generalize to never seen patterns. However, further investigation is needed in order to understand what are the architectural peculiarities and limits as far as abstract reasoning is concerned.**

***Index Terms*—AI, Deep Learning, Abstract Reasoning, Relational Reasoning, Convolutional Neural Networks**

I. Introduction

Artificial intelligence and in particular deep neural networks have recently shown impressive results in key domains such as vision, language, control, and decision-making. In particular, with the work carried out on the ImageNet challenge, Krizhevsky et al. [1] demonstrated major capabilities of deep neural networks in the field of image processing. Deep learning architectures, and in particular Convolutional Neural Networks (CNNs) [2], constitute now de-facto standard approaches to image processing and understanding.

Recently, deep convolutional architectures have defined the state-of-the-art on multiple computer vision tasks, such as image classification [1], [3], [4], object detection and segmentation [5], [6], multimedia and cross-media information retrieval [7], [8], and detection of adversarial examples [9].

Despite their success, there are still many open problems with current deep architectures. In fact, it is known that they cannot generalize well to unseen objects and they lack human-like reasoning capabilities.

Final authenticated publication © 2019 IEEE:
https://doi.org/10.1109/CBMI.2019.8877412

Humans, as well as animals, are able to abstract visual perception in order to recognize some shape patterns never seen before. Unlike humans, convolutional networks are able to learn very precise representations, usually very difficult to generalize to abstract concepts.

For this reason, in this work, we concentrate on a very specific task designed to test abstract reasoning capabilities of neural networks. This task is called *same-different* and consists in predicting if two shapes inside the same image are the same or not. It is a challenging task for convolutional architectures since it is required not to learn specific shape patterns in order to solve the problem.

A clear understanding of the *same-different* concept, and the ability to actuate complex relational reasoning are considered key features in many tasks. For example, these concepts may have a great impact in classification problems, in the research of particular patterns in cultural heritage, and in the detection of patterns defining aesthetic beauty in images and even music. In fact, the world is often perceived by humans as a set of recurrent structures composited together, such as the human eyes in a face or the repeating chorus in a song.

In this work we employ a carefully designed dataset, called SVRT [10], containing multiple visual problems pertaining to the *same-different* category. We study the behavior of modern standard visual deep-learning architectures by training and validating them on the SVRT dataset.

More in detail, we extend the study carried out by Kim et al. [11], by testing a variety of state-of-the-art deep convolutional architectures, originally designed for image classification, on some challenging *same-different* problems in the SVRT dataset. In particular, we show that residual networks are able to correctly solve the problem, with a remarkable generalization margin.

We also discovered that different image interpolation algorithms used during the data augmentation process have an important effect on the final performance. This finding remarks the overall vulnerability of neural networks to subtle perturbations of the input image.

Our contribution is two-fold: first, we train state-of-the-art deep image classification networks on the *same-different* task, treating it as a binary classification problem; second, we perform an extensive study upon the generalization capabilities of the probed networks, in order to understand to what

extent current models are able to deal with higher-level visual abstractions.

The rest of the paper is organized as follows: in Section II we review some of the related work concerning abstract reasoning capabilities of deep neural networks; in Section III we go into details with the SVRT dataset and we recall the models that will be probed against the *same-different* problem; in Section IV we will present our experimental setup and we will discuss the obtained results; finally, in Section V we will remark on the importance of these studies and we will propose future directions for this work.

II. Related Work

It has been proven that deep neural networks can obtain remarkable performance on different computer vision fields. However, relatively few works have tackled the cognitive and abstract reasoning capabilities of state-of-the-art deep neural networks.

Recently, the introduction of on-purpose generated benchmarks such as CLEVR and Sort-of-CLEVR [12] has paved the way towards fine-grained studies on relational capabilities of neural networks. In particular, [13] developed a relational reasoning module able to solve the CLEVR visual-question-answering task, by augmenting a standard convolutional network with a reasoning module able to process couples of objects. A slight variation of this network, the 2S-RN [14], has been used for relational content-based image retrieval (R-CBIR), in order to extract relationship-aware visual features for indexing purposes.

Transparency and explainability have been considered key objectives for understanding what kind of reasoning neural networks are internally performing. To this aim, on the CLEVR dataset, [15] used explicit module composition in order to build an explicit reasoning pipeline, while [16] proposed a set of primitives which, when composed, manifest as a model capable of performing complex reasoning tasks in an explicitly-interpretable manner. In particular, [16] reached more than 99% accuracy on the CLEVR test set.

Other than CLEVR, other difficult tasks, such as Raven's Progressive Matrices (RPM), have been proposed to test in great detail, and under fully-controlled environments, reasoning capabilities of a given architecture.

In particular, [17] worked on RPMs and tried to establish a semantic link between vision and reasoning by providing structured representation. Similarly, [18] used an on-purpose created dataset similar to RPM, called Procedurally Generated Matrices (PGM). They demonstrated that popular models such as ResNets perform poorly on this benchmark, and they presented a novel architecture demonstrating quite effective reasoning capabilities. Differently from [18], our work concentrates on a very simple yet challenging problem, the *same-different* challenge.

The work by [19] introduced a synthetic dataset composed of sequences of 2D images in order to test memorization capabilities of neural networks.

[10] introduced SVRT, a simple dataset composed of 2D shapes in order to test comparison and relational capabilities of neural networks. [20] first showed that problems involving comparisons between SVRT shapes were difficult for convolutional architectures like LeNet and GoogLeNet [21]. By contrast, instead, the traditional boosting method used by [10] was able to reach very good results even on comparison problems.

Even [11], in their recent work, found that the *same-different* problem strains simple feed-forward convolutional networks. In this paper we extend their work, giving some insights about the behavior of very-deep convolutional networks on the *same-different* task.

III. Method

This work takes into consideration state-of-the-art deep convolutional neural networks originally developed for image classification tasks. Our aim consists in studying their performance when dealing with higher-level abstraction problems.

First of all, we give some insights into the SVRT benchmark; then, we will discuss the architectures that we will employ to tackle this problem.

A. SVRT Dataset

SVRT [10] is an extensive benchmark designed to test abstract reasoning capabilities of machine learning algorithms. SVRT requires more than trivial local descriptors in order to be solved correctly. It consists of simple 2D images containing simple closed curves. It is clear that brute-force memorization cannot solve the task since the shape of curves is always randomly generated.

SVRT collects 23 different sub-problems, that can be further divided into two clusters: problems related to the spatial arrangement of shapes, and problems regarding comparisons between shapes. The latter set of sub-problems pertain to the *same-different* challenge, and this is the set that we are interested in.

Kim et al. [11] probed simple CNNs upon all the problems in the SVRT dataset, and they found some of them particularly straining for their simple setup. In particular, following the findings by Kim et al., problems 1, 5, 20, 21, are among the most difficult to solve for a convolutional neural network.

In particular, these problems pose the following challenges:

- **Problem 1 (P.1)**: Detecting the very same shapes, randomly placed in the image, but having the same rotation and scale.
- **Problem 5 (P.5)**: Detecting two couples of identical shapes, randomly placed in the image. The two images inside every couple have the same rotation and scale.
- **Problem 20 (P.20)**: Detecting the same shape, translated and flipped along a randomly chosen axis.
- **Problem 21 (P.21)**: Detecting the same shape, randomly translated, rotated and scaled.

All the images from the proposed problems contain two shapes except problem 5, where images contain two couples

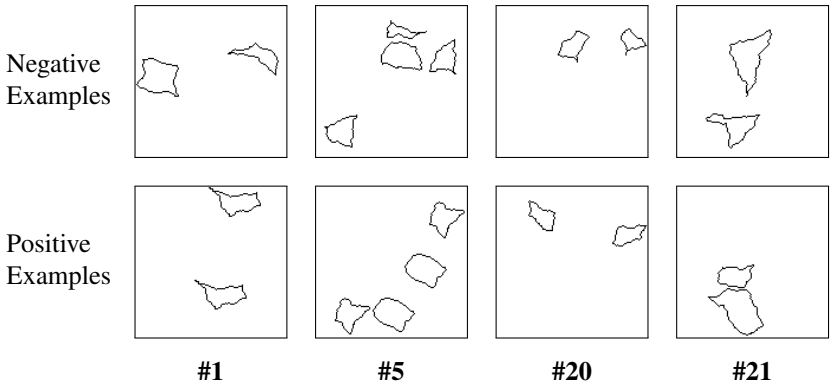


Fig. 1: Positive and negative examples from the four considered SVRT problems.

of shapes. Figure 1 shows some positive and negative examples from each one of the above-listed problems.

In none of these benchmarks are shapes overlapping. They are always well separated. Despite being a quite straightforward dataset, built of simple shapes and with no color information, SVRT can help to highlight some intrinsic limitations of current neural network models.

B. Models

In this work we take into consideration the following state-of-the-art architectures for image classification: AlexNet [1]; VGG19 [4]; three variants of the Resnet [3], in order of increasing complexity ResNet-18, ResNet-34 and ResNet-101; a recently introduced biologically inspired network called CorNet-S [22], and its simpler version, Cornet-Z, a simple feed-forward convolutional network used as baseline.

AlexNet and CorNet-Z are intended to be our baselines. In fact, they are straightforward convolutional networks, with relative few layers with respect to VGG and ResNets. In particular, CorNet-Z is a lightweight version of the AlexNet. It consists of only four convolutional layers, with ReLU activations and MaxPooling, with a single fully-connected layer as a classifier, outputting probabilities for the two target classes.

VGG19 still contains a simple convolutional architecture comparable with the AlexNet structure, but it is significantly deeper. ResNets, on the other hand, introduce residual connections. These skip connections force the architecture to learn incremental differences in the stored representations, refining the information multiple times until it can be used for the downstream task.

Following the work by Kar et al. [23], it seems that ResNets can be considered, from a functional point of view, biologically inspired networks. In fact, they continuously refine the information coming from pixels, just like the visual cortex processes the information flow coming from the eyes. In particular, [23] found some experimental evidence on primates brain claiming that visual cortex could be comprised of recurrent connections. In light of this, they developed CorNet-S [22]. CorNet-S is composed of four blocks, mimicking different brain cortical areas involved with vision; each one of these four blocks contains a recurrent connection together with a skip connection, taking inspiration from ResNets. Basically, ResNets are unrolled versions of CorNet-S. Compared to ResNet101, that has a comparable number of weights with CorNet-S, this model is also lighter to train since, differently from ResNets, recurrent connections make most of neurons weights shared among different timesteps.

IV. Experimental setup

We train the networks discussed in Section III on all the four *same-different* problems from the SVRT dataset. Concerning AlexNet, ResNets, and VGG19, we use the models provided by the PyTorch framework. Instead, for CorNet-S and CorNet-Z, we employ the public available implementation provided by the authors.

For each of the benchmarks, we use 400k training examples and 100k images for testing. All the positive and negative examples are perfectly balanced in both sets. For all the probed models, we use SGD as optimization algorithm, with a momentum of 0.9, weight decay of 1e-4 and an initial learning rate of 0.1. We use an exponential decay schedule for the learning rate, that halves it every 20 epochs. We do not use pre-trained weights if they are available.

A. Experiment 1

a) Description: Our primary objective consists in trying to correctly learn an SVRT problem, measuring the accuracy on the test set of the same problem.

Together with the obtained accuracies, in this experiment we are also interested in understanding what is the convergence speed of the model, as a measure of the strain perceived by the network during the training phase. For this reason, we keep track of the epoch in which the test accuracy reaches 90%. From this moment on, we will refer to this particular point as the *convergence epoch* (CE). For performance reasons we validate the model every half epoch, so we are able to provide the CE with a resolution of 0.5 epochs. This resolution is enough to capture substantial differences in the training curves.

	Problem 1		Problem 5		Problem 20		Problem 21	
Model	Acc. (%)	CE	Acc. (%)	CE	Acc. (%)	CE	Acc. (%)	CE
LeNet [20]	57.0	n.a.	54.0	n.a.	55.0	n.a.	51.0	n.a.
GoogLeNet [20]	50.0	n.a.	50.0	n.a.	50.0	n.a.	51.0	n.a.
AdaBoost [10]	98.0	n.a.	87.0	n.a.	70.0	n.a.	50.0	n.a.
Human [10]	98.0	n.a.	90.0	n.a.	98.0	n.a.	83.0	n.a.
AlexNet	50.0	-	50.0	-	50.0	-	50.0	-
CorNet-Z	50.0	-	50.0	-	50.0	-	50.0	-
VGG-19	50.0	-	50.0	-	50.0	-	50.0	-
ResNet-18	99.0	2.0	**99.8**	2.5	95.8	2.0	96.1	17.5
ResNet-34	99.4	0.5	98.7	1.5	94.6	6.5	**96.9**	13.0
ResNet-101	99.1	3.5	97.8	3.5	**95.9**	4.0	91.1	20.5
CorNet-S	**99.5**	2.0	96.8	2.0	95.3	2.0	95.9	13.0

TABLE I: Accuracy values measured on the probed architectures, for each of the four SVRT problems. The values from the first experiments are reported as they are from [10], [20]. They did not report any convergence information (CE is n.a.).

Together with our measurements on deep convolutional networks, we also report the values as measured by [10], [20] on LeNet, GoogLeNet and AdaBoost (using feature group 3).

Table I summarizes the accuracies reached from all the architectures, over the test sets of the respective problems. For the architectures trained by us, we report also the measured CEs, only when the architecture converged.

Learning rotation invariance is known to be one of the major sources of strain in convolutional networks. For this reason, we also validate ResNet18, ResNet101, and Cornet-S on different instances of the test set, where all the images from the same instance are rotated by the same amount. For these trials, we take as reference the models trained on P.21.

Rotation is inherently lossy for digital images, especially if images do not have high resolution, as in our case. For this reason, we try to rotate images with and without interpolation, in order to appreciate how much the model is robust to the noise added during the transform.

Figure 2 collects accuracy values measured on a particular rotated instance of the test set, with and without pixel interpolation.

b) Discussion: Table I shows that, among all the probed architectures, only ResNets and Cornet-S are able to learn all the four problems correctly, perhaps with a very few error rate, defeating all the state-of-the-art results previously reached with AdaBoost [10].

AlexNet, CorNet-Z, and VGG19 are unable to learn. On the validation set, they remain on the chance level accuracy of 50%. AlexNet and CorNet-Z are not very deep and they are quite straightforward. Hence, the performance obtained with these models seems to be in line with results shown by Kim et al. [11] on their simple convolutional architectures.

One interesting finding is the fact that a very-deep convolutional network, VGG19, is unable to learn. Specifically, it always outputs *same* for all the test samples of all the four problems. This particular outcome confirms the inability of the VGG architecture to discern discriminant data from the training examples.

By contrast, residual networks (ResNet-18, ResNet-34, ResNet-101 and CorNet-S) are able to obtain best performances on all the problems. Even the small ResNet-18 behaves well, both in terms of reached accuracies and CEs. Considering P.1, all the ResNets reach almost perfect performance on the validation set, approaching more than 99% accuracy with pretty fast convergence. Only P.21 seems to cause some strain in the training process since more epochs are required in order for the models to converge.

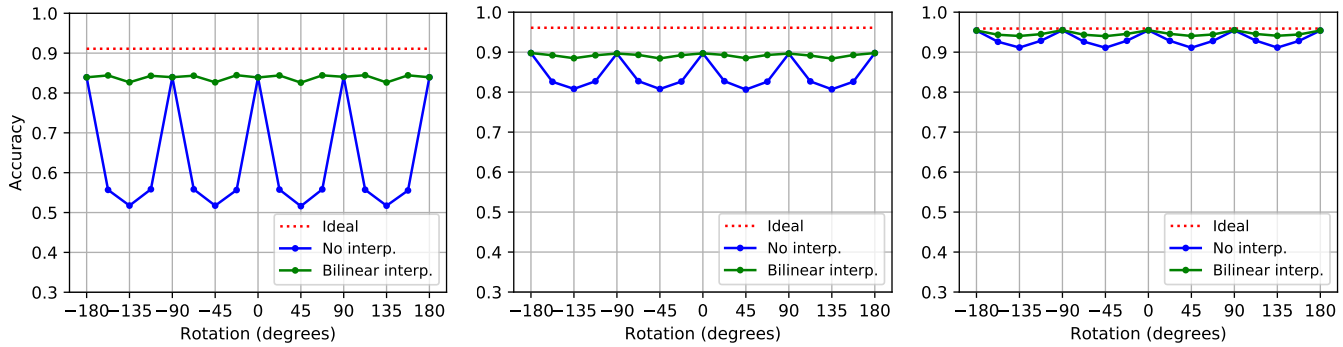
Overall, ResNets and Cornet-S are able to defeat humans on three of the four tasks.

These pieces of evidence on the probed models suggest that the cause of the convergence is not related to the depth of the network, but to its architecture instead. Given that also the very-deep inception architecture of GoogLeNet seems unable to converge [20], the results obtained with ResNets make us realize that their residual architecture may have an important role in the convergence when addressing the *same-different* task.

Looking at Figure 2, it seems that converged models are also able to correctly handle rotated test set images. The peaks at 0, 90, 180 and 270 degrees tell us that the models are robust to rotation, and they are able to generalize to a brand new set of rotated figures never seen before.

On the other hand, the fluctuations in the accuracy at intermediate angles are a clear signal that the noise added while rotating images strains these models. In particular, ResNet18 and ResNet101 show more sensitivity to this noise. ResNet101 is the most vulnerable to this disturbance. Its accuracy falls to full chance when the image is rotated by 45 degrees and no interpolation is used. Resnet18 and Resnet101 suffer in the same way from the padding and center-crop transforms needed in order to preserve shapes during the rotation (see Figure 2 caption for details), meaning that ResNets have also stronger scale-dependent problems with respect to Cornet-S.

By contrast, CorNet-S seems very stable, both to rotation



(a) ResNet101 (b) ResNet18 (c) CorNet-S

Fig. 2: Accuracy values when the test set is rotated by the given angle. *Ideal* refers to the accuracy value as measured on the original validation-set, without the re-scaling operations used for rotating the images. *No interp* is measured by removing the interpolation during the rotation; the shape acquires at most 1-pixel wide distortion; *Bilinear interp.* uses bilinear interpolation during rotation; in this case, rotation artifacts are strongly attenuated. At zero degrees, *Ideal* can diverge from *No interp.* and *Bilinear Interp.*. In fact, in order to be rotated, the image has been first padded and then center-cropped, in order to avoid cutting the shapes during the rotation. These preliminary transforms are responsible for the gap in the case of zero rotation. A gap at zero degrees is a clear signal indicating that the model is strained by this simple pre-scaling procedure.

noise and to image re-scaling.

B. Experiment 2

a) Description: In this scenario, we test generalization capabilities of the converged models by measuring their performance on the test set of other problems. In particular, we set up the following trials:

1) Train on P.21, test on P.20: we test the capability of the model to generalize to mirroring transformations, never seen when training on P.20.
2) Train on P.21, test on P.5: we test the capability of the model to generalize to multiple instances of the *same-different* problem.
3) Train on P.1, test on P.20: we test the capability of the model to generalize to mirroring transformations. It is more challenging with respect to point 1), since P.1 does not even give the possibility to learn the notion of rotation and scale invariance.
4) Train on P.1, test on P.21: we test the capability of the model to generalize to arbitrary rotations and scales.

Table II summarizes the accuracies of all the converged models in the above-proposed configurations.

b) Discussion: Looking at Table II it turns out that almost all the architectures trained on P.21 are also able to correctly solve P.20. This means that P.21 gave the ResNets the ability to correctly understand shape flips, starting from the notion of scale and rotation.

By contrast, there is no notion of rotation or flip invariance in the architectures trained on the simple P.1. This is an expected result since convolutional architectures are not able to natively deal with invariance to affine transformations, with the exception of the plain translation.

Also, tests on P.5 reveal that discerning multiple instances of the *same-different* problem is a difficult task when the architectures have been trained to detect only one instance. This problem requires the models to understand that objects should be clustered into two couples of possibly identical shapes. Instead, an architecture trained on P.1 most likely learned to output a positive result only if all the shapes are equal.

Model	**Train P.21 Test P.20**	**Train P.21 Test P.5**	**Train P.1 Test P.20**	**Train P.1 Test P.21**
ResNet-18	95.9	57.5	**67.8**	**51.8**
ResNet-34	**96.5**	**66.1**	60.1	51.5
ResNet-101	90.5	56.2	57.3	51.3
CorNet-S	95.7	54.7	55.4	51.6

TABLE II: Accuracy values measured on the probed architectures, by training and testing them on different SVRT problems.

V. Conclusions

In this study we tackled the problem of understanding to what extent very-deep convolutional neural networks are able to deal with the *same-different* challenging task. We think that developing abstract and relational abilities of neural networks is an important step towards the achievement of some interesting new tasks, such as the discovery of particular patterns in cultural heritage, or the search for aesthetic beauty patterns in images and even music.

In particular, we stuck to the work by Kim et al. [11]. They showed how the *same-different* problem strain simple feed-forward convolutional neural networks. We extended their study to very-deep state-of-the-art neural networks for image classification.

Considering the SVRT visual challenge, our results show that, despite some difficulties, ResNets and CorNet-S (a biologically-inspired architecture similar to ResNet architecture) are able to correctly understand and generalize to never seen shapes. We also found that Cornet-S is able to reach strong stability with respect to 1-pixel wide noise, in the case of shapes rotation without interpolation.

By contrast, not very-deep models such as AlexNet and CorNet-Z are not able to learn any of the proposed problems. This evidence is aligned with Kim et al. findings.

However, with the evidence that also deep neural networks such as VGG19 and GoogLeNet are not able to converge, we hypothesized that the residual connections in the ResNet architecture are important key points for the successful convergence of these models.

There are still many open problems in this research direction. As of now, abstract reasoning capabilities of neural networks are tested on very simple and low distribution variability datasets, such as SVRT or Sort-of-CLEVR. One of the important steps to enhance this research would be to consider more complex images collecting real-world shapes, possibly seen from different perspective angles.

In our work, we tested CorNet-S, a model that directly draws inspiration from neuroscientific evidence on the primates brain. We think that it would be interesting to access more neuroscience research, in order to understand what are the functional components of the human brain that contribute to complex abstract reasoning.

ACKNOWLEDGMENTS

This work was partially supported by the AI4EU project, funded by the EC (H2020 - Contract n. 825619), and Automatic Data and documents Analysis to enhance human-based processes (ADA), CUP CIPE D55F17000290009.

We also gratefully acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 GPU used for this research.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105.

[2] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov 1998.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[4] S. Liu and W. Deng, "Very deep convolutional neural network based image classification using small training sample size," in *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, Nov 2015, pp. 730–734.

[5] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *CoRR*, vol. abs/1804.02767, 2018.

[6] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2980–2988.

[7] F. Carrara, A. Esuli, T. Fagni, F. Falchi, and A. M. Fernández, "Picture it in your mind: Generating high level visual representations from textual descriptions," *Information Retrieval Journal*, vol. 21, no. 2-3, pp. 208–229, 2018.

[8] L. Vadicamo, F. Carrara, A. Cimino, S. Cresci, F. Dell'Orletta, F. Falchi, and M. Tesconi, "Cross-media learning for image sentiment analysis in the wild," in *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, Oct 2017, pp. 308–317.

[9] F. Carrara, F. Falchi, R. Caldelli, G. Amato, R. Fumarola, and R. Becarelli, "Detecting adversarial example attacks to deep neural networks," in *Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing*. ACM, 2017, pp. 38:1–38:7. [Online]. Available: http://doi.acm.org/10.1145/3095713.3095753

[10] F. Fleuret, T. Li, C. Dubout, E. K. Wampler, S. Yantis, and D. Geman, "Comparing machines and humans on a visual categorization test," *Proceedings of the National Academy of Sciences*, vol. 108, no. 43, pp. 17 621–17 625, 2011.

[11] J. Kim, M. Ricci, and T. Serre, "Not-so-CLEVR: Visual relations strain feedforward neural networks," 2018. [Online]. Available: https://openreview.net/forum?id=HymuJz-A-

[12] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. Lawrence Zitnick, and R. Girshick, "Clevr: A diagnostic dataset for compositional language and elementary visual reasoning," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[13] A. Santoro, D. Raposo, D. G. T. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. P. Lillicrap, "A simple neural network module for relational reasoning," *CoRR*, vol. abs/1706.01427, 2017. [Online]. Available: http://arxiv.org/abs/1706.01427

[14] N. Messina, G. Amato, F. Carrara, F. Falchi, and C. Gennaro, "Learning relationship-aware visual features," in *Computer Vision – ECCV 2018 Workshops*, L. Leal-Taixé and S. Roth, Eds. Cham: Springer International Publishing, 2019, pp. 486–501.

[15] J. Johnson, B. Hariharan, L. van der Maaten, J. Hoffman, L. Fei-Fei, C. Lawrence Zitnick, and R. Girshick, "Inferring and executing programs for visual reasoning," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[16] D. Mascharka, P. Tran, R. Soklaski, and A. Majumdar, "Transparency by design: Closing the gap between performance and interpretability in visual reasoning," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[17] C. Zhang, F. Gao, B. Jia, Y. Zhu, and S.-C. Zhu, "RAVEN: A Dataset for Relational and Analogical Visual rEasoNing," *arXiv e-prints*, Mar 2019.

[18] D. Barrett, F. Hill, A. Santoro, A. Morcos, and T. Lillicrap, "Measuring abstract reasoning in neural networks," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. Stockholmsmssan, Stockholm Sweden: PMLR, 10–15 Jul 2018, pp. 511–520.

[19] G. R. Yang, I. Ganichev, X.-J. Wang, J. Shlens, and D. Sussillo, "A dataset and architecture for visual reasoning with a working memory," in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, 2018, pp. 729–745.

[20] S. Stabinger, A. Rodríguez-Sánchez, and J. Piater, "25years of cnns: Can we compare to human abstraction capabilities?" in *Artificial Neural Networks and Machine Learning – ICANN 2016*, A. E. Villa, P. Masulli, and A. J. Pons Rivero, Eds. Cham: Springer International Publishing, 2016, pp. 380–387.

[21] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[22] J. Kubilius, M. Schrimpf, A. Nayebi, D. Bear, D. L. K. Yamins, and J. J. DiCarlo, "Cornet: Modeling the neural mechanisms of core object recognition," *bioRxiv*, 2018.

[23] K. Kar, J. Kubilius, K. Schmidt, E. B. Issa, and J. J. DiCarlo, "Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior," *Nature Neuroscience*, 2019. [Online]. Available: https://doi.org/10.1038/s41593-019-0392-5