Right for the Right Concept: Revising Neuro-Symbolic Concepts by Interacting with their Explanations

Technical Appendix

1 CLEVR-Hans data set

For CLEVR-Hans classes for which class rules contain more than three objects, the number of objects to be placed per scene was randomly chosen between the minimal required number of objects for that class and ten, rather than between three and ten, as in the original CLEVR data set.

Each class is represented by 3000 training images, 750 validation images, and 750 test images. The training, validation, and test set splits contain 9000, 2250, and 2250 samples, respectively, for CLEVR-Hans3 and 21000, 5250, and 5250 samples for CLEVR-Hans7. The class distribution is balanced for all data splits.

CLEVR-Hans7 The first, second, and seventh class rules of CLEVR-Hans7 correspond to classes one, two, and three of CLEVR-Hans3. Images of the third class of CLEVR-Hans7 contain a small cyan object in front of two red objects. The cyan object is a small cube in all images of the training and validation set, yet it can be any shape and size within the test set. Images of the fourth class contain at least five small objects. One of these must be green, one brown, and one purple. There are no constraints on the remaining small objects. This class is not confounded. Images of class five consist of two rules. There are three spheres present in the left half of the image (class rule 5a), or there are three spheres present in the left half of the image and three metal cylinders in the right half of the image (class rule 5b). Within all data splits, including the test split, class rule 5a occurs 90% of the time and class rule 5b 10% of the time. The class rule of the sixth class is contained in class rule 5b, namely three metal cylinders in the right half of the image. This is the same for all splits.

Preprossing details We downscaled the CLEVR-Hans images to visual dimensions 128 x 128 and normalized the images to lie between -1 and 1. For training the Slot-Attention module, an object is represented as a vector of binary values for the shape, size, color, and material attributes and continuous values between 0 and 1 for the x, y, and z positions. We refer to [5] for more details.

2 ColorMNIST Experiment

The model used for the ColorMNIST data set is described in Tab 1.

This model was trained with an initial learning rate of 1.0 for 14 epochs with a batch size of 64 using a step learning rate scheduler with step size 1 and $\gamma = 0.7$ and Adadelta [8] as optimizer.

Type	Size/Channels	Activation	Comment
Conv 3 x 3	32	ReLu	stride 1
Conv 3 x 3	64	ReLu	stride 1
AdaptiveAvgPool (2D)	14×14	-	-
Dropout	-	-	p = 0.25
Flatten	-	-	dim = 1
Linear	128	-	-
Dropout	-	-	p = 0.5
Linear	10	-	-

Table 1: CNN used for ColorMNIST experiments.

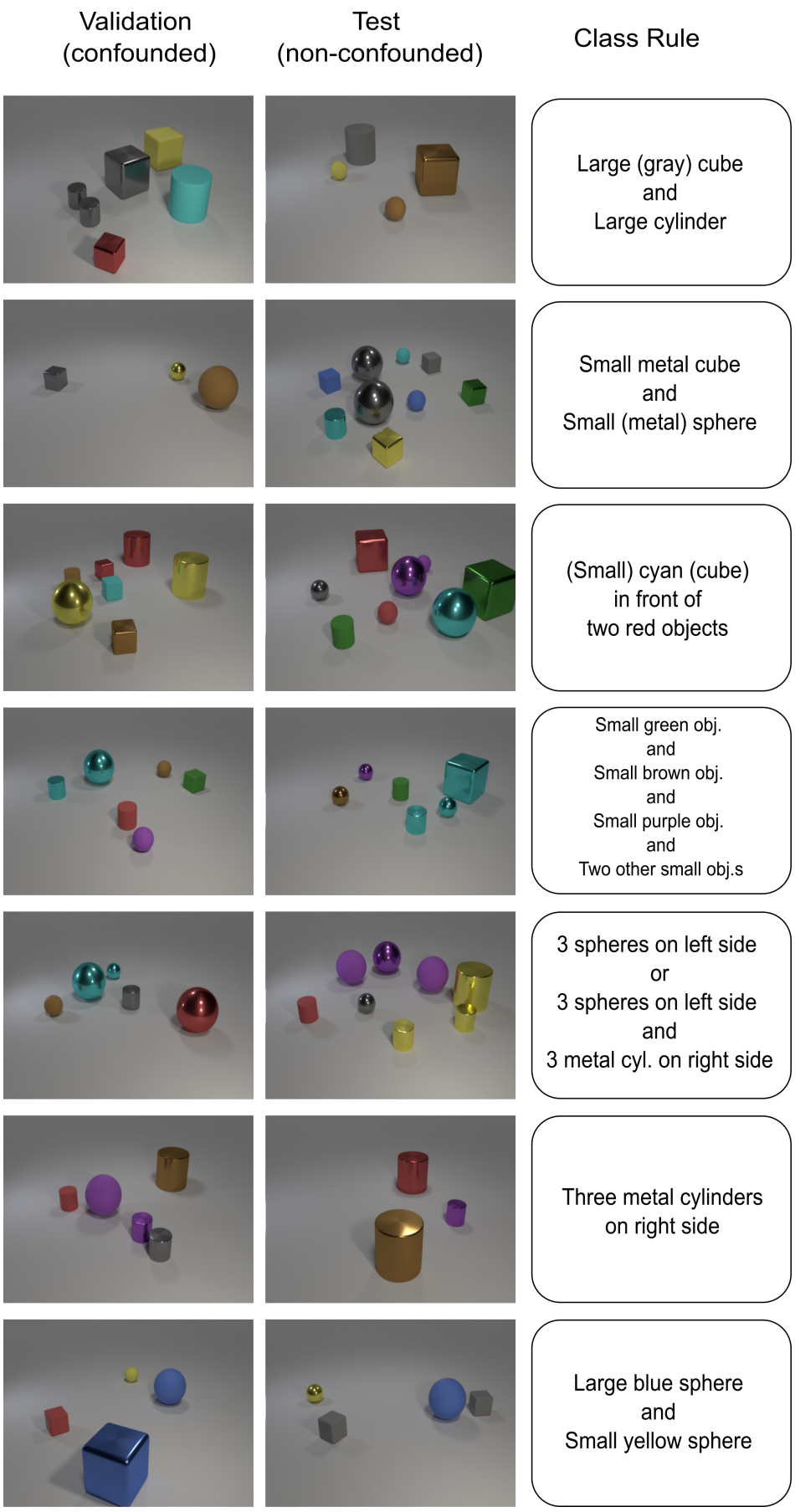


Figure 1: **CLEVR-Hans7 data set overview.** Please refer to the main text for a more detailed description of the data set.

Type	Dim Out	Numb. Heads	Comment
SAB	128	4	-
SAB	128	4	-
Dropout	-	-	p = 0.5
PMA	128	4	-
Dropout	-	-	p = 0.5
Linear	3/7	-	-

Table 2: Set Transformer architecture used for reasoning module. Depending on whether CLEVR-Hans3 or CLEVR-Hans7 was used the final output varied between 3 and 7.

3 Experiment and Model Details

Cross-validation We ran all experiments with five random parameter initializations and reported the mean classification accuracy with standard deviation over these runs. We used the seeds: 0, 1, 2, 3, 4.

Reasoning Module For our reasoning module, we used the recently proposed Set Transformer, an attention-based neural network designed to handle unordered sets. Our implementation consists of two stacked Set Attention Blocks (SAB) as encoder and a Pooling by Multihead Attention (PMA) decoder. Architecture details can be found in Tab 2

Concept Embedding Module For our concept embedding module, we used the set prediction architecture of Locatello *et al.* [5] that the authors had used for the experiments on the original CLEVR data set. We refer to their paper for architecture parameters and details rather than duplicating these here.

We pre-trained this set prediction architecture on the original CLEVR data set with a cosine annealing learning rate scheduler for 2000 epochs, minimum learning rate $1e-5$, initial learning rate $4e-4$, batch size 512, 10 slots, 3 internal slot-attention iterations and the Adam optimizer [3] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, $\epsilon = 1e-08$ and zero weight decay.

Neuro-Symbolic Concept Learner To summarize, we thus have the two modules, as stated above. For our experiments, we passed an image through the pre-trained concept embedding module. For simplicity, we binarized the output of the concept embedding module for the attributes shape, size, and color, before passing it to the reasoning module by computing the argmax of each attribute group. This way, each object is represented by a one-hot encoding of each of these attributes.

The architecture parameters of the concept embedding and reasoning module were as stated above, and the same for both training settings, i.e., default and XIL.

In the default training setting, using the cross-entropy classification loss, we used the Adam optimizer ($\beta_1 = 0.9$ and $\beta_2 = 0.999$, $\epsilon = 1e-08$ and zero weight decay) in combination with a cosine annealing learning rate scheduler with initial learning rate $1e-4$, minimal learning rate $1e-6$, 50 epochs and batch size of 128.

For training our concept learner using the HINT [7] loss term on the symbolic explanations in addition to cross entropy term we used the Adam optimizer ($\beta_1 = 0.9$ and $\beta_2 = 0.999$, $\epsilon = 1e-08$ and zero weight decay) in combination with a cosine annealing learning rate scheduler with initial learning rate $1e-3$, minimal learning rate $1e-6$, 50 epochs and batch size of 128. We used $\lambda_s = 1000$ for the XIL experiments on CLEVR-Hans3 and $\lambda_s = 10$ for the XIL experiments on CLEVR-Hans7. For the global rule experiments, using the RRR term of Ross et al. [6], we set $\lambda_s = 20$ with all other hyperparameters the same as previously.

CNN Model Details Our CNN model is based on the popular ResNet34 model of [1]. The visual explanations generated by Grad-CAM are in the visual dimensions of the hidden feature maps. As these dimensions of the ResNet34 model were very coarse given our data pre-processing, we decreased the number of layers of the ResNet34 model by removing the last six convolutional layers (i.e., fourth of the four ResNet blocks) and adjusting the final linear layer accordingly.

For training the CNN in default cross-entropy mode, we used a constant learning rate of $1e-4$ for 100 epochs and a batch size of 64. We used the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, $\epsilon = 1e-08$ and zero weight decay. For training the CNN with an additional HINT explanation regularization, we used the same training parameters, as in the default case, and a $\lambda_v = 10$. These parameters were the same for CLEVR-Hans3 and CLEVR-Hans7.

4 Explanation Loss Terms

For our experiments, we used two different types of explanation loss terms (Eq. 4). For all experiments, apart from those with a single global rule, we simulated the user feedback as positive feedback. In other words, the user feedback indicated what features the model should be focusing on. For simplicity in our experiments, we simulated the user to have full knowledge of the

Model	Global Test Average	Class 1	Class 2	Class 3
NeSy (Default)	4.99 ± 0.16	6.57 ± 0.95	4.58 ± 0.68	3.81 ± 0.34
NeSy XIL	$\mathbf{1.37 \pm 0.2}$	$\mathbf{2.37 \pm 0.62}$	$\mathbf{1.09 \pm 0.08}$	$\mathbf{0.68 \pm 0.1}$
True Positive Rate				
NeSy (Default)	2.56 ± 0.05	2.97 ± 0.27	2.13 ± 0.23	2.57 ± 0.2
NeSy XIL	$\mathbf{0.85 \pm 0.09}$	$\mathbf{1.26 \pm 0.31}$	$\mathbf{0.77 \pm 0.06}$	$\mathbf{0.52 \pm 0.12}$

Table 3: **L1 error between symbolic user feedback (i.e. ground-truth (GT) symbolic explanations) and the respective model's symbolic explanations for CLEVR-Hans3.** Presented are the average L1 error over all samples of the test set and the average L1 error separately over all samples of individual classes. Note: a lower value is preferable. The best (lowest) errors are in bold. The first two rows present the L1 error over all classification errors. The bottom two rows present the error by comparing only for relevant GT elements (i.e. have a value of one).

Model	Global Test Average	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7
NeSy (Default)	7.26 ± 0.32	6.14 ± 0.65	7.61 ± 0.66	6.64 ± 0.33	6.93 ± 1.83	10.1 ± 0.45	7.66 ± 0.73	5.77 ± 0.51
NeSy XIL	$\mathbf{3.35 \pm 0.13}$	$\mathbf{2.28 \pm 0.09}$	$\mathbf{2.86 \pm 0.08}$	$\mathbf{4.72 \pm 0.72}$	$\mathbf{1.88 \pm 0.27}$	$\mathbf{7.19 \pm 0.45}$	$\mathbf{2.90 \pm 0.24}$	$\mathbf{1.59 \pm 0.09}$
True Positive Rate								
NeSy (Default)	4.55 ± 0.12	3.35 ± 0.27	4.12 ± 0.26	4.41 ± 0.16	3.13 ± 0.55	7.6 ± 0.08	5.57 ± 0.38	3.65 ± 0.22
NeSy XIL	$\mathbf{2.43 \pm 0.12}$	$\mathbf{1.38 \pm 0.04}$	$\mathbf{1.87 \pm 0.11}$	$\mathbf{3.22 \pm 0.51}$	$\mathbf{1.48 \pm 0.26}$	$\mathbf{6.03 \pm 0.5}$	$\mathbf{2.00 \pm 0.28}$	$\mathbf{1.04 \pm 0.06}$

Table 4: **L1 error between symbolic user feedback (i.e. ground-truth (GT) symbolic explanations) and the respective model's symbolic explanations for CLEVR-Hans7.** Presented are the average L1 error over all samples of the test set and the average L1 error separately over all samples of individual classes. Note: a lower value is preferable. The best (lowest) errors are in bold. The first two rows present the L1 error over all classification errors. The bottom two rows present the error by comparing only for relevant GT elements (i.e. have a value of one).

task and give the fully correct rules or visual regions as feedback. For this positive feedback, we applied a simple mean-squared error between the model explanations and user feedback as an explanation loss term:

$$L(\theta,\ X,\ y,\ A) = \lambda_1 \frac{1}{N} \sum_{i=1}^{N} \sum_{d=1}^{D} (A_{id} - \hat{e}_{id}^{g})^2 \tag{1}$$

This was applied to the XIL experiments with the standard CNN model, for which the explanations were in the form of Grad-CAMs, and for revising the Neuro-Symbolic model. In the case of revising the CNNs, the user annotation masks were downscaled to match the Grad-CAM size resulting from the last hidden layer of the CNN.

For handling the negative feedback of the experiments with the single global rule, in which the user indicated which features are not relevant, rather than which are, we reverted to the RRR term of Ross et al. [6]:

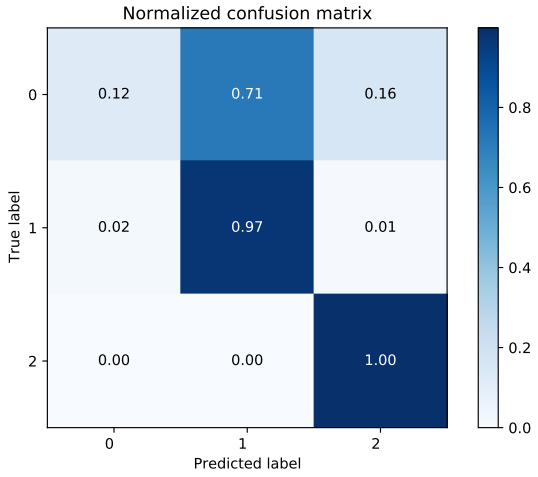
$$L(\theta,\ X,\ y,\ A) = \lambda_1 \sum_{i=1}^{N} \sum_{d=1}^{D} \left(A_{id} \frac{\delta}{\delta \hat{z}_{id}} \sum_{k=1}^{N_c} \log(\hat{y}_{ik}) \right)^2 \tag{2}$$

5 Quantitative Analysis of Improved Symbolic Explanations

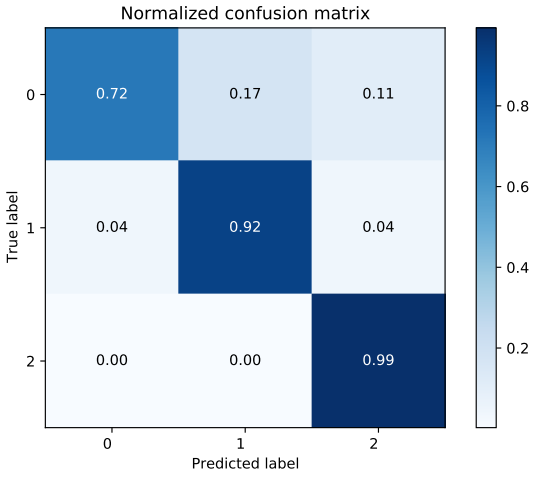
To more quantitatively assess the improvements of the symbolic explanations of our NeSy model using XIL we measured the absolute difference (L1 error) for each sample between the ground-truth (GT) explanations and the symbolic explanations of the NeSy Default trained with cross-entropy and NeSy XIL, respectively. Specifically, we computed the difference for an individual sample in the following. Given the GT explanation $e_i^{GT} \in [0,\ 1]^D$ and symbolic explanation of the model $\hat{e}_i^g \in [0,\ 1]^D$ of sample i we computed the L1 error as: $\sum_j^D |e_{ij}^{GT} - \hat{e}_{ij}^g|$. We finally averaged the error over all samples of the test set, as well as all samples of a specific sample class, separately.

Due to that within e_i^{GT} only few attributes are marked as relevant (i.e. have a value of one) we measured the absolute L1 error here over all possible classification errors, i.e. true positives, true negatives, false positives and false negatives. The results can be found in the top two rows of Tab. 3 and Tab. 4 for CLEVR-Hans3 and CLEVR-Hans7, respectively. Note here that a lower error corresponds to a stronger correspondence between the GT explanation and model explanation.

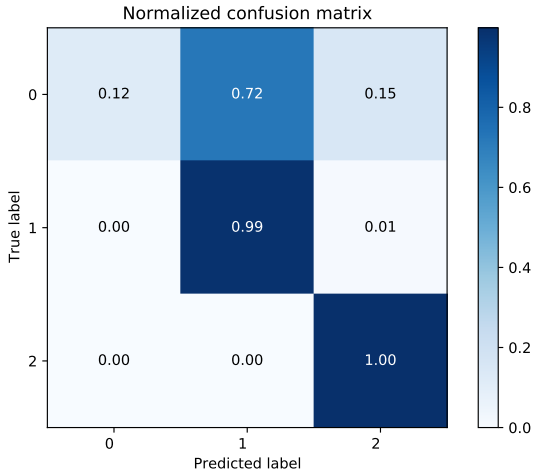
Additionally we computed the absolute L1 error only over the relevant GT attributes, yielding the true positive rate. The results can be found in the bottom two rows Tab. 3 and Tab. 4 for CLEVR-Hans3 and CLEVR-Hans7, respectively. One can observe that in fact with XIL the symbolic model explanations more strongly correspond to the GT explanations, thus further



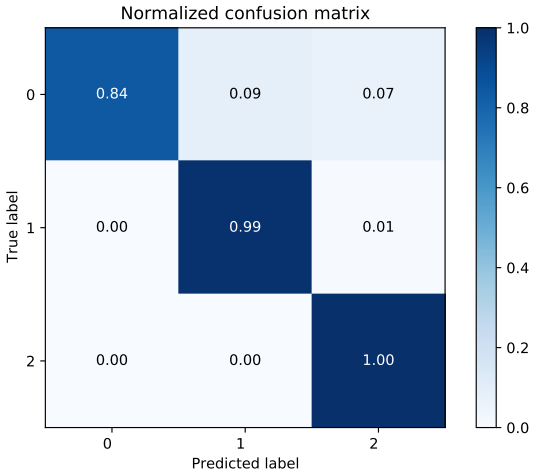
(a) CNN Default



(b) NeSy Default



(c) CNN XIL



(d) NeSy XIL

Figure 2: **Confusion matrices of the different models and training settings for the test set of CLEVR-Hans3.** Note: label 0 corresponds to class 1, etc..

supporting the results indicated by the balanced accuracies for validation and test sets of the main text as well as the qualitative results of the main text and supplementary materials that using XIL on the symbolic explanations the model explanations could be improved to more strongly correspond to the underlying GT symbolic explanations.

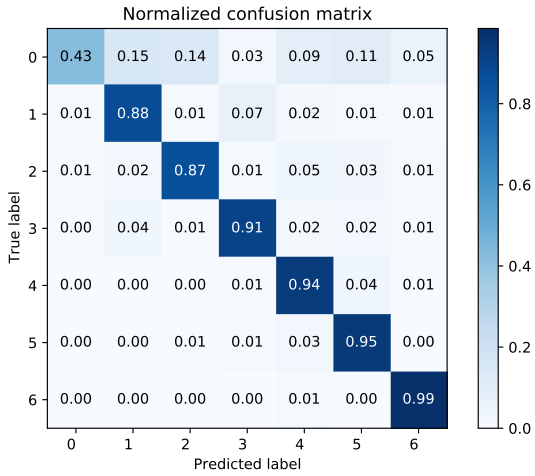
For CLEVR-Hans7 NeSy XIL resulted in a reduction in relative L1 error compared to NeSy (Default) of: 54% (total), 63% (class 1), 62% (class 2), 29% (class 3), 73% (class 4), 29% (class 5), 62% (class 6) and 72% (class 7).

One particularly interesting result to take from Tab. 4 is the difficulty of improving the symbolic explanations for those classes of CLEVR-Hans7 which require counting the occurrences of specific attribute combinations, i.e. classes 3 and 5 (see Fig. 1 for an overview of the class rules). The improvement in L1 error for NeSy XIL is not as strong for class 3 and class 5 as for the other classes. We believe this to indicate a shortcoming in the counting ability of the Set Transformer module.

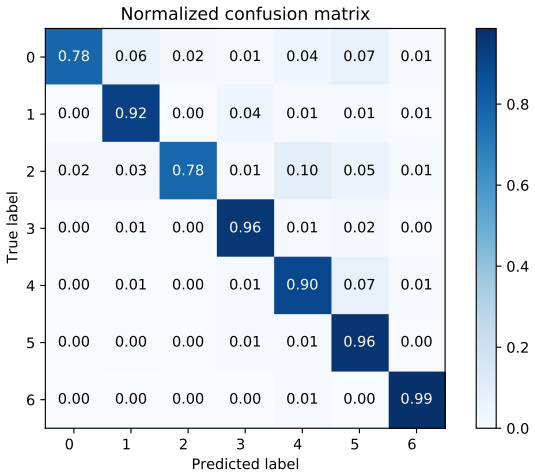
6 Detailed Analysis of Confusion Matrices

Fig. 2 presents the confusion matrix for the all model and training settings on the test set of CLEVR-Hans3. Note the default CNN's difficulty especially with the color confounder of class one rather than the material confounder of class two.

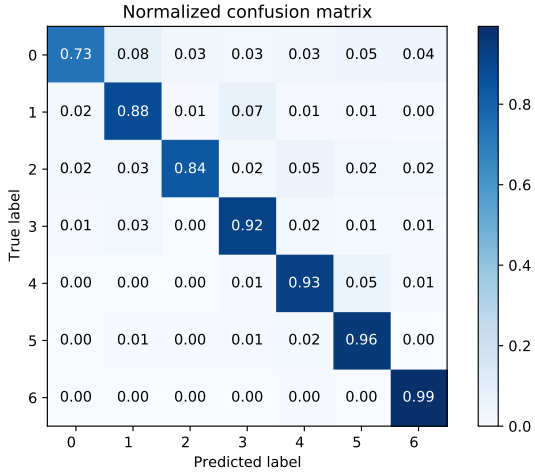
Fig. 3 presents the confusion matrix for the all model and training settings on the test set of CLEVR-Hans7. Quite surprisingly, in comparison to Fig. 2 one can see that within the classes also present in CLEVR-Hans3 all models reach a higher class accuracy than when trained with CLEVR-Hans3. We suggest this is caused by the nonexclusive nature of the CLEVR-Hans data generation. As an example: though a large gray cube and large cylinder will never appear in combination in any other image than of class 1, each object separately may appear in images of other classes. Thus with more images available in which an individual large gray cube may appear, the confounding factor, the color gray, may not carry as much weight as with fewer



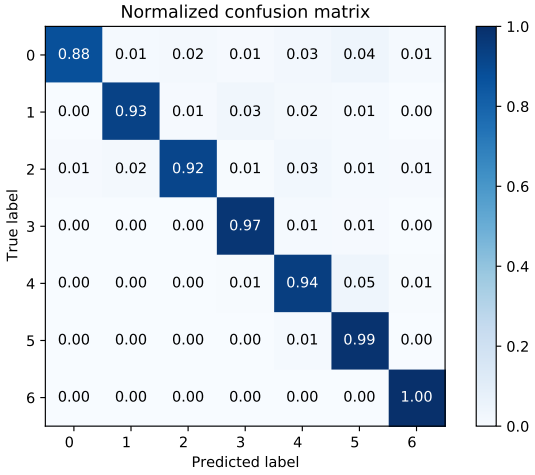
(a) CNN Default



(b) NeSy Default



(c) CNN XIL



(d) NeSy XIL

Figure 3: **Confusion matrices of the different models and training settings for the test set of CLEVR-Hans3.** Note: label 0 corresponds to class 1, etc..

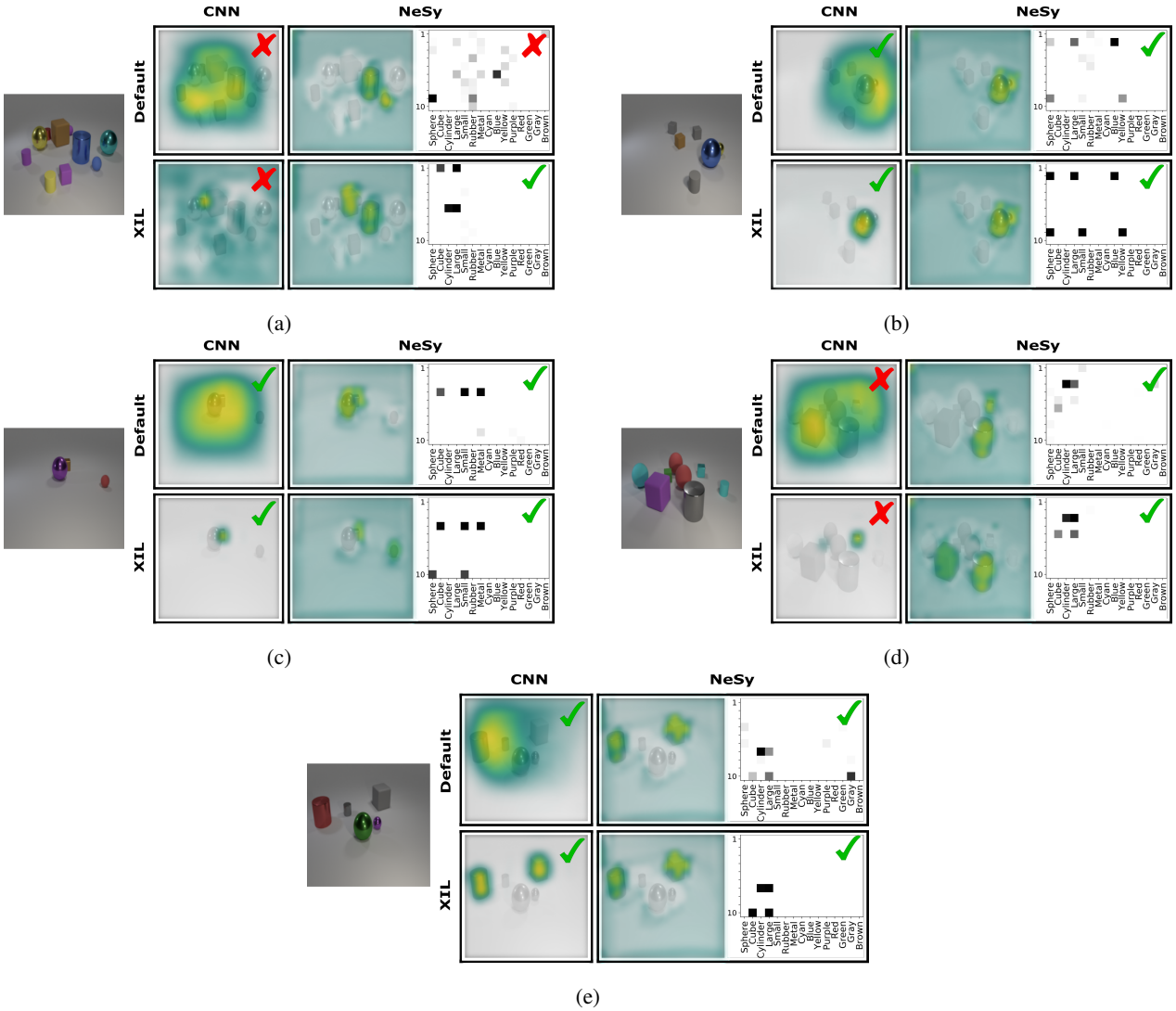


Figure 4: Additional explanations of the various model types for test samples. Green checks represent correct class predictions, red crosses incorrect predictions.

classes and images. Thus the generalizability to the test set is from the start easier to accomplish with CLEVR-Hans7.

7 Additional Explanation Visualizations

Fig. 4 shows additional qualitative results of NeSy XIL in addition to those of the main text. The top left example (a) presents another example where only via interacting with Neuro-Symbolic explanations can get the correct prediction for the correct reason. Top right (b) shows an example where all model configurations make the correct prediction. However, it does not become clear whether the CNN is indeed focusing on both relevant objects. With the NeSy model, this becomes clearer, though only using NeSy XIL are the correct objects and attributes identified as relevant for prediction. A similar case can be found in the middle left (c), where NeSy XIL aids in focusing on both relevant objects. The middle right shows a case where already NeSy shows advantages for creating correct predictions, yet not entirely for the correct concept. The bottom example (e) exemplifies that solely from a visual explanation, it does not become clear that the model is focusing on the color confounder, gray.

8 Further Concluding Remarks

The presented CLEVR-Hans benchmarks are challenging data sets due to the complex logic concepts that underlie the visual scenes, we also strive towards an evaluation on real world data sets. Since, Koh et al. [4] and Kim et al. [2] show that the performance of concept based models on real world data sets are en par with popular black-box models —however, don't investigate revising these models— we expect good results here as well.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 770–778. IEEE Computer Society, 2016.

[2] Kim, Kim, and Bengio. Visual concept reasoning networks. *arXiv preprint*, 2020.

[3] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[4] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning*. PMLR, 2020.

[5] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 11525–11538. Curran Associates, Inc., 2020.

[6] Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations. In *Proceedings of International Joint Conference on Artificial Intelligence IJCAI*, pages 2662–2670, 2017.

[7] Ramprasaath Ramasamy Selvaraju, Stefan Lee, Yilin Shen, Hongxia Jin, Shalini Ghosh, Larry P. Heck, Dhruv Batra, and Devi Parikh. Taking a HINT: leveraging explanations to make vision and language models more grounded. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV*, pages 2591–2600. IEEE, 2019.

[8] Matthew D. Zeiler. ADADELTA: an adaptive learning rate method. *CoRR*, abs/1212.5701, 2012.