Semantics-Informed Group Interventions for Concept Bottleneck Models

Michał Mikuta [*1] **Mikael Makonnen** [*1] **Sari Issa** [*1] **Max Buckley** [*2]

Abstract

Concept Bottleneck Models (CBMs) enhance interpretability by predicting human-understandable concepts that directly inform the final predictions. Interventions allow for manual corrections of predicted concepts. This work introduces a semantics-informed clustered realignment framework for CBMs, leveraging hierarchical clustering to constrain intervention realignment within semantically meaningful groups. The approach provides interpretability guarantees. Model performance remains comparable with the baselines considered. Furthermore, the clustered-realignment method demonstrates robustness to noisy concepts by isolating them in single clusters.

1. Introduction

Concept bottleneck models (CBMs) (Koh et al., 2020) are neural networks designed to enhance interpretability by incorporating human-understandable concepts. CBMs are typically trained on triplet data $(\mathbf{x}, \mathbf{c}, \mathbf{y})$, comprising covariates $\mathbf{x} \in \mathcal{X}$, targets $\mathbf{y} \in \mathcal{Y}$, and human-annotated concepts $\mathbf{c} \in \mathcal{C}$. In CBMs, the input features $\mathbf{x}$ are mapped to a predicted set of concepts $\hat{\mathbf{c}} = h_{\theta_1}(\mathbf{x})$. The map $g_{\theta_2}(\hat{\mathbf{c}})$ is used to predict the output $\hat{\mathbf{y}}$. Nevertheless, the nature of using human-annotated concepts, makes this task difficult (Marcinkevičs et al., 2024). In general g_{θ_2} is a simple function preserving interpretability.

CBMs allow for interventions at test time, wherein faulty concept values are manually corrected by experts to improve downstream predictions. This constitutes a key advantage of this architecture (Koh et al., 2020; Marcinkevičs et al., 2024; Shin et al., 2023; Vandenhirtz et al., 2024). Concepts are usually correlated (Vandenhirtz et al., 2024), so changing only a single entry would not capture the entire information encoded in the intervention. Moreover, single interventions are time-consuming and have a small effect on predictions.

To address this Singhi et al. (2024) propose a concept realignment model, capturing correlations and adjusting the entire concept vector. We argue, that their approach could obscure the model's reasoning and lose semantic structure and interpretability. This occurs as the extent of realignment is not bounded across the $\mathbf{c}$, allowing re-alignment of unrelated concepts.

We propose a semantics-informed clustering based approach instead. We apply Divisive Analysis Clustering (DIANA) (Kaufman & Rousseeuw, 1990), which we argue produces a semantically-informed clustering. Concept realignment is subsequently constrained within clusters, bounding the effect of interventions and maintaining a high level of interpretability. We apply our approach to the same datasets used in Singhi et al. (2024) to establish clear comparisons (Wah et al., 2011; Liu et al., 2018; Xian et al., 2018).

We show that our model performs similarly to an Uncertainty-based Concept Picking (UCP) policy without realignment and over-performs the end-to-end CBM prediction baseline. As expected we under-perform with respect to the unconstrained realignment from Singhi et al. (2024). However, we argue that our model guarantees additional interpretability via semantically-bounded realignment. We implement ablations on the realignment architecture and the number of maximum interventions. The conclusions regarding the relative performance of clustered realignment remain consistent. A placebo test, which appends Bernoulli noise to the ground-truth concepts showcased robustness to noise, by effectively isolating noise in a single cluster. Our implementation can be viewed here.

2. Models and Methods

2.1. Datasets

Although our initial research design encompassed the datasets utilized in Singhi et al. (2024) - namely CUB, CelebA, and AwA2 (Wah et al., 2011; Liu et al., 2018; Xian et al., 2018) - we ultimately concentrated our investigation on the CUB dataset to ensure methodological rigor and robust empirical findings. This strategic focus stemmed from structural incompatibilities in the alternative datasets and computational constraints inherent in processing substantially larger datasets. Table A1 shows the long training

[*]Equal contribution [1]Department of Mathematics, ETH Zurich [2]Google Zurich. Correspondence to: Michał Mikuta <mmikuta@ethz.ch>.

times and extreme accuracies making realignment testing unsuitable. The training of DIANA is much cheaper, therefore the clustering performance is displayed on all datasets nonetheless.

The Caltech-UCSD Birds-200-2011 (CUB) dataset (Wah et al., 2011) comprises 11,788 avian images spanning 200 subcategories, with 5,994 samples allocated to training and 5,794 to testing. Each image is annotated with both a category label and 312 binary attributes serving as concepts.

The CelebA dataset Liu et al. (2018) consists of 202,599 facial images representing 10,177 distinct celebrities, annotated with 40 binary concepts. While suitable for CBM training, the dataset presents significant methodological challenges for celebrity identity classification tasks. The high cardinality of classes (exceeding 10,000) is compounded by sparse representation, with some classes containing only a single instance. Furthermore, the concept-label relationship exhibits considerable variability - while certain attributes like "pointy nose" demonstrate relative consistency across a given identity, others such as "glasses" or "wavy hair" show substantial intra-subject variation.

The AwA2 dataset Xian et al. (2018) contains 37,322 images distributed across 50 animal categories. Its conventional experimental protocol restricts training to 40 classes while reserving 10 classes exclusively for testing. This configuration presents inherent limitations for decoder and realignment training, as the absence of examples from test classes precludes direct prediction. The standard methodological approach involves training a concept classifier on the 40 training classes while separately developing a decoder to map from human-annotated concepts to test classes. The resulting system integrates the training data-derived concept predictor with the test class-specific decoder.

2.2. Concept Bottleneck Model

Our implementation closely follows the architecture proposed by Koh et al. (2020), utilizing a pre-trained ResNet18 encoder (He et al., 2015; Vandenhirtz et al., 2024) to map image inputs to the concept bottleneck layer. This bottleneck layer employs a sigmoid activation function, constraining all output values to the interval $(0, 1)$.

For class prediction, we implement a compact multilayer perceptron (MLP) with batch normalization as the decoder, mapping the learned concepts to class labels while preserving interpretability. The final classification is determined by the maximum logit value, without applying softmax transformation. We adopted an independent training strategy, which had more stable performance, wherein the encoder is trained to completion before the decoder is trained using ground truth concepts as input.

The training protocol consisted of 150 epochs for both the image-to-concept encoder and the concept-to-class decoder, executed on an NVIDIA A100 GPU (40GB) via Google Colab. We employed a batch size of 512 and implemented an adaptive training strategy: when the validation loss exhibited no improvement over fifteen consecutive evaluation epochs, the model weights were restored to the best performing checkpoint and training continued with the learning rate reduced by a factor of 0.5.

2.3. Divisive Analysis Clustering (DIANA)

Palumbo et al. (2024) and Lipton (2018); Marcinkevičs et al. (2024) argue that hierarchical clustering algorithms reflect the inherent organization of real data and provide more informative and interpretable representations, respectively. We use DIANA, a divisive hierarchical clustering approach which employs a top-down approach (Kaufman & Rousseeuw, 1990). All data-points are initially assigned to a global cluster and subsequently splitting is performed on a given cluster.

The choice of DIANA is motivated by the ability to predefine the number of clusters, as opposed to typical hierarchical methods, where the number of clusters is inferred (Palumbo et al., 2024). We need guarantees that the resulting clusters are not too granular to prevent overfitting, so the flexibility of choosing the number of resulting clusters is required.

2.4. Clustered Realignment Module

Based on the realignment module introduced in Singhi et al. (2024), we propose a clustered realignment module where the effect of interventions is constrained to clusters. (Iterative) interventions $\mathcal{S} = \mathcal{S}_t \subseteq \{1, \dots, k\}$ generate an initially updated $\mathbf{c}' = \{c_{\mathcal{S}}, c_{\setminus\mathcal{S}}\}$, where $c_{\mathcal{S}}$ are assumed to be the adjusted values provided by the intervention. Denote the per-cluster interventions $\mathcal{S}^l = \mathcal{S} \cap \mathcal{K}^l$, and the restricted vectors $\mathbf{c}'^l = \mathbf{c}'|_{\mathcal{S}^l}$. The realignment networks u_l then update the representations on a per-group basis:

$$u_l(\mathbf{c}'^l)_i = \begin{cases} c_i'^l, & \text{if } i \in \mathcal{S}^l \\ u_l(\hat{c}^l)_i, & \text{if } i \notin \mathcal{S}^l \end{cases}$$

The group-specific updates are combined to produce the fully updated concept vector. By applying realignment on a per-group basis, the downstream effects of interventions are restricted to semantically linked concepts. This approach introduces a semantically informed realignment module that captures correlations while preserving interpretability.

Building on Singhi et al. (2024), we implement autoregressive neural network architectures (RNN, LSTM, or GRU) across all clusters, with one instance of the network assigned per cluster to model the mapping $u_l(\mathbf{c}'^l)$. This mapping updates the initially intervened vector $\mathbf{c}'^l$ to its refined representation. Using separate networks for each cluster

prevents cross-cluster interference. A single global mapping could allow information flow between semantically distinct clusters, leading to confounding and reduced interpretability. Furthermore, the use of a global mapping may lead to shortcut learning (Geirhos et al., 2020), where spurious correlations are exploited, causing leakage between clusters and compromising their semantic independence.

In the proposed architecture, a single global loss is computed after all intervention rounds for a batch by comparing the final corrected concept vectors to the ground truth using binary cross-entropy loss. Backpropagation is performed once per batch, ensuring synchronized and efficient parameter updates across all cluster-specific models. We aggregate errors across clusters into a single loss, simplifying the training process. For selecting concepts to intervene on, we adopt a UCP policy, which has demonstrated strong performance in prior studies (Shin et al., 2023; Singhi et al., 2024). UCP identifies the most uncertain concepts for intervention based on their prediction confidence. These uncertain concepts are replaced with ground truth values during the intervention process.

As reference, we implement a realignment module from Singhi et al. (2024), UCP without realignment and a baseline vanilla CBM prediction. We expect the baseline and unconstrained realignment to represent lower- and upper-bounds to the accuracy of our model, respectively. The degree to which accuracy is lost with respect to Singhi et al. (2024) indicates the trade-off between interpretability and accuracy in intervention realignment. However, we argue that realignment within semantically informed clusters retains higher interpretability. Figure 1 shows the final flow diagram describing our approach. The policy choice corresponds to a cluster assignment within which realignment is subsequently applied.

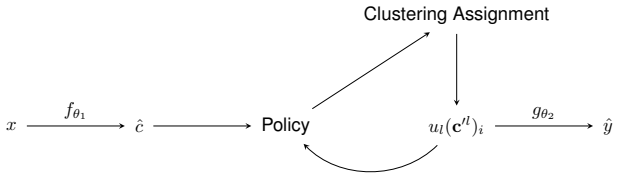

Figure 1. Flow diagram illustrating the semantics-informed group interventions realignment model. Policy interacts with Clustering Assignment, which guides the group-specific updates applied by $u_l(\mathbf{c}'^l)_i$.

3. Results

3.1. Clustering

As an illustrative example, consider the CUB dataset (Wah et al., 2011). We replicate previous work's findings that particularly hierarchical methods, such as DIANA, yield semantically well-informed groupings. On the CUB dataset, typically groupings are along color, i.e. red wing, red head, etc., for most non-degenerate hierarchically computed clusters. The algorithm is a good choice in this setting as even this type of color-based grouping remains consistent regardless of the number of clusters chosen (Figure 2). Even with 5 clusters(and thus also the less granular groupings), the clustering occurs along color, which we argue is semantically meaningful. We believe this to be the case as if one body part of a bird is of the same color, it is more likely that another body part spatially close to it is also of that color. As DIANA is a divisive algorithm, specifying a low number of splits does not yield an overly granular partition of concept space. The approach is computationally very inexpensive and scalable, running in several seconds on an Intel Celeron N4500 CPU.
We argue that the obtained color-based groupings provide a reasonable semantically-informed algorithm which allows us to subsequently bound the realignment module.
Similarly, groupings on both CelebA (Liu et al., 2018) (Table A2), and AwA2 (Xian et al., 2018) (Table A3) manage to extract reasonable distinctions. We argue that a heuristic analysis of the clustering results reveals semantic coherence across datasets. For CelebA, clusters differentiate facial aesthetics and features from concepts related to age (e.g., 'Gray Hair', 'Receding Hairline'). For AwA2, clusters group physical traits (e.g., 'Quadrupedal', 'Longneck'), habitat features (e.g., 'Cave', 'Jungle') and so on, showcasing the method's ability to identify meaningful patterns. In all cases the number of clusters we chose is the largest for which no grouping of less than 10 concepts exists.

3.2. Realignment module

Table 1 displays the results across the realignment models, the UCP policy and the baseline under a maximum of 10 interventions. Overall, consistently across model architectures, the clustered-models perform better than the baseline with respect to top1 label accuracy. Moreover, the bounded-realignment architectures (with the prefix "Multi") perform similarly to UCP. The unconstrained implementations from Singhi et al. (2024) over-perform our accuracies, as expected. These results indicate the scale of the interpretability-accuracy trade-off present. We find that without a major loss in accuracy, the clustered-models allow concept realignment with additional interpretability guarantees.

3.2.1. Ablations and Placebo Test

We run ablations on the maximum number of interventions possible on the concept vector. In Figure 3 we find that the performance of the Multi-RNN stagnates with a higher number of interventions, showing the need for the devel-

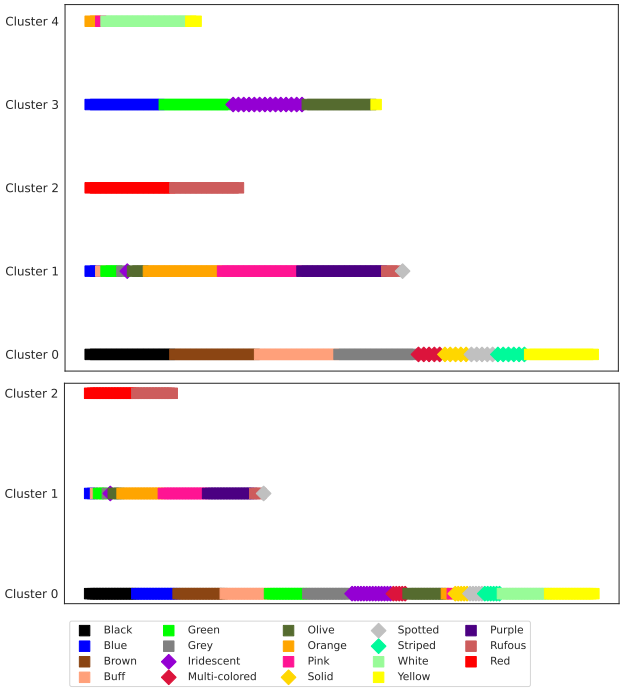

Figure 2. Exemplary clustering results on CUB data using DIANA. Body parts of the birds in question are differentiated by color.

Model Type	Test Accuracy (%)
RNN	35.93
LSTM	36.69
GRU	33.41
MultiRNN	29.79
MultiLSTM	27.05
MultiGRU	26.67
UCP policy	31.74
Baseline	26.37

Table 1. Top1 Accuracy with 200 classes across model and realignment types.

opment of additionally expressive and computationally efficient Multi-architectures. Additionally, we recreate the finding that a single intervention is insufficient in realignment frameworks (Singhi et al., 2024).

To gauge robustness to noise, we added noise equivalent to 10% of the size of the original concept vector, generated as Bernoulli noise ($p = 0.5$). We ran the clustering algorithm with the same heuristics discussed in Section 2.3. We study the model's ability to properly isolate noise within clusters and subsequently bound the effect of interventions to only a small subset of concepts. Figures A1 to A3 show that in all but one dataset, noise is isolated perfectly, while in Awa2, we cannot isolate one noise concept. We argue this showcases our method's novel ability to bound the effect of noisy concepts in intervention realignment.

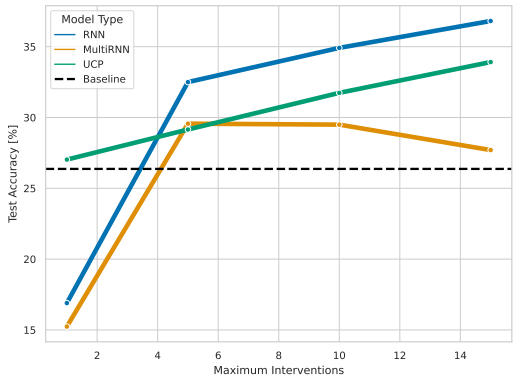

Figure 3. In an ablation on the maximum number of interventions, test accuracy is evaluated across different model types to assess performance with increasing interventions.

4. Discussion

The approach presented introduces interpretability guarantees and a novel ability to isolate and subsequently bound the effect of noise in concept realignment. We argue that the performance is similar to a UCP policy and retains concept specificity possibly lost in the higher accuracy models from (Singhi et al., 2024).

Admittedly, our method does not perform better than UCP in this example, while still improving accuracy over the vanilla CBM. Moreover, training our models is computationally more expensive than UCP and standard realignment methods. However, we argue that access to better computational resources could improve the performance of the clustered models further. We find complex model architectures were necessary to obtain meaningful results for all realignment architectures. Therefore, we claim that our approach provides a semantically-informed realignment architecture which does not sacrifice interpretability. Additional effort is required to identify architectural choices that can outperform UCP, thereby justifying the higher computational expense.

5. Conclusion

This work presents a semantics-informed clustered realignment framework for Concept Bottleneck Models. The approach emphasizes interpretability and robustness against noise while maintaining comparable performance to existing configurations across architectures. Hierarchical clustering creates semantically informed groupings. Within these groupings, the effect of interventions is bounded, which ensures interpretability guarantees.

References

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

Kaufman, L. and Rousseeuw, P. J. *Divisive Analysis (Program DIANA)*, chapter 6, pp. 253–279. John Wiley Sons, Ltd, 1990. ISBN 9780470316801. doi: https://doi.org/10.1002/9780470316801.ch6. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470316801.ch6.

Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., and Liang, P. Concept bottleneck models. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5338–5348, Virtual, 2020. PMLR. URL https://proceedings.mlr.press/v119/koh20a.html.

Lipton, Z. C. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.

Liu, Z., Luo, P., Wang, X., and Tang, X. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15 (2018):11, 2018.

Marcinkevičs, R., Reis Wolfertstetter, P., Klimiene, U., Chin-Cheong, K., Paschke, A., Zerres, J., Denzinger, M., Niederberger, D., Wellmann, S., Ozkan, E., Knorr, C., and Vogt, J. E. Interpretable and intervenable ultrasonography-based machine learning models for pediatric appendicitis. *Medical Image Analysis*, 91:103042, January 2024. ISSN 1361-8415. doi: 10.1016/j.media.2023.103042. URL http://dx.doi.org/10.1016/j.media.2023.103042.

Palumbo, E., Vandenhirtz, M., Ryser, A., Daunhawer, I., and Vogt, J. E. From logits to hierarchies: Hierarchical clustering made simple. *arXiv preprint arXiv:2410.07858*, 2024.

Shin, S., Jo, Y., Ahn, S., and Lee, N. A closer look at the intervention procedure of concept bottleneck models. In *International Conference on Machine Learning*, pp. 31504–31520. PMLR, 2023.

Singhi, N., Roth, K., Kim, J. M., and Akata, Z. Improving intervention efficacy via concept realignment in concept bottleneck models. In *ICLR 2024 Workshop on Representational Alignment*, 2024. URL https://openreview.net/forum?id=7bQmU2rukF.

Vandenhirtz, M., Laguna, S., Marcinkevičs, R., and Vogt, J. E. Stochastic concept bottleneck models. *arXiv preprint arXiv:2406.19272*, 2024.

Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. The caltech-ucsd birds-200-2011 dataset. 2011.

Xian, Y., Lampert, C. H., Schiele, B., and Akata, Z. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2251–2265, 2018.

A. Appendix: Additional Figures and Tables

Dataset	Training Time	Task Accuracy	Concept Accuracy
CUB	21min 19s	52.71	90.66
CelebA	5h 22min 50s	0.0	81.49
AwA2	9h 29min 10s	99.93	99.99

Table A1. Performance Metrics Across Datasets

Cluster 0	Cluster 1
5 o Clock Shadow	Bald
Arched Eyebrows	Blurry
Attractive	Chubby
Bags Under Eyes	Double Chin
Bangs	Eyeglasses
Big Lips	Goatee
Big Nose	Gray Hair
Black Hair	Mustache
Blond Hair	Receding Hairline
Brown Hair	Sideburns
Bushy Eyebrows	Wearing Hat
Heavy Makeup	Wearing Necktie
High Cheekbones	
Male	
Mouth Slightly Open	
Narrow Eyes	
No Beard	
Oval Face	
Pale Skin	
Pointy Nose	
Rosy Cheeks	
Smiling	
Straight Hair	
Wavy Hair	
Wearing Earrings	
Wearing Lipstick	
Wearing Necklace	
Young	

Table A2. Clustering results on CelebA dataset using DIANA. Note how Cluster 1 is associated with maturity, in contrast to Cluster 0,which consists of comparatively youthful features.

Cluster 0	Cluster 1	Cluster 2	Cluster 3
Black	Orange	Blue	Small
White	Red	Spots	Paws
Brown	Yellow	Hairless	Buckteeth
Gray	Stripes	Flippers	Claws
Patches	Hands	Strainteeth	Hops
Furry	Pads	Tusks	Tunnels
Toughskin	Flys	Swims	Weak
Big	Bipedal	Fish	Nocturnal
Bulbous	Insects	Plankton	Hibernate
Lean	Scavenger	Skimmer	Forest
Hooves	Stalker	Arctic	Nestspot
Longleg	Desert	Coastal	
Longneck	Bush	Ocean	
Tail	Jungle	Water	
Chewteeth	Tree		
Meatteeth	Cave		
Horns			
Smelly			
Walks			
Fast			
Slow			
Strong			
Muscle			
Quadrapedal			
Active			
Inactive			
Agility			
Meat			
Vegetation			
Forager			
Grazer			
Hunter			
Newworld			
Oldworld			
Plains			
Fields			
Mountains			
Ground			
Fierce			
Timid			
Smart			
Group			
Solitary			
Domestic			

Table A3. Clustering results on AwA2 dataset using DIANA. Note how Cluster 2 groups together aquatic features.

Figure A1. Clustering results on birds dataset *with added* 10% *Bernoulli noise* ($p = 0.5$), using DIANA.

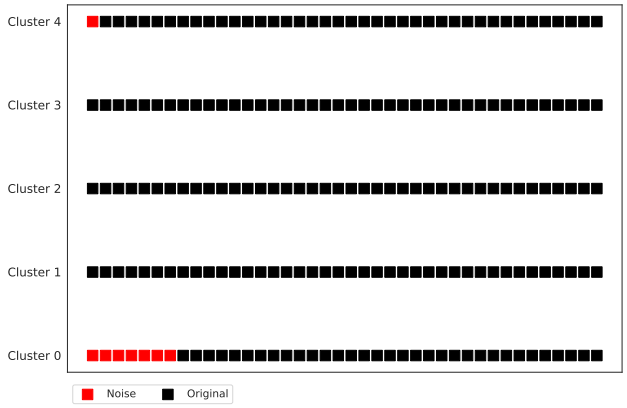

Figure A2. Clustering results on AwA2 dataset *with added* 10% *Bernoulli noise* ($p = 0.5$), using DIANA.

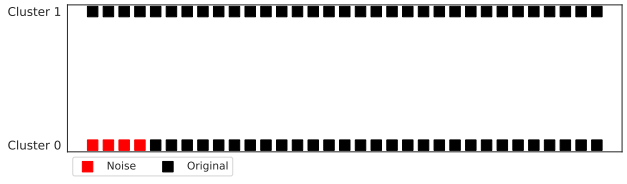

Figure A3. Clustering results on CelebA dataset *with added* 10% *Bernoulli noise* ($p = 0.5$), using DIANA.