

# Object-Centric Case-Based Reasoning via Argumentation

Gabriel de Olim Gaul<sup>1</sup>, Adam Gould<sup>1,\*</sup>, Avinash Kori<sup>1</sup> and Francesca Toni<sup>1</sup>

<sup>1</sup>*Department of Computing, Imperial College London, United Kingdom*

## Abstract

We introduce Slot Attention Argumentation for Case-Based Reasoning (SAA-CBR), a novel neuro-symbolic pipeline for image classification that integrates object-centric learning via a neural Slot Attention (SA) component with symbolic reasoning conducted by Abstract Argumentation for Case-Based Reasoning (AA-CBR). We explore novel integrations of AA-CBR with the neural component, including feature combination strategies, casebase reduction via representative samples, novel count-based partial orders, a One-Vs-Rest strategy for extending AA-CBR to multi-class classification, and an application of Supported AA-CBR, a bipolar variant of AA-CBR. We demonstrate that SAA-CBR is an effective classifier on the CLEVR-Hans datasets, showing competitive performance against baseline models.

## Keywords

Computational Argumentation, Slot Attention, Case-Based Reasoning, Neuro-Symbolic AI

## 1. Introduction

Over the past decade, deep learning models have become powerful tools across a wide range of domains. Despite their impressive performance, these models are often regarded as black boxes due to their lack of interpretability. Meanwhile, there is growing consensus among researchers, ethicists, policymakers, and the broader public about the critical importance of explainability — especially in high-stakes applications such as medicine, healthcare, and autonomous driving [1, 2]. Interpreting the decisions made by deep learning classifiers not only highlights their internal mechanisms but also helps identify potential biases [3], thereby deepening our understanding of the underlying data-generating process [4]. Many techniques have been developed to enhance interpretability, including feature attribution methods [5], network dissection approaches [6], mechanistic analyses of neural networks [7, 8], and causal or counterfactual explanations [9]. Explainability methods are generally categorised as either ante-hoc or post-hoc explanations [10], with the majority falling into the post-hoc category.

However, most of these approaches fail to align with the concept-based reasoning that humans naturally employ [11]. Some recent methods aim to bridge this gap—[12] highlights the presence of conceptual representations, while [13] explores their interactions to produce explanations. To further this direction, debate-based explanation generation was introduced in [14, 15], where the rationale for a prediction emerges from simulated agent dialogues. Nonetheless, these methods often fall short in producing clear human-interpretable concepts.

Neuro-symbolic methods have increasingly focused on addressing this particular challenge by integrating neural networks with symbolic, logic-based reasoning frameworks [16, 17]. These hybrid models combine the representational power of neural networks, which excel at learning complex data patterns, with the structured reasoning capabilities of symbolic systems, leading to more interpretable outputs [18, 19]. Building on this foundation, we propose a novel and scalable neuro-symbolic framework that leverages object-centric learning—specifically Slot Attention (SA) [20] as the neural backbone for unsupervised, human-understandable concept discovery. Here, the neural component extracts a set of object representations, which are then classified to derive object-level attributes. These attributes generate arguments within a symbolic argumentation model to predict outcomes for input images. Unlike prior work that combines SA with argumentation [19], our approach is the first to incorporate

*ArgXAI-25: 3rd International Workshop on Argumentation for eXplainable AI*

✉ gabriel.de-olim-gaul21@imperial.ac.uk (G. d. O. Gaul); adam.gould19@imperial.ac.uk (A. Gould); a.kori21@imperial.ac.uk (A. Kori); f.toni@imperial.ac.uk (F. Toni)

✉ 0009-0008-0972-7501 (A. Gould); 0000-0002-5878-3584 (A. Kori); 0000-0001-8194-1459 (F. Toni)



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

the Abstract Argumentation for Case-Based Reasoning (AA-CBR) framework [21], which leverages prior argumentation cases to resolve new instances—offering enhanced scalability.

**Contributions:** Our key contributions are as follows. (1) We propose Slot Attention Argumentation for Case-Based Reasoning (SAA-CBR), a novel neuro-argumentative framework for image classification, and evaluate its performance on the CLEVR-Hans datasets [18], demonstrating competitive results with state-of-the-art models. (2) We explore novel integrations of AA-CBR into the SAA-CBR pipeline. These include using a One-Vs-Rest approach to apply the model in a multi-class classification setting; experimenting with Supported AA-CBR [22], a novel variant that introduces support relations in the casebase; using case clustering and uncertainty filtering to reduce the size of the casebase; exploring novel methods for feature combination to overcome AA-CBR’s inherent lack of feature weighting. (3) Additionally, we propose novel counting-based partial orders and characterisations for AA-CBR, enabling object-type counting within scenes as part of the proposed pipeline.

## 2. Related Work

**Object-centric representation learning.** The primary goal of Object-Centric Learning (OCL) is to achieve spatial disentanglement—that is, to learn independent representations for each object present in an image. Early approaches to OCL [23, 24, 25] predominantly leveraged (iterative) variational inference procedures [26, 27, 28], typically built upon the Variational Autoencoder (VAE) framework [29]. More recent advances in the field have shifted toward using an iterative attention mechanism known as Slot Attention (SA) [20] to fulfil the OCL objective. A growing body of follow-up work [30, 31, 32, 33, 34, 35, 36, 37] has since extended the Slot Attention framework, making it applicable to a range of real-world settings. In this work, we adopt the iterative Slot Attention approach to spatially disentangle image content and leverage the resulting object-centric representations for downstream reasoning tasks.

**Neuro-symbolic learning.** [17] provides an overview of several approaches that integrate logic-based learning with neural models, including for image classification. One such method, NeSyFold [38], employs the rule-based learning algorithm FOLD-SE-M to transform binarized kernels from a trained convolutional neural network into a set of Answer Set Programming (ASP) rules [39] with abstract predicates. It then applies semantic labeling to associate these predicates with human-interpretable concepts, thereby producing global explanations for image classifications. Another method, Embed2Sym [40], combines clustered neural embeddings with a symbolic reasoner, encoded with predefined rules, to generate explainable predictions. Additional approaches incorporate argumentation-based reasoning for image classification. For example, [41] classifies images through argumentative debates derived from neural encoders, while [14, 15] generate argumentative explanations based on quantized image features. However, most of these approaches either depend on densely labeled data or lack the ability to produce truly human-understandable concepts and explicit symbolic reasoning. This limitation motivated the development of the Neuro-Symbolic Concept Learner (NS-CL) [18], which leverages object-centric learning via Slot Attention to identify distinct objects in images, while using set transformers in its reasoning module to support explanation generation. This line of work was further extended to incorporate fully symbolic reasoning layer in Object-Centric Neuro-Argumentative Learning (OC-NAL) [19], which enables more complex argumentative reasoning over the slot-based object representations extracted from visual input. In this work, inspired by OC-NAL, we enable case-based reasoning on the extracted object representations, providing access to the model’s reasoning for making a certain decision.

**Argumentative Case-Based Reasoning** Computational argumentation has been used to support case-based reasoning approaches, in which previously observed cases are cited as evidence for/against a classification. *Abstract Argumentation for Case-Based Reasoning* (AA-CBR) [21] is one such approach wherein data points are characterised as *cases* that argue about the outcome of an unlabelled case. However, as a purely symbolic method, this method cannot easily be deployed with complex data types,

such as images, and cannot automatically select features in the data. The introduction of user-specified feature preferences [42] to AA-CBR was proposed to aid feature selection, but this does not solve the issue of more complex data types. The *Data-Empowered Argumentation (DEAr)* [43] was proposed, in which a characterisation extractor converts complex data into symbols that can be reasoned over with AA-CBR. However, DEAr has not been tested with a neural component extracting features from images. Our proposed approach uses this pipeline methodology with an SA module as the feature extractor. This is unlike Gradual AA-CBR [44], which attempts to learn the argumentation debate end-to-end with the feature extractor but has yet to be applied to more complex data types like images.

### 3. Background

#### 3.1. Slot-Attention

The Slot Attention (SA) module [20] is an architectural mechanism that interfaces with perceptual spatial representations, as estimated by a convolutional neural network (CNN) encoder—to produce a set of task-specific abstract representations referred to as slots. These slots dynamically bind to arbitrary objects in the input through a competitive, attention-based refinement process that unfolds over multiple iterations. Slot attention takes a set of  $N$  feature embeddings  $\mathbf{z} = \phi(\mathbf{x}) \in \mathbb{R}^{N \times d_z}$ , as input, where  $d_z$  is the dimensionality of  $\mathbf{z}$ ,  $\mathbf{x}$  corresponds to an observed image, with CNN encoder  $\phi$ , and applies an iterative attention to produce  $K$  object representations called slots  $\mathbf{s} \in \mathbb{R}^{K \times d_s}$  ( $d_s \leq d_z$ ), where  $d_s$  is the dimensionality of each slot. For this the input features  $\mathbf{z}$  are projected with a linear layer resulting in a *query*, *key*, and *value* representations, denoted by  $\mathbf{q}$ ,  $\mathbf{k}$  and  $\mathbf{v}$ , respectively. To simplify let  $f$  and  $g$  be shorthand notation for the *slot update* and *attention* functions respectively:

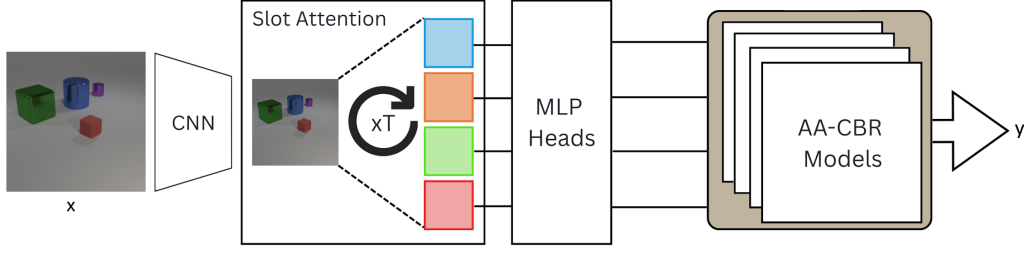
$$f(\mathbf{A}, \mathbf{v}) = \mathbf{A}^T \mathbf{v}, \quad \mathbf{A}_{ij} = \frac{g(\mathbf{q}, \mathbf{k})_{ij}}{\sum_{l=1}^K g(\mathbf{q}, \mathbf{k})_{lj}} \quad \text{and} \quad g(\mathbf{q}, \mathbf{k}) = \frac{e^{\mathbf{M}_{ij}}}{\sum_{l=1}^N e^{\mathbf{M}_{il}}}, \quad \mathbf{M} = \frac{\mathbf{k}\mathbf{q}^T}{\sqrt{d_s}}, \quad (1)$$

where  $\mathbf{A} \in \mathbb{R}^{N \times K}$  reflects the cross-attention matrix capturing the competition across  $K$  slots to attend to a  $N$  image tokens. Unlike self-attention [45], the queries in slot attention are a function of the slots, initialised by sampling from a normal distribution  $\mathcal{N}(\mathbf{s}; \mu, \sigma)$ , i.e.,  $\mathbf{s} \sim \mathcal{N}(\mathbf{s}; \mu, \sigma) \in \mathbb{R}^{K \times d_s}$ , and are iteratively refined over  $T$  attention iterations [20]. Methodologically, the slots are randomly initialized at  $t = 0$ . The queries at iteration  $t$  are given by  $\hat{\mathbf{q}}^t$  and are generated as a linear projection of current state slots  $\mathbf{s}^t$ , and the slot update process can be summarized as:  $\mathbf{s}^{t+1} = f(g(\hat{\mathbf{q}}^t, \mathbf{k}), \mathbf{v})$ . Lastly, a Gated Recurrent Unit (GRU) is applied to the slot representations  $\mathbf{s}^{t+1}$  at the end of each SA iteration, followed by a generic MLP skip connection.

#### 3.2. Abstract Argumentation for Case-Based Reasoning

An abstract *argumentation framework (AF)* [46] is a pair  $\langle \mathcal{A}, \rightsquigarrow \rangle$ , where  $\mathcal{A}$  is a set of arguments and  $\rightsquigarrow \subseteq \mathcal{A} \times \mathcal{A}$  is a binary relation defining *attacks* between arguments. We can visualise an AF as a directed graph with arguments as nodes and attacks as edges. A set of arguments  $E \subseteq \mathcal{A}$  *defends* an argument  $b \in \mathcal{A}$  if for all  $a \rightsquigarrow b$  there exists  $c \in E$  such that  $c \rightsquigarrow a$ . To determine the set of *acceptable* arguments, we apply the *grounded* semantics. The *grounded extension*  $\mathbb{G}$  is computed as  $\mathbb{G} = \bigcup_{i \geq 0} G_i$ , where  $G_0$  is the set of unattacked arguments and  $\forall i \geq 0, G_{i+1}$  is the set of all arguments that  $G_i$  defends.

*Abstract Argumentation for Case-Based Reasoning (AA-CBR)* [21] is a machine learning approach for binary classification in which an argumentation framework is constructed from the training data. Each argument in the resulting AF is built from a single data point and is referred to as a *case*. A partial order defining a notion of *exceptionality* between data points allows the AF to be constructed utilising defeasible reasoning, wherein each data point is an exception to those they attack. An unlabelled case,  $N$ , is added to the framework by attacking any cases considered *irrelevant* to it. We then apply the grounded semantics to compute which arguments are accepted. If a *default argument* - representing our



**Figure 1:** An overview of the Slot Attention Argumentation for Case-Based Reasoning (SAA-CBR) architecture.

expected outcome given no other information - is accepted, then its corresponding outcome is predicted for  $N$ ; otherwise, the opposing class is predicted. Formally:

**Definition 3.1.** Let  $D_{cb} \subseteq X \times Y$  be a finite casebase of labelled examples where  $X$  is a set of characterisations and  $Y = \{\delta, \bar{\delta}\}$  is the set of possible outcomes. Each example is of the form  $(x, y)$ . Let  $(x_\delta, \delta)$  be the default argument with  $\delta$  the default outcome. Let  $N$  be an unlabelled example of the form  $(x_N, y_?)$  with  $y_?$  an unknown outcome. Let  $\succsim$  and  $\approx$  be a partial order and binary relation defined over  $X$  representing exceptionality and irrelevance respectively. The argumentation framework  $AF(D_{cb}, x_N)$  mined from  $D_{cb}$  and  $x_N$  is  $\langle \mathcal{A}, \rightsquigarrow \rangle$  in which:

- $\mathcal{A} = D_{cb} \cup \{(x_\delta, \delta)\} \cup \{N\}$
- for  $(x_a, y_a), (x_b, y_b) \in D_{cb} \cup \{(x_\delta, \delta)\}$ , it holds that  $(x_a, y_a) \rightsquigarrow (x_b, y_b)$  iff
  1.  $y_a \neq y_b$ , and
  2.  $x_a$  is more exceptional than  $x_b$  and there is minimal difference between them, i.e.:
    - a)  $x_a \succ x_b$  and
    - b)  $\nexists (x_c, y_c) \in D_{cb} \cup \{(x_\delta, \delta)\}$  with  $y_c = y_a$  and  $x_a \succ x_c \succ x_b$ ;
- for  $(x_a, y_a) \in D_{cb} \cup \{(x_\delta, \delta)\}$ , it holds that  $N \rightsquigarrow (x_a, y_a)$  iff  $x_N \approx x_a$ .

Finally, we have  $AA\text{-}CBR(D_{cb}, x_N) = \delta$  if  $(x_\delta, \delta) \in \mathbb{G}$  and  $\bar{\delta}$  otherwise, where  $\mathbb{G}$  is the grounded extension of  $AF(D_{cb}, N)$ .

In this work, we make use of *regular* AA-CBR [47] in which the default characterisation,  $x_\delta$ , is the least element of the partial order  $\succsim$  and irrelevance is the negation of exceptionality, that is  $\approx \neq$ .

## 4. Slot-Attention for Case-Based Reasoning

Figure 1 showcases our proposed architecture, comprised of a neural-component governed by a slot-attention mechanism and a symbolic-component that reasons over the learned representations using AA-CBR. Effectively, the slot-attention mechanism acts as a characterisation extractor for AA-CBR.

**Notations.** Similar to OC-NAL [19], our architecture operates on an observed dataset  $D \subseteq \mathcal{X} \times \mathcal{Y} \times \mathcal{L}$  consisting of labelled images, where  $\mathcal{X}$  denotes the set of input images,  $\mathcal{L} = \{c_1, \dots, c_L\}$  represents a set of  $L$  class labels, and  $\mathcal{Y} = \{0, 1\}^{K \times (P+1)}$  is metadata encoded as one-hot vectors. Here,  $K$  is the maximum number of objects (or slots) that can appear in an image, and  $P$  is the number of distinct properties each object may possess. An additional dimension is included to account for the absence of an object. Importantly, the neural component of the architecture does not utilize the class labels in  $L$ ; instead, it processes the image data in  $x \sim \mathcal{X}$  in a *weakly supervised* fashion. This is weakly supervised because we are using object-level labels instead of pixel-wise dense labels. Additionally, semantics object segmentation is obtained in an unsupervised manner, while supervision is only used for classification. Conversely, the symbolic reasoning component does not operate on the raw images but instead consumes an abstract representation generated by the neural component.

## 4.1. Neural Component

The neural component is responsible for converting an image input  $\mathbf{x} \in \mathcal{X}$  into a set of slot representations  $\mathbf{s}$  that describe individual objects present in an image. These representations form the basis of the cases used by the argumentative reasoning component. To achieve this, we utilize the Slot Attention (SA) module, as described in the background section. This component matches with the neural components described in [19].

The image  $\mathbf{x}$  is first passed through a feature extractor  $\phi$ , yielding a feature map  $\mathbf{z}$ . The SA module then processes  $\mathbf{z}$  in an iterative manner, producing a set of slot representations  $\mathbf{s} = \{\mathbf{s}_1, \dots, \mathbf{s}_K\}$ . Due to the permutation equivariant nature of SA, these slots are not inherently grounded with respect to object identity or properties. Therefore, we employ multi-layer perceptron (MLP) heads to classify the slot representations and derive object-level attributes (e.g., color, shape, size), using weak supervision from the available metadata  $\mathcal{Y}$ .

Our framework adopts a two-stage training approach, where the neural component is trained independently from the argumentative component. The object discovery mechanism within the neural component is optimized via self-supervision using a reconstruction objective. Specifically, we minimize the mean squared error (MSE) loss between the original image  $\mathbf{x}$  and its reconstruction  $\hat{\mathbf{x}}$ . In parallel, the MLP heads are trained using a binary cross-entropy (BCE) loss between the predicted attributes  $\hat{\mathbf{y}}$  and the ground-truth labels  $\mathbf{y} \in \mathcal{Y}$ , accounting for the multi-label nature of the classification problem.

As noted, the permutation equivariance of SA leads to an unknown slot order during inference. To address this, we apply the Hungarian Matching (HM) algorithm during training to find the optimal alignment  $\tau$  between the predicted slots and ground-truth annotations before computing the BCE loss:

$$\mathcal{L} = \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 + \alpha \cdot \min_{\tau} \sum_{j=1}^P \text{BCE}(\mathbf{y}_j, \hat{\mathbf{y}}_{\tau(j)}) \quad (2)$$

where  $\tau$  denotes the permutation provided by the HM algorithm, and  $\alpha$  is a weighting hyperparameter. It is important to note that the HM step is used only during training for alignment with ground-truth annotations. During inference, this step is omitted. The final output of the neural component is a tensor containing the predicted attributes for each object in the input image.

## 4.2. Argumentative Component

The attributes for each object in the input image can be used to construct characterisations that the AA-CBR can reason with. Attributes may include the shape, colour, size, material, and position of each object in the image. Labelled data points can be passed through the SA component, characterised and used to construct a casebase that can be used to construct an AF that makes predictions for unlabelled data points. We now consider how to integrate AA-CBR with a neural component for image classification.

**Feature Combination.** AA-CBR has no inherent ability to conduct feature weighting. Whilst the slot-attention module is useful for extracting sets of features that can be reasoned with using AA-CBR, as these two components are trained separately, we have no way of knowing whether these features are useful for the downstream classification task. To address this issue, the DEAr pipeline [43] employs feature selection, discarding features deemed unimportant, e.g as determined by the weights of an autoencoder. However, discarding features may result in the removal of those necessary for classifying a specific class. This is a problem if, for example, a class depends on the existence of both a small metal cube and a small sphere whose material is irrelevant to the classification. The material of an object is important in one instance but unimportant in the other, so we cannot simply discard the attribute, but we cannot keep it for all instances either. Instead, we propose combining the individual features into super-features and then using feature selection to determine which super-features to use. For example, for an image that contains a small (*sm*) metal (*m*) cube (*cu*) and a small (*sm*) rubber (*ru*) sphere (*sp*), we may use two super-features:  $\{sm\_m\_cu, sm\_sp\}$ . By applying feature selection over



super-features, we can extract only the properties necessary for a classification<sup>1</sup>. So if the small sphere super-feature is selected, any time the model encounters a small rubber sphere, it will only ever be represented as simply a small sphere. The useful feature combinations can be tuned for a given AA-CBR Model.

**Casebase Reduction.** Furthermore, as traditional AA-CBR uses each data point as an argument, the number of arguments can be very large. This can be problematic as the outputs from the neural-component can introduce noise, if, for example, an object is incorrectly predicted in the image or there are confounding variables in the image. Additionally, for large datasets, constructing the AF with AA-CBR may be a slow process given the algorithm has a worst-case time complexity of  $O(n^3)$ . Moreover, AFs with a large number of arguments are cognitively intractable, that is, they can be difficult for humans to interpret. To reduce the number of arguments, we can first apply a clustering algorithm (e.g., kMeans clustering [48]) and use the cluster centroids as the data points in the casebase. We can further filter the casebase by removing centroids where the sum of the confidence scores of each predicted object does not meet a set threshold. Both these approaches filter out noisy and uncertain predictions from the neural-component whilst reducing the casebase size. The number of centroids and the value of the thresholds can be tuned as hyperparameters for an AA-CBR model.

**Multi-Class Classification.** As AA-CBR computes a grounded extension which only determines if an argument is accepted or not, it is only suitable for binary classification. To apply it to a multi-class classification task, we can use a tournament-style classification strategy in which we chain various AA-CBR models together. We decided to use a One-Vs-Rest (OvR) approach in which the AA-CBR model selects one class as the *focus class* and aims to distinguish a data point between this focus class and all other classes. For those data points classified as ‘other’, we can again select a new focus class and deploy the same strategy with the focus class excluded. We can continuously apply this until the final two classes remain, in which case the final model simply distinguishes between these two classes. We can tune the order of application of the focus classes as well as whether each model uses the focus class or ‘other’ as the default outcome to optimise model performance.

**Supported AA-CBR.** We also experiment with Supported AA-CBR [22], in which we modify the AA-CBR model to additionally include supporting relationships between arguments with the same label. Supports allow for more arguments to be included in the classification process, as the AA-CBR minimality condition can otherwise result in excluded cases [47]. The support relation is interpreted as an indirect attack: if argument  $a$  supports argument  $b$ , and  $b$  attacks argument  $c$ , then  $a$  is also considered to indirectly attack  $c$ . We can then apply the grounded semantics with the set of attacks and indirect attacks. By including supports, we ensure the cases in the casebase are used more effectively. Each model in the tournament can treat whether to use supports or not as an additional hyperparameter.

**Instantiating AA-CBR.** AA-CBR relies on the use of a partial order over the case characterisations to be defined. We experiment with two possible characterisation methods and corresponding partial orders. For the first characterisation approach, we let  $X = \mathbb{F}$ , where  $\mathbb{F}$  is the set of (super-)features, and define the exceptionality partial order with the subset relation  $\subseteq$ . For example, consider that the slot attention module determines that an image contains a cube ( $cu$ ) and a large ( $l$ ) cylinder ( $cy$ ). If we know such an image belongs to class 0, it can be represented as  $a = (\{cu, l\_cy\}, 0)$  and would be considered an exception to a image with just a cube, belonging to class 1:  $b = (\{cu\}, 1)$ . This approach is as defined in the original proposal of AA-CBR [21].

However, this characterisation approach may prove problematic, given that an image may contain multiple instances of the same object with the same properties. Instead, we propose a novel count-based characterisation. Each case is therefore characterised as  $(f_i : n_{a,i} | f_i \in \mathbb{F})$ , where  $n_{a,i}$  is the number

<sup>1</sup>This approach comes at the expense of a combinatorial number of super-features to select over. Future work should look at feature combination methods that can be learned with the model.

of times the feature appears in case  $a$ . For example, if an image in class 0 contains two cubes and a single large cylinder, it would be represented as  $((cu : 2, l\_cy : 1), 0)$ <sup>2</sup>. We define exceptionalism as  $(f_i : n_{a,i} | f_i \in \mathbb{F}) \prec (f_i : n_{b,i} | f_i \in \mathbb{F})$  iff  $\forall i \in [0, m], n_{a,i} \leq n_{b,i}$  and  $\exists i \in [0, m], n_{a,i} < n_{b,i}$  where  $m$  is the total number of features. For example, a class 0 image with two cubes and a single large cylinder would be considered an exception to an image in class 1 with just two cubes represented as  $((cu : 2), 1)$  because the image in class 0 has at least as many cubes as the image in class 1 but has more large cylinders. We can treat the choice of partial order as a hyperparameter for each AA-CBR in the tournament, swapping between different characterisations and partial orders as necessary for the classification task.

Figure 2 shows a small example AA-CBR model with just four cases in the casebase using the first characterisation approach<sup>3</sup>. The new case,  $C_N$  contains a large cylinder ( $l\_cy$ ), a large cube ( $l\_cu$ ) and another cube ( $cu$ ) but no small sphere ( $s\_sp$ ). Cases in the casebase attack all those with a different outcome and a subset of features, subject to the minimality condition (black arrows). Under the grounded semantics, case  $C_0$  is accepted, so  $C_N$  is predicted to be in class 0. This can be explained as follows: the debate starts with  $C_0$ , the default rule assigning class 0. Case  $C_1$  challenges this, since it has a large cylinder but belongs to another class.  $C_3$  then defeats  $C_1$ , as it includes both a large cylinder and a large cube but belongs to class 0.  $C_2$  also tries to argue for the other class, but it is dismissed since it contains a small sphere, making it irrelevant to  $C_N$ . Similarly,  $C_4$  is dismissed for the same reason.

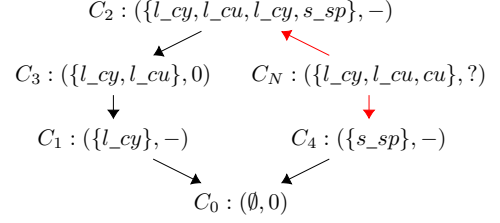


Figure 2: An example AA-CBR model.

## 5. Experiments

**Datasets:** We evaluate our approach on the CLEVR-Hans3 and (a modified variant of) the CLEVR-Hans7 datasets [18]. The CLEVR-Hans datasets contain images with objects of varying shapes, sizes, materials and colours grouped into three and seven classes respectively. These classes are constructed from rules about the objects; for example, any image that contains a large cube and a large cylinder is in the first class. However, CLEVR-Hans has the additional difficulty of including confounding variables, which are designed to assess if the system can ideally learn the most general rules about a class, or instead learns more specific rules based on its training data. For example, in the training/validation sets of CLEVR-Hans, all large cubes in the images of the first class are grey, but the rule itself is not dependent on the colour of the cube. Meanwhile, the test set contains images labelled as the first class with large cubes and large cylinders of any colour.

CLEVR-Hans7 introduces four additional classes to CLEVR-Hans3, which include more complex rules that consider the number of objects, the absolute position of objects (e.g. left/right side of the image) and the relative position of the objects (e.g. is object A in front of object B). However, with our feature extraction method, relative positioning would lead to a combinatorial number of features with respect to the number of objects in the image. For this reason, we chose to exclude the one class from the dataset that relies on relative positioning in its class rule for our testing. We refer to our modified dataset as ‘CLEVR-Hans7 (modified)’.

**Hyperparameter Tuning** The integrations of AA-CBR into our SAA-CBR pipeline result in several possible parameters to tune. This includes the order of classes in which we apply the OvR strategy. Additionally, for each AA-CBR model, we must tune which class is the default, how cases are characterised, the number of representative samples used to build the AF, the threshold value, and whether the model

<sup>2</sup>We chose to exclude features with a count of 0 from the notation. So in this case, we know implicitly that this image has, for example, 0 spheres.

<sup>3</sup>The real models contain the number of cluster centres in their casebase.

uses supports. Due to the large number of parameters, we employed a random search to identify which could lead to the best performance. Selected hyperparameters are specified in Appendix A.

**Baseline models** We compare our proposed SAA-CBR method against four baseline models. The first is a ResNet18 [49], which is a deep neural network consisting of convolutional layers and residual (skip) connections. We selected this model as it offers a simple baseline for a neural-based approach that makes no use of symbolic reasoning to resolve the classification. The model was trained for 100 epochs using the Adam Optimiser [50] with a learning rate of 0.0001. The second baseline selected consists of a Slot Attention module in a pipeline with an MLP handling the final classification. This allows us to demonstrate whether the MLP is capable of learning the reasoning mechanism that a symbolic module is capable of. We train this in a pipeline fashion, with the SA module trained separately from the MLP to enable direct comparison with our approach. The MLP consists of a three-layer network with ReLU activations for the hidden layers and a softmax activation for the final layer. The MLP was trained for 100 epochs with the Adam optimiser with a learning rate of 0.001. We used OC-NAL [19] for the third baseline model for a direct comparison of a pipeline using an SA module followed by an argumentative reasoner. This pipeline uses Assumption-Based Argumentation (ABA) [51, 52] for the symbolic reasoning. This is a form of structured argumentation, corresponding in OC-NAL to logic programs with negation as failure. The ABA framework is itself learned from positive and negative examples by iteratively applying rules from the ABALearn algorithm [53]. Finally, we compare our approach to NS-CL [18], to compare against a neurosymbolic model that does not make use of argumentation and is trained end-to-end. We used the pre-trained weights for the SA module for NS-CL and the set transformer trained with an Adam optimiser, with a learning rate of 0.0001 for 50 epochs.

## 5.1. Evaluation

Validation Set				
Model	CLEVR-Hans3		CLEVR-Hans7 (modified)	
	Accuracy	F1	Accuracy	F1
ResNet18	$59.02 \pm 0.00$	$59.04 \pm 0.00$	$30.96 \pm 0.00$	$31.09 \pm 0.00$
Slot Attention w/ MLP	$55.00 \pm 0.44$	$54.36 \pm 0.44$	<u><math>62.90 \pm 0.20</math></u>	$61.24 \pm 0.25$
OC-NAL	$66.87 \pm 0.32$	$65.47 \pm 0.29$	$15.57 \pm 0.46$	$11.62 \pm 0.43$
NS-CL	<b><math>98.86 \pm 0.07</math></b>	<b><math>98.86 \pm 0.07</math></b>	<b><math>96.00 \pm 0.18</math></b>	<b><math>96.00 \pm 0.18</math></b>
SAA-CBR (Ours)	<u><math>75.59 \pm 2.23</math></u>	<u><math>77.22 \pm 2.42</math></u>	$62.16 \pm 2.81$	<u><math>62.56 \pm 3.00</math></u>

**Table 1**

Model performances on the CLEVR-Hans3 and CLEVR-Hans7 (modified) validation sets. The best-performing model is shown in **bold**, and the second best is underlined.

Test Set				
Model	CLEVR-Hans3		CLEVR-Hans7 (modified)	
	Accuracy	F1	Accuracy	F1
ResNet18	$48.71 \pm 0.00$	$47.13 \pm 0.00$	$29.44 \pm 0.00$	$29.57 \pm 0.00$
Slot Attention w/ MLP	$52.97 \pm 0.77$	$52.22 \pm 0.78$	$58.33 \pm 0.57$	$56.37 \pm 0.56$
OC-NAL	$68.76 \pm 0.41$	$67.57 \pm 0.40$	$15.75 \pm 0.72$	$11.79 \pm 0.54$
NS-CL	<b><math>82.43 \pm 0.12</math></b>	<b><math>81.42 \pm 0.13</math></b>	<b><math>91.29 \pm 0.12</math></b>	<b><math>91.22 \pm 0.12</math></b>
SAA-CBR (Ours)	<u><math>75.13 \pm 2.16</math></u>	<u><math>74.77 \pm 2.28</math></u>	<u><math>62.87 \pm 2.81</math></u>	<u><math>63.19 \pm 3.00</math></u>

**Table 2**

Model performances on the CLEVR-Hans3 and CLEVR-Hans7 (modified) test sets. The best-performing model is shown in **bold**, and the second best is underlined.

Table 2 showcases the macro-averaged accuracy and F1-score evaluated on the test set of both datasets across five runs. We see that SAA-CBR significantly outperforms all baselines except the NS-CL. We



see that purely neural approaches cannot perform well on the CLEVR-Hans classification task, with the ResNet18 achieving an accuracies of  $48.71\% \pm 0.00$  and  $29.44\% \pm 0.00$  and the Slot Attention with MLP achieving  $52.97\% \pm 0.77$  and  $58.33\% \pm 0.57$  on CLEVR-Hans3 and CLEVR-Hans7 (modified), respectively. In particular, these models perform better on the validation sets, as shown in Table 1, than on the test sets, suggesting that these models fail to generalise beyond the confounding variables.

A form of symbolic learning reasoning appears necessary to achieve higher performance on this task without large model scaling. The introduction of argumentative reasoning with OC-NAL demonstrates how performance can be improved for CLEVR-Hans3, achieving an accuracy of  $66.87\% \pm 0.32$ . However, this approach fails to generalise to CLEVR-Hans7 (modified), with an accuracy of  $15.57\% \pm 0.46$ . It is of note that the performance of OC-NAL increased between the validation set to the test set, suggesting that the introduction of argumentative reasoning allows for this pipeline to learn rules that generalise beyond the confounders. One of the key limitations of learning ABA frameworks is ABALearn’s inability to learn from a large number of examples, wherein only 10 representative samples per class could be used to learn the framework. On the other hand, the largest of the AA-CBR models in the SAA-CBR pipeline was capable of using over 900 representative samples in its learning process. As a result, the SAA-CBR pipeline performs considerably better, achieving accuracies of  $75.13\% \pm 2.16$  and  $62.87\% \pm 2.81$  on CLEVR-Hans3 and CLEVR-Hans7 (modified), respectively. Notably, the average performance change between the SAA-CBR pipeline on the validation set compared to the test set is the smallest compared to any other model, with only a small decrease in accuracy for CLEVR-Hans3 and an increase in accuracy for CLEVR-Hans7 (modified). As with OC-NAL, this suggests that one of the key benefits of applying argumentative reasoning here is the ability to learn a more general set of rules for classifying the data. However, the variance for SAA-CBR is greater, which we attribute to the randomisation resulting from the initialisation of the k-Means centroids in the casebase clustering.

The NS-CL appears to have greater learning capacity, and the set transformers are capable of learning from the full training set compared to using a subset of representative samples with AA-CBR. However, the NS-CL model exhibits a decrease in performance on the test set compared to the validation set for both CLEVR-Hans datasets (although not as pronounced for CLEVR-Hans7 (modified)), suggesting that, despite its greater learning capacity, this model struggles to generalise beyond the confounding variables. Our modifications to AA-CBR are key for our pipeline to perform as it does, demonstrated by our ablation study in Appendix B. NS-CL is trained end-to-end, rather than in a pipeline fashion with separate training for the neural and symbolic components, which may explain its improved performance. Future work can explore end-to-end training of both AA-CBR and slot attention.

## 6. Conclusion

We introduce Slot-Attention Argumentation for Case-Based Reasoning, a novel neuro-argumentative framework integrating SA and AA-CBR for image classification. We showcase how the model outperforms neural-based baseline models and other proposed methods of combining SA and (structured) argumentation. Our novel extensions to AA-CBR, namely the introduction of a One-Vs-Rest model for multi-class classification, the application of Supported AA-CBR, the use of clustering and uncertainty filtering for casebase reduction, feature combination methods and the introduction of count-based partial orders, improve the pipeline, allowing for a successful integration of symbolic argumentative reasoning via case-based reasoning with slot attention.

Further research into the combination of SA and AA-CBR should investigate how the integration between these two components can be further enhanced. For example, instead of using a feature combination method and applying feature selection after the fact, research should explore how to directly learn which feature combinations to use. Additionally, improvements to constructing representative samples that consider the downstream argumentation process may prove more effective than using k-means clustering. Such an approach may also prove to scale better, allowing for fewer necessary representative samples that can be learned from a larger dataset. Additionally, other AA-CBR variants such as Preference-Based AA-CBR [42], Cumulative AA-CBR [47], or Gradual AA-CBR [44] could be

experimented with on this task. This latter model, which scores argument acceptability with a value between 0 and 1 rather than existence in an extension, may prove useful for improving multi-class classification beyond the One-Vs-Rest strategy and allow for end-to-end learning.

## Acknowledgments

Gould was supported by UK Research and Innovation [UKRI Centre for Doctoral Training in AI for Healthcare grant number EP/S023283/1]. Kori was supported by UKRI through the CDT in Safe and Trusted Artificial Intelligence and acknowledges support from the EPSRC Doctoral Prize. Toni was partially funded by the ERC under the EU's Horizon 2020 research and innovation programme (grant agreement No. 101020934). Toni was also partially funded by J.P. Morgan and the Royal Academy of Engineering, UK, under the Research Chairs and Senior Research Fellowships scheme.

## Declaration on Generative AI

The authors have not employed any Generative AI tools.

## References

- [1] F. Doshi-Velez, R. Budish, M. Kortz, The Role of Explanation in Algorithmic Trust, Technical Report, Technical report, Artificial Intelligence and Interpretability Working Group ..., 2017.
- [2] J. A. Kroll, Accountable algorithms, Ph.D. thesis, Princeton University, 2015.
- [3] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, et al., Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav), in: International conference on machine learning, PMLR, 2018, pp. 2668–2677.
- [4] A. Narayanaswamy, S. Venugopalan, D. R. Webster, L. Peng, G. S. Corrado, P. Ruamviboonsuk, P. Bavishi, M. Brenner, P. C. Nelson, A. V. Varadarajan, Scientific discovery by generating counterfactuals using image translation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2020, pp. 273–283.
- [5] M. T. Ribeiro, S. Singh, C. Guestrin, " why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.
- [6] D. Bau, B. Zhou, A. Khosla, A. Oliva, A. Torralba, Network dissection: Quantifying interpretability of deep visual representations, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 6541–6549.
- [7] C. Olah, A. Mordvintsev, L. Schubert, Feature visualization, Distill (2017). doi:10.23915/distill.000007, <https://distill.pub/2017/feature-visualization>.
- [8] C. Olah, N. Cammarata, L. Schubert, G. Goh, M. Petrov, S. Carter, Zoom in: An introduction to circuits, Distill (2020). doi:10.23915/distill.00024.001, <https://distill.pub/2020/circuits/zoom-in>.
- [9] A. Sauer, A. Geiger, Counterfactual generative networks, arXiv preprint arXiv:2101.06046 (2021).
- [10] Z. C. Lipton, The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery., Queue 16 (2018) 31–57.
- [11] S. L. Armstrong, L. R. Gleitman, H. Gleitman, What some concepts might not be, Cognition 13 (1983) 263–308.
- [12] A. Ghorbani, J. Wexler, J. Zou, B. Kim, Towards automatic concept-based explanations, arXiv preprint arXiv:1902.03129 (2019).
- [13] A. Kori, P. Natekar, G. Krishnamurthi, B. Srinivasan, Abstracting deep neural networks into concept graphs for concept level interpretability, arXiv preprint arXiv:2008.06457 (2020).

- [14] A. Kori, B. Glocker, F. Toni, Explaining image classifiers with visual debates, in: *Discovery Science - 27th International Conference, DS 2024, 2024*, pp. 200–214. URL: [https://doi.org/10.1007/978-3-031-78980-9\\_13](https://doi.org/10.1007/978-3-031-78980-9_13).
- [15] A. Kori, A. Rago, F. Toni, Free argumentative exchanges for explaining image classifiers, *CoRR abs/2502.12995* (2025). URL: <https://doi.org/10.48550/arXiv.2502.12995>. doi:10.48550/ARXIV.2502.12995. arXiv:2502.12995.
- [16] A. d. Garcez, L. C. Lamb, Neurosymbolic ai: the 3rd wave, *Artificial Intelligence Review* 56 (2023) 12387–12406. URL: <http://dx.doi.org/10.1007/s10462-023-10448-w>. doi:10.1007/s10462-023-10448-w.
- [17] M. Proietti, F. Toni, A roadmap for neuro-argumentative learning, in: *Proceedings of the 17th International Workshop on Neural-Symbolic Learning and Reasoning, La Certosa di Pontignano, Siena, Italy, July 3-5, 2023*, volume 3432 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 1–8. URL: <https://ceur-ws.org/Vol-3432/paper1.pdf>.
- [18] W. Stammer, P. Schramowski, K. Kersting, Right for the right concept: Revising neuro-symbolic concepts by interacting with their explanations, in: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, Computer Vision Foundation / IEEE, 2021, pp. 3619–3629. URL: <https://10.1109/CVPR46437.2021.00362>.
- [19] A. R. Jacob, A. Kori, E. D. Angelis, B. Glocker, M. Proietti, F. Toni, Object-centric neuro-argumentative learning, in: *19th International Conference on Neurosymbolic Learning and Reasoning*, 2025. URL: <https://openreview.net/forum?id=dBB58hqTXG>, to appear.
- [20] F. Locatello, D. Weissenborn, T. Unterthiner, A. Mahendran, G. Heigold, J. Uszkoreit, A. Dosovitskiy, T. Kipf, Object-centric learning with slot attention, *Advances in Neural Information Processing Systems* 33 (2020) 11525–11538.
- [21] K. Čyras, K. Satoh, F. Toni, Abstract argumentation for case-based reasoning, in: *Fifteenth International Conference on the Principles of Knowledge Representation and Reasoning*, AAAI Press, 2016.
- [22] A. Gould, G. d. O. Gaul, F. Toni, Supported abstract argumentation for case-based reasoning, 2025. URL: <https://arxiv.org/abs/2507.04994>. doi:10.48550/ARXIV.2507.04994.
- [23] C. P. Burgess, L. Matthey, N. Watters, R. Kabra, I. Higgins, M. Botvinick, A. Lerchner, Monet: Unsupervised scene decomposition and representation, *arXiv preprint arXiv:1901.11390* (2019).
- [24] M. Engelcke, A. R. Kosiorek, O. P. Jones, I. Posner, Genesis: Generative scene inference and sampling with object-centric latent representations, *arXiv preprint arXiv:1907.13052* (2019).
- [25] K. Greff, R. L. Kaufman, R. Kabra, N. Watters, C. Burgess, D. Zoran, L. Matthey, M. Botvinick, A. Lerchner, Multi-object representation learning with iterative variational inference, in: *International Conference on Machine Learning*, PMLR, 2019, pp. 2424–2433.
- [26] J. Marino, Y. Yue, S. Mandt, Iterative amortized inference, in: *International Conference on Machine Learning*, PMLR, 2018, pp. 3403–3412.
- [27] S. Van Steenkiste, K. Kurach, J. Schmidhuber, S. Gelly, Investigating object compositionality in generative adversarial networks, *Neural Networks* 130 (2020) 309–325.
- [28] Z. Lin, Y.-F. Wu, S. V. Peri, W. Sun, G. Singh, F. Deng, J. Jiang, S. Ahn, Space: Unsupervised object-oriented scene representation via spatial attention and decomposition, *arXiv preprint arXiv:2001.02407* (2020).
- [29] D. P. Kingma, M. Welling, Auto-encoding variational bayes, *arXiv preprint arXiv:1312.6114* (2013).
- [30] M. Engelcke, O. Parker Jones, I. Posner, Genesis-v2: Inferring unordered object representations without iterative refinement, *Advances in Neural Information Processing Systems* 34 (2021) 8085–8094.
- [31] G. Singh, F. Deng, S. Ahn, Illiterate dall-e learns to compose, *arXiv preprint arXiv:2110.11405* (2021).
- [32] Y. Wang, L. Liu, J. Dauwels, Slot-vae: Object-centric scene generation with slot attention, *arXiv preprint arXiv:2306.06997* (2023).
- [33] G. Singh, Y. Kim, S. Ahn, Neural block-slot representations, *arXiv preprint arXiv:2211.01177* (2022).

- [34] P. Emami, P. He, S. Ranka, A. Rangarajan, Slot order matters for compositional scene understanding, arXiv preprint arXiv:2206.01370 (2022).
- [35] A. Kori, F. Locatello, F. D. S. Ribeiro, F. Toni, B. Glocker, Grounded object-centric learning, in: The Twelfth International Conference on Learning Representations (ICLR 2024), 2024. URL: <https://openreview.net/forum?id=pBxeZ6pVUD>.
- [36] A. Kori, F. Locatello, F. Toni, B. Glocker, F. D. S. Ribeiro, Identifiable object centric representations via probabilistic slot attention, in: Part of Advances in Neural Information Processing Systems 37 (NeurIPS 2024), 2024.
- [37] M. Seitzer, M. Horn, A. Zadaianchuk, D. Zietlow, T. Xiao, C.-J. Simon-Gabriel, T. He, Z. Zhang, B. Schölkopf, T. Brox, et al., Bridging the gap to real-world object-centric learning, arXiv preprint arXiv:2209.14860 (2022).
- [38] P. Padalkar, H. Wang, G. Gupta, Nesifold: A framework for interpretable image classification, in: Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, February 20-27, 2024, Vancouver, Canada, AAAI Press, 2024, pp. 4378–4387. URL: <https://doi.org/10.1609/aaai.v38i5.28235>.
- [39] M. Gelfond, V. Lifschitz, Classical negation in logic programs and disjunctive databases, *New Generation Computing* 9 (1991) 365–385. URL: <https://doi.org/10.1007/BF03037169>.
- [40] Y. Aspis, K. Broda, J. Lobo, A. Russo, Embed2Sym - Scalable neuro-symbolic reasoning via clustered embeddings, in: Proceedings 19th International Conference on Principles of Knowledge Representation and Reasoning (KR), volume 19, 2022, pp. 421–431. URL: <https://doi.org/10.24963/kr.2022/44>, Special Session on KR and Machine Learning.
- [41] D. Thauvin, S. Herbin, W. Ouerdane, C. Hudelot, Interpretable image classification through an argumentative dialog between encoders, in: ECAI 2024 - 27th European Conference on Artificial Intelligence, volume 392 of *Frontiers in Artificial Intelligence and Applications*, 2024, pp. 3316–3323. URL: <https://doi.org/10.3233/FAIA240880>.
- [42] A. Gould, G. Paulino-Passos, S. Dadhanian, M. Williams, F. Toni, Preference-Based Abstract Argumentation for Case-Based Reasoning, in: KR, 2024, pp. 394–404. URL: <https://doi.org/10.24963/kr.2024/37>. doi:10.24963/kr.2024/37.
- [43] O. Cocarascu, A. Stylianou, K. Čyras, F. Toni, Data-empowered argumentation for dialectically explainable predictions, in: ECAI 2020, IOS Press, 2020, pp. 2449–2456.
- [44] A. Gould, F. Toni, Neuro-argumentative learning with case-based reasoning, 2025. URL: <https://arxiv.org/abs/2505.15742>. doi:10.48550/ARXIV.2505.15742.
- [45] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, *Advances in neural information processing systems* 30 (2017).
- [46] P. M. Dung, On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games, *Artificial Intelligence* 77 (1995) 321–357. URL: <https://www.sciencedirect.com/science/article/pii/000437029400041X>. doi:[https://doi.org/10.1016/0004-3702\(94\)00041-X](https://doi.org/10.1016/0004-3702(94)00041-X).
- [47] G. Paulino-Passos, F. Toni, Monotonicity and noise-tolerance in case-based reasoning with abstract argumentation, in: Proceedings of the Eighteenth International Conference on Principles of Knowledge Representation and Reasoning, 2021, pp. 508–518. URL: <https://doi.org/10.24963/kr.2021/48>. doi:10.24963/kr.2021/48.
- [48] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, J. Heming, K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data, *Information Sciences* 622 (2023) 178–210. URL: <http://dx.doi.org/10.1016/j.ins.2022.11.139>. doi:10.1016/j.ins.2022.11.139.
- [49] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
- [50] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014. URL: <https://arxiv.org/abs/1412.6980>. doi:10.48550/ARXIV.1412.6980.
- [51] A. Bondarenko, P. Dung, R. Kowalski, F. Toni, An abstract, argumentation-theoretic approach to default reasoning, *Artificial Intelligence* 93 (1997) 63–101. URL: <http://dx.doi.org/10.1016/>

- [52] K. Cyraś, X. Fan, C. Schulz, F. Toni, Assumption-Based Argumentation: Disputes, Explanations, Preferences, 2018.
- [53] E. De Angelis, M. Proietti, F. Toni, ABA learning via ASP, in: Proceedings 39th International Conference on Logic Programming, ICLP 2023, Imperial College, London, UK, 9th July 2023 - 15th July 2023, volume 385 of *EPTCS*, 2023, pp. 1–8. URL: <https://doi.org/10.4204/EPTCS.385.1>.

## A. Hyperparameters

Hyperparameters	AA-CBR Models	
	Model 1	Model 2
Characterisation	Count	Set
Model Classifies	Class 2 vs. rest	Class 1 vs. Class 0
Default Outcome	Rest	Class 1
No. K-means Centroids	900	500
Threshold	✗	0.9
Uses Supports	✗	✓

**Table 3**

SAA-CBR argumentative component’s hyperparameters for CLEVR-Hans3 dataset

Hyperparameters	AA-CBR Models		
	Model 1	Model 2	Model 3
Characterisation	Count	Count	Count
Model Classifies	Class 0 vs. rest	Class 6 vs. rest	Class 3 vs. rest
Default Outcome	Rest	Rest	Rest
No. K-means Centroids	300	500	300
Threshold	0.75	0.7	0.75
Uses Supports	✗	✗	✗
Hyperparameters	AA-CBR Models		
	Model 4	Model 5	
Characterisation	Count	Position Count	
Model Classifies	Class 1 vs. rest	Class 4 vs. Class 5	
Default Outcome	Class 1	Class 4	
No. K-means Centroids	300	300	
Threshold	0.85	0.75	
Uses Supports	✗	✗	

**Table 4**

SAA-CBR argumentative component’s hyperparameters for CLEVR-Hans7 (modified) dataset

Note that the difference between the Count characterisation and the Position Count characterisation is that Position Count also considers whether an object is on the left or right side of the image as part of the feature combination process, whereas the Count characterisation does not. Note that only Class 5 in CLEVR-Hans 7 (modified) uses the absolute position of the object, so it makes sense that our hyperparameter tuning would find the best performance is to only use the Position Count partial order when classifying this class.

## B. Ablation Study

We conducted an ablation study with the CLEVR-Hans3 dataset to assess which aspects of our integrations to the SAA-CBR pipeline led to improved model performance. We choose to explore how the



model would perform without using a feature combination approach, without uncertainty thresholding to reduce the casebase size and without using the support relations. Note that we chose not to include the k-means clustering from the ablation study, as using the full casebase for training the AA-CBR models was infeasible given the  $O(n^3)$  time complexity.

Methods	Test Set			
	Accuracy	Precision	Recall	F1
SAA-CBR	$75.13 \pm 2.16$	$75.31 \pm 2.10$	$75.13 \pm 2.16$	$74.77 \pm 2.28$
w/o feature combination	$54.68 \pm 3.69$	$62.97 \pm 1.53$	$54.68 \pm 3.69$	$47.11 \pm 3.92$
w/o thresholding	$68.04 \pm 4.61$	$70.70 \pm 6.00$	$68.04 \pm 4.61$	$66.02 \pm 4.76$
w/o supports	$74.30 \pm 2.96$	$74.79 \pm 2.39$	$74.30 \pm 2.96$	$73.73 \pm 3.44$
w/o supports, thresholding, feature combination	$55.16 \pm 3.81$	$62.37 \pm 2.12$	$55.16 \pm 3.81$	$48.88 \pm 4.76$

**Table 5**

Ablation study of the SAA-CBR model on the CLEVR-Hans3 test set

Table 5 shows our results. From our ablation study on CLEVR-Hans3, we observe that the largest increase in model performance was a result of the feature combination strategy, without which the addition of supports or thresholding surprisingly reduces model performance. However, when used in combination, we see that both the thresholding and the support relation are necessary for improving the model’s classification performance. Interestingly, the model’s variance is considerably lower when all of these integration methods are applied, suggesting that the training variability is also dependent on making use of these integrations, even though the variability only comes from the initial weights of the Slot Attention module and the initial position of the k-means cluster centroids.