Concept-Centric Transformers: Enhancing Model Interpretability through Object-Centric Concept Learning within a Shared Global Workspace

Jinyung Hong[1], Keun Hee Park[1], and Theodore P. Pavlic[1, 2]
[1]School of Computing and Augmented Intelligence
[2]School of Life Sciences
Arizona State University
Tempe, AZ 85281
{ jhong53, kpark53, tpavlic }@asu.edu

Abstract

Many interpretable AI approaches have been proposed to provide plausible explanations for a model's decision-making. However, configuring an explainable model that effectively communicates among computational modules has received less attention. A recently proposed shared global workspace theory showed that networks of distributed modules can benefit from sharing information with a bottlenecked memory because the communication constraints encourage specialization, compositionality, and synchronization among the modules. Inspired by this, we propose Concept-Centric Transformers, a simple yet effective configuration of the shared global workspace for interpretability, consisting of: i) an object-centric-based memory module for extracting semantic concepts from input features, ii) a cross-attention mechanism between the learned concept and input embeddings, and iii) standard classification and explanation losses to allow human analysts to directly assess an explanation for the model's classification reasoning. We test our approach against other existing concept-based methods on classification tasks for various datasets, including CIFAR100, CUB-200-2011, and ImageNet, and we show that our model achieves better classification accuracy than all baselines across all problems but also generates more consistent concept-based explanations of classification output.

1. Introduction

Although state-of-the-art machine-learning models have achieved remarkable performance across a wide range of applications, their intrinsic lack of transparency due to their many degrees of training freedom limits their usage in safety-critical areas—such as medical diagnostics, healthcare, public infrastructure safety, and visual inspection for civil engineering—where trustworthy domain-specific knowledge is crucial for decision making. Recently, several developed methods provide *post hoc* explanations that identify relevant features that a trained model uses to make predictions [58, 64, 70, 78], but these are commonly criticized for focusing only on low-level features [1, 43, 44, 79]. In contrast, *intrinsically interpretable models* [67] have been proposed to make decisions based on human-understandable "concepts," the foundation of domain expertise [1,6,12,13,25,32,42,43,45,50,65,91,94]. The gap between *post hoc* explainability and intrinsically interpretable models is also discussed in the NLP community concerning interpreting *attention mechanisms* [3]. In particular, the debate over what degree of interpretability can be ascribed to attention weights over input tokens still needs to be settled to help meet the need for interpretability of attention mechanisms [1, 9, 40, 71, 86].

Ideally, an intrinsically interpretable model will generate explanations that are compositions of individually meaningful modules. Modular explanations may improve the human understanding, and natural neuronal systems that continue to inspire AI development are often described as having modular architectures themselves [4, 5, 8, 66]. Thus, structuring algorithms to promote learning of modular latent structures may also lead to better overall performance. Motivated by improving modularity in interpretable models, we propose the *Concept-Centric Transformer (CCT)*, a framework of intrinsically interpretable models inspired by the Shared Global Workspace (SGW) [30], a new conceptual framework meant to generally encourage modularity by forcing parallel specialized components to compete for bottlenecked access to a shared memory. The configuration of CCT allows trained models to have simple, modular structures that can extract semantic concepts with or without the guidance of ground-truth explanations of the concepts.

In what follows, we frame our CCT as a novel extension of the SGW concept to interpretable model development, and we describe how CCT is implemented using three key components: i) **Concept-Slot-Attention (CSA) module** that interfaces with image embedding from a backbone model and produces a set of task-dependent embeddings for concepts, ii) **Cross-Attention (CA) module** that generates classification outputs using cross-attention between input features and the CSA module's embeddings, and iii) specialized loss penalties, including **Explanation Loss**, when expert's knowledge can be leveraged, and **Sparsity Loss**, an entropy-based loss to enforce the sparsity to determine the importance of features during training. The CCT architecture is designed to augment existing deep-learning backbones to add explainability to them. Consequently, we validate our approach on three image benchmark datasets—CIFAR100 Super-class [23], CUB-200-2011 [83], and ImageNet [19]—combining it with various deep-learning backbones, such as Vision Transformer (ViT) [22], Swin Transformer (SwinT) [55], and ConvNeXt [56].

2. Related Work

Significant advances have recently been made in devising explainable and interpretable models to measure the importance of individual features for predictive output. The *post hoc* analysis is one general approach to analyzing a trained model by matching explanations to classification outputs [1, 58, 64]. For example, activation maximization [62, 81, 92] and saliency visualization [70, 78, 80] are well-known methods for CNNs. Attention-based interpretable approaches have also been introduced to identify the most relevant parts of the input that the network focuses on when making a decision [24, 26, 27, 37, 96–99].

In addition, designing methods that explain predictions with high-level, human-understandable concepts [6, 12, 13, 25, 32, 42, 43, 45, 50, 65, 89, 91, 94] is one of the recent advancements in the field of interpretability. These intrinsically interpretable methods focus on identifying common activation patterns in the nodes of the last layer of the neural network corresponding to human-understandable categories or constraining the network to learn such concepts. Among them, our work is most similar to *Concept Transformers (CTs)* [65], a framework that learns high-level concepts defined with a set of related dimensions. Those concepts, which can be part-specific or global, typically can boost the performance of the learning task while offering explainability at no additional cost to the network. However, that approach relies on extracting concepts based on provided image patches even though each image patch may be an unreliable predictor of high-level concepts. Our CCT formulation generalizes this approach beyond image patches.

Historically, it has been argued that it is better to build an intelligent system from many interacting specialized modules rather than a single "monolithic" entity to deal with a broad spectrum of conditions and tasks [29, 59, 66]. Thus, there has been significant effort on synchronization between computationally specialized modules via a shared global workspace [16, 31, 39, 60, 69]. Furthermore, work on the integration of modular computational architectures with working memories takes inspiration from biology, neuroscience, and cognitive science [35, 36, 63, 87]. The recently proposed shared global workspace [30] shows how to utilize the attention mechanism to encourage the most helpful information to be shared among neural modules in modern AI frameworks. This approach is the inspiration for our use of an explicit working memory in the CCT to improve the generalization of Transformer- and object-centric-based models in the context of explainable models.

3. Preliminary

The Shared Global Workspace in AI Models. Inspired by the Global Workspace Theory (GWT) from cognitive science [2, 17, 18, 72–75], the Shared Global Workspace (SGW) [30] explores how GWT can be manifested in modern AI models to possess communication and coordination schemes where several sparsely communicating *specialists* (specific computing modules dealing with the input) interact via a *shared workspace* (a shared working memory module). To do so, the transformer and slot-based methods were extended by adding a shared workspace and allowing the modules to compete for write access in each computational stage (Fig. 1). Replacing pairwise communications among the modules with interaction facilitated by the shared workspace allows for: i) higher-order interaction among the modules, ii) dynamic filtering due to the memory persistence, and iii) computational sophistication of using shared workspace for synchronizing different specialists. Motivated by the SGW, we aim to discover an efficient configuration for intrinsically interpretable AI frameworks and propose a simple but efficient way of interacting between a shared working space and specialists to improve interpretability and performance.

Variants of Slot-Attention. Due to its simple yet effective design, Slot-Attention (SA) [57] has gained significant attention in unsupervised object-centric learning to mimic the development of symbolic understanding in human cognition. The iterative attention mechanism allows SA to learn and compete between slots for explaining parts of the input, showing a soft clustering effect on visual inputs [57]. However, as revealed by recent studies, the vanilla SA module as innately limited in that: i) the random initialization for slots hampers addressing object-binding in input and ii) it

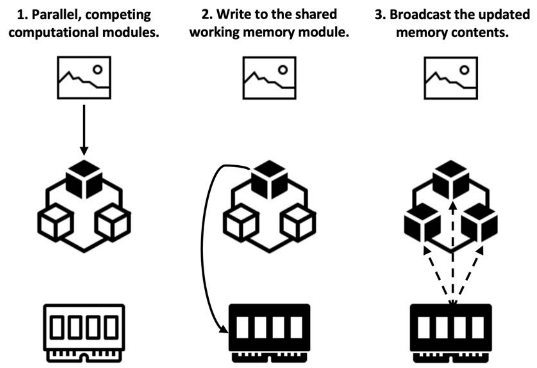

1. Parallel, competing computational modules.
2. Write to the shared working memory module.
3. Broadcast the updated memory contents.

Figure 1. The shared global workspace [30] emerges from three steps: 1) A collection of computational modules (or *specialists*) perform standard processing, and a subset of the specialists becomes active at a particular computational stage depending upon the input; 2) The active specialist writes information in a shared working memory module (or *shared workspace*); 3) The updated contents of the workspace are broadcast to all specialists. We explore how these steps can be used to add explainability in AI frameworks. The generic figure above is inspired by [30, Fig. 1].

heavily depends on hyperparameter tuning so that it cannot generally be applicable in many domains. Thus, some variants of SA, including I-SA [10] and BO-QSA [41], have been proposed recently to address those issues[1].

There are several existing examples on leveraging slot-based methods in explainable models to extract semantic concepts [49, 84]. However, few studies have emphasized the perspective of modular architecture to foster communication between the slot-based model and other modules. We leverage the three SA variants above as the shared workspace of the SGW and explore how to encourage the interactions among them to achieve better interpretability and performance.

4. Concept-Centric Transformers

For supervised classification tasks, we introduce Concept-Centric Transformers (CCTs), an instantiation of the SGW for configuring an intrinsically interpretable model, and we will describe the connection between the SGW steps (Fig. 1) and our formulation. Our model consists of: i) *Concept-Slot-Attention (CSA)* module that acts as a shared memory module and extracts the latent concept embedding specific to each batch of input, ii) *Cross-Attention (CA)* module for broadcasting between input embedding and the extracted concept embedding from the CSA module so that it produces classification output as well as faithful and plausible concept-based explanations and encourages pairwise interactions among them, and iii) specialized losses, including *Explanation Loss* and *Sparsity Loss*, which is our information broadcast scheme to encourage interpretability. The CCT architecture, summarized in Fig. 2, is described in the following sections. Further details, including limitations (Appendix D), are given in the appendix.

4.1. Between Specialists and the Shared Workspace

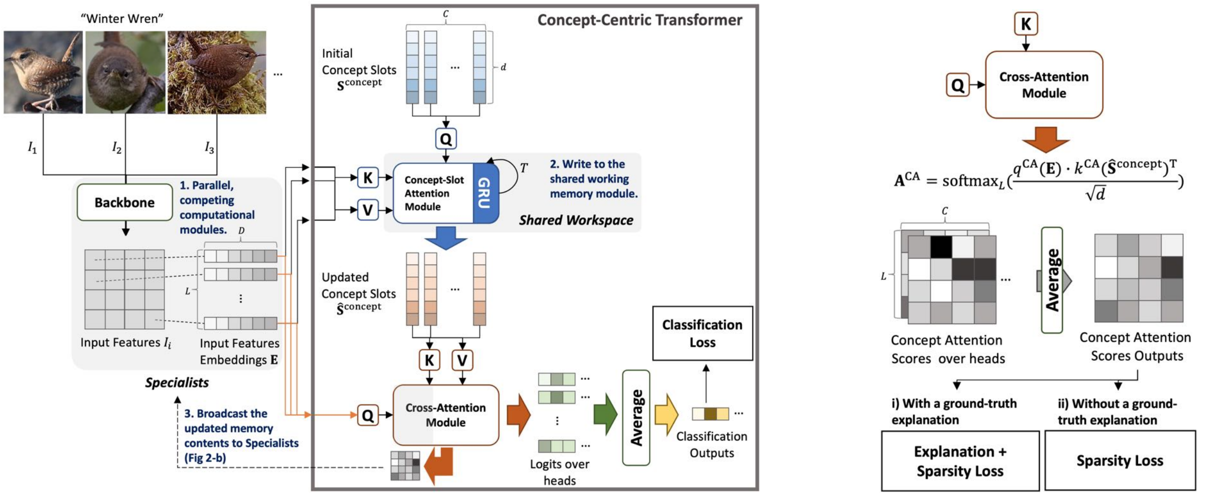
Following the general structure of the SGW, we leverage *specialists*, which are the computational modules of our backbone, and a *shared workspace* by utilizing slot-based methods in our CSA module, including SA [57], I-SA [10], and BO-QSA [41]. Because of the modularity in our formulation, those three SA variants are interchangeable, and we demonstrate the performance comparison among them in our experiments. We use a conventional key–query–value attention mechanism to implement the competition between specialists to write into the workspace, similar to the SGW. The CSA module encodes a set of L input feature vectors $\mathbf{E}$ into concept representations $\mathbf{S}^{\texttt{concept}}$, which we refer to as *concept slots*.

Concept Binding Specific to Each Input Batch. With the number of concepts C, the concept slots $\mathbf{S}^{\texttt{concept}} \in \mathbb{R}^{C\times d}$ first perform competitive attention [57] on the input features $\mathbf{E} \in \mathbb{R}^{L\times D}$. For this, we apply linear projection $q^{\texttt{CSA}}$ on the concept slots to obtain the queries and projections $k^{\texttt{CSA}}$ and $v^{\texttt{CSA}}$ on the inputs to obtain the keys and the values, all having the same size d [2]. Then, we perform a dot product between the queries and keys to get the attention matrix $\mathbf{A}^{\texttt{CSA}} \in \mathbb{R}^{C\times L}$. In $\mathbf{A}^{\texttt{CSA}}$, each entry $\mathbf{A}^{\texttt{CSA}}_{c,l}$ is the attention weight of concept slot c for attending over the input vector l. We normalize $\mathbf{A}^{\texttt{CSA}}$ by applying softmax across concept slots, i.e., along the axis C. This implements a form of competition among slots for attending to each input l.

We then seek to group and aggregate the attended inputs and obtain the attention readout for each concept slot. Intuitively, this represents how much the attended inputs contribute to semantically representing each concept. For this, we normalize the attention matrix $\mathbf{A}^{\texttt{CSA}} \in \mathbb{R}^{C\times L}$ along the axis L and multiply it with the input values $v^{\texttt{CSA}}(\mathbf{E}) \in \mathbb{R}^{L\times d}$. This produces the attention readout in the form of a matrix $\mathbf{U} \in \mathbb{R}^{C\times d}$ where each row $u_c \in \mathbb{R}^d$ is the readout corresponding to concept slot c; $\mathbf{A}^{\texttt{CSA}} = \mathrm{softmax}_C(q^{\texttt{CSA}}(\mathbf{S}^{\texttt{concept}})\cdot k^{\texttt{CSA}}(\mathbf{E})^\top/\sqrt{d})$, $\mathbf{A}^{\texttt{CSA}}_{c,l} = \mathbf{A}^{\texttt{CSA}}_{c,l}/\sum_{l=1}^{L}\mathbf{A}^{\texttt{CSA}}_{c,l}$, and $\mathbf{U} = \mathbf{A}^{\texttt{CSA}} \cdot v^{\texttt{CSA}}(\mathbf{E})$. We use the readout information obtained from concept binding and update each concept slot. The aggregated updates $\mathbf{U}$ are finally used to update the concept slots via a learned recurrent function, for which we use a Gated Recurrent Unit (GRU) [14] with d hidden units so that $\mathbf{S}^{\texttt{concept}} = \texttt{GRU}(\mathbf{S}^{\texttt{concept}}, \mathbf{U})$. The processes above form one refinement iteration. The concept slots obtained from the last iter-

[1] We omit details of their technical differences as they are out of scope.

[2] For simplicity, embedding size d is shared equally in our method.


"Winter Wren"
...
I_1
I_2
I_3
Backbone
1. Parallel, competing computational modules.
D
L
Input Features I_i
Input Features Embeddings E
Specialists
3. Broadcast the updated memory contents to Specialists (Fig 2-b)
Concept-Centric Transformer
Initial Concept Slots $S^{concept}$
C
d
Q
K
V
Concept-Slot Attention Module
GRU
T
2. Write to the shared working memory module.
Shared Workspace
Updated Concept Slots $\hat{S}^{concept}$
K
V
Q
Cross-Attention Module
Logits over heads
Average
Classification Loss
Classification Outputs
K
Q
Cross-Attention Module
$A^{CA} = \mathrm{softmax}_L(\frac{q^{CA}(E) \cdot k^{CA}(\hat{S}^{concept})^T}{\sqrt{d}})$
C
L
Concept Attention Scores over heads
Average
Concept Attention Scores Outputs
i) With a ground-truth explanation
ii) Without a ground-truth explanation
Explanation + Sparsity Loss
Sparsity Loss

(a) Concept-Centric Transformers via a Shared Workspace

(b) Interpretable Broadcast to Specialists

Figure 2. Overall architecture of Concept-Centric Transformers (CCTs). (a) The CCT (in a solid gray line) is a drop-in replacement for the classifier head of any backbone architecture, such as a ViT or a CNN, and consists of two modules: (1) Concept-Slot-Attention module and (2) Cross-Attention module. (b) The process of training uses losses to induce an interpretable broadcast scheme.

ation are considered final. The overall module is described in Algorithm 1 in pseudo-code in Appendix B.

4.2. Broadcast Updated Memories to Specialists

At this stage of the SGW, each specialist must update its status using information broadcast from the shared workspace. We also leverage the cross-attention mechanism (called the CA module) to make specialists queries (step 3 in Figs 1 and 2) and perform dot products between them and the values from the updated concept slots to update the state of each specialist. However, because we aim to configure an interpretable model and perform classification tasks, we modify the iterative process by combining it with our desired downstream classification task: i) guided by expert knowledge if ground-truth concept explanations are available, or ii) using only sparsity loss to enforce minimizing the entropy of the broadcasting information.

A set of L input feature vectors $\mathbf{E} \in \mathbb{R}^{L \times D}$ are reused with a linear projection $q^{\mathtt{CA}}$ to attain the queries, and projections $k^{\mathtt{CA}}$ and $v^{\mathtt{CA}}$ are applied to the extracted concept slots with position embedding $\hat{\mathbf{S}}$ from the CSA module. The resulting keys and values are used in a cross-attention mechanism with the queries, and the cross-attention then outputs an attention weight $\mathbf{A}^{\mathtt{CA}} = \mathrm{softmax}_L \left(q^{\mathtt{CA}}(\mathbf{E}) \cdot k^{\mathtt{CA}}(\hat{\mathbf{S}}^{\mathtt{concept}})^\top / \sqrt{d} \right) \in \mathbb{R}^{L \times C}$ between each patch–concept slot pair. The final output of the CA module is the product obtained by multiplying the attention map $\mathbf{A}^{\mathtt{CA}}$, the values $v^{\mathtt{CA}}(\hat{\mathbf{S}}^{\mathtt{concept}}) \in \mathbb{R}^{C \times d}$, and an output matrix $\mathbf{O} \in \mathbb{R}^{d \times n_c}$ that projects onto the (unnormalized) n_c logits over the output classes and then averaging over input features[3]; that is, for $i = 1, \ldots, n_c$,

$$\mathrm{logit}_i = \frac{1}{L} \textstyle\sum_{l=1}^{L} \mathbf{A}_l^{\mathtt{CA}} \cdot v^{\mathtt{CA}}(\hat{\mathbf{S}}^{\mathtt{concept}}) \cdot \mathbf{O}_{:,i} \qquad (1)$$

So, given an input $\mathbf{x}$ to the network, the conditional probability of output class $i \in \{1, \ldots, n_c\}$ is:

$$\Pr(i|\mathbf{x}) = \mathrm{softmax}_i \left(\textstyle\sum_{c=1}^{C} \beta_c \gamma_c(\mathbf{x}) \right) \qquad (2)$$

with β_c components $\beta_{ci} := (v^{\mathtt{CA}}(\hat{\mathbf{S}}^{\mathtt{concept}}) \cdot \mathbf{O})_{c,i}$ where $\gamma_c(\mathbf{x})$ are non-negative relevance scores that depend on $\mathbf{x}$ through the averaged attention weights; that is, $\gamma_c(\mathbf{x}) = (1/L) \sum_{l=1}^{L} \mathbf{A}_{l,c}^{\mathtt{CA}}$. We can interpret the equations above from the two following perspectives:

1) Faithful Concept-slot-based Explanations by Design. The CA module output is a multinomial logistic regression model over positive variables $\gamma_c(\mathbf{x})$ that measures the contribution of each concept slot. *Faithfulness* is the degree to which explanation reflects the decision and aims to ensure that the explanations are indeed explaining model operation [33,48]. As shown in [65], the result of the CA module follows the linear relation between the value vectors and the classification logits and comes from the design choices of computing outputs from the value matrix $v^{\mathtt{CA}}(\hat{\mathbf{S}}^{\mathtt{concept}})$ through the linear projection $v^{\mathtt{CA}}(\hat{\mathbf{S}}^{\mathtt{concept}}) \cdot \mathbf{O}$ and aggregating patch contributions by averaging. So, our CA module is also guaranteed to be faithful by design by satisfying Proposition 1 in [65] and the technical definitions of *faithfulness* from [1].

[3] For simplicity, we describe a single-head attention model here; a multi-head version [82] is available and is also used in our experiments.

2) Dynamic State Update for Specialists with Information Broadcast. In the original definition of the SGW, $\mathbf{A}^{\texttt{CA}} \cdot v^{\texttt{CA}}(\hat{S}^{\texttt{concept}})$ is the formal computation of the update for specialists (step 3 from [30, Sec 2.1]). Instead of applying an additional iterative process of updating specialists, Eqs 1 and 2 are to produce classification outputs using the weight $\mathbf{O}$. So, the attention mask $\mathbf{A}^{\texttt{CA}}$ can contain not only information from the updated memory but also classification error. Furthermore, by directly manipulating the mask $\mathbf{A}^{\texttt{CA}}$, we finally define explanation loss and sparsity loss to enhance the model's explainability.

4.3. Training Objectives for Interpretability

Plausibility by Construction with Explanation Loss. *Plausibility* refers to how convincing the interpretation is to humans [9,33]. To provide plausible human-understandable explanations, we leverage the idea of explicitly guiding the attention heads to focus on concepts in the input based on domain expertise that are important for correctly classifying the input. Similar to [20], given a desired distribution of attention $\mathbf{H}$ provided by domain knowledge, the attention weights from the CA module of the CCT $\mathbf{A}^{\texttt{CA}}$ are used as a regularization term by adding an *explanation cost* to the objective function that is proportional to $\mathcal{L}_{expl} = \|\mathbf{A} - \mathbf{H}\|_F^2$, where $\|\cdot\|$ is the Frobenius norm. The ground-truth explanation $\mathbf{H}$ can indicate global (e.g., sub-class or dominant attribute of a bird) or spatial/image-patch-level information (e.g., eye color of a bird). Below, we demonstrate the effectiveness of this loss in the experiments of CIFAR100 Super-class and CUB-200-2011.

Sparsity Loss based on Entropy. The advantage of richer, more informative labels can increase interpretability, but this comes at the expense of additional annotation effort. A methodology that can bypass this trade-off would be particularly worthwhile. We can attain this capability through our configuration that involves interactions between specialists and a shared workspace. We introduce the sparsity loss based on minimizing the entropy of the attention mask $\mathbf{A}^{\texttt{CA}}$; $\mathcal{L}_{sparse} = H(\mathbf{A}) = H(a_1, \ldots, a_{|\mathbf{A}|}) = (1/|\mathbf{A}|) \sum_i -a_i \cdot \log(a_i)$. This loss can be used in the experiments with/without the ground-truth explanations.

Final Loss. Thus, the final loss to train the model becomes $\mathcal{L} = \mathcal{L}_{cls} + \lambda_{expl}\mathcal{L}_{expl} + \lambda_{sparse}\mathcal{L}_{sparse}$, where $\mathcal{L}_{cls}$ denotes the conventional classification loss. Notice that the constant $\lambda_{expl} \geq 0$ controls the relative contribution of the explanation loss to the total loss so our model can be applied with ground-truth explanations ($\lambda_{expl} > 0$) or without them ($\lambda_{expl} = 0$). Finally, the constant λ_{sparse} handles the intensity of sparsity loss. Our experiments demonstrate that our model can perform well without using additional complicated losses, such as contrastive or reconstruct losses.

Model	F.C. Acc. (%)	S.C. Acc. (%)
Vanilla ResNet†	NA	$83.2_{\pm 0.2}$
Vanilla ViT-T	NA	$86.2_{\pm 0.3}$
Hierarchical Model†	$71.2_{\pm 0.2}$	$84.7_{\pm 0.1}$
DL2† [23]	$75.3_{\pm 0.1}$	$84.3_{\pm 0.1}$
MultiplexNet† [34]	$74.4_{\pm 0.2}$	$85.4_{\pm 0.3}$
PIP-Net [61]	NA	$83.9_{\pm 0.2}$
ProtoPFormer [89]	NA	$81.7_{\pm 0.1}$
ProtoPool [68]	NA	$82.9_{\pm 0.4}$
ProtoPNet [11]	NA	$82.3_{\pm 0.1}$
Deform-ProtoPNet [21]	NA	$83.7_{\pm 1.0}$
BotCL [84]	NA	$56.9_{\pm 10}$
CT [65]	$73.3_{\pm 2.9}$	$92.1_{\pm 0.2}$
CCT: ViT-T+SA	$80.3_{\pm 0.4}$	$92.6_{\pm 0.1}$
CCT: ViT-T+I-SA	$83.3_{\pm 0.1}$	$92.8_{\pm 0.1}$
CCT: ViT-T+BO-QSA	$\mathbf{83.4}_{\pm 0.1}$	$\mathbf{93.0}_{\pm 0.1}$

Table 1. Test accuracy on fine-grained class (F.C.) and super-class (S.C.) label prediction on CIFAR100. Notice that the classification of super-classes and fine-grained classes are performed simultaneously and that this kind of experiment can be done in deep learning with constraints, but our method and CT are only among concept-based approaches. † indicates results from [34].

5. Experiments

We evaluate the performance of our CCT on three distinct datasets: CIFAR100 Super-class [23], CUB-200-2011 [83], and ImageNet [19]. Specific objectives guide the selection of these datasets. Firstly, the CIFAR100 Super-class is a testing ground to demonstrate our model's exceptional capabilities under fully supervised conditions, where complete global concept explanations are available. Secondly, CUB-200-2011 acts as an intermediary dataset, allowing us to showcase our model's prowess in both supervised and unsupervised explanation setups. Lastly, we challenge our CCT with the ImageNet dataset, which represents a fully unsupervised scenario due to the absence of concept explanations. Despite this hurdle, we adapt our model to achieve remarkable performance even without any concept explanations. This comprehensive setup underscores our CCT's adaptability and versatility across various use cases, visual backbones, and data scenarios. Our experimental results are robustly validated through three different random seeds and 95% confidence intervals, with additional details provided in Appendix C.

5.1. Evaluation on CIFAR100 Super-class

The CIFAR100 Super-class dataset is a variant of the CIFAR100 [47] image dataset. It consists of 100 fine-grained classes (F.C.) of images that are further grouped into 20 super-classes (S.C.). For instance, the five fine-grained classes *baby, boy, girl, man,* and *woman* belong to the super-class *people*. Since the introduction of *deep-learning models with logical constraints* [23], this dataset has been used as one of the benchmark datasets to assess the effectiveness of embedding the constraints into neural networks (in-depth surveys can be found in [15, 28]). Our study highlights that the concepts within our CCT (and closely related CT [65]) can be understood as *constraints*, with differences outlined in Appendix C.4. We leverage individual fine-grained image classes as global concept explanations for a multi-class prediction task with 20 super-classes, employing a Vision Transformer-Tiny (ViT-T)[4] as the backbone[5].

Following [34], our CCT's performance is compared against three baseline groups. The first group includes vanilla backbone models like *Wide ResNet 28-10* [93] and *ViT-T*. The second group involves neural-network models with logical constraints, including *Hierarchical Model* [34], *DL2* [23], and *MultiplexNet* [34]. The third group comprises concept-based explainable models; except for CT, these models cannot handle both fine-grained and super-class tasks concurrently. For our CCT's CSA module, we configure three SA variants—SA, I-SA, and BO-QSA—with the backbone and evaluate fine-grained class accuracy by comparing ground-truth concept explanations with top-1 predicted concepts based on the resulting attention scores.

Table 1 presents the experimental outcomes, showing our CCT's substantial outperformance of all baselines. It significantly enhances ViT-T's backbone performance and achieves a remarkable increase in fine-grained class accuracy compared to CT. Notably, CT performs less effectively than logic-constraint-based approaches, highlighting our module's superior role in shaping global concepts.

Figure 3 shows two examples that demonstrate where our CCT's concept learning excels both quantitatively and qualitatively over CT. CT's poor fine-grained class accuracy might be the result of making *hallucinations* to greedily achieve super-class accuracy without forming sound latent concepts. Though both models correctly classify super-classes on the shown examples, the explanatory decision-making processes perform entirely differently in the two models. Although the ground-truth class of the left image in Fig. 3 is *boy*, the best-matching class concept with the highest attention score by CT was *baby*, which is the incorrect fine-grained class but also belongs to the correct super-class *people*. In the right example image, we observe a similar behavior of CT. In contrast, our CCT's predicted concepts for both examples correctly match their ground-truth fine-grained classes with sparser concept attention scores than CT. Further details and results are described in Appendix C.4.

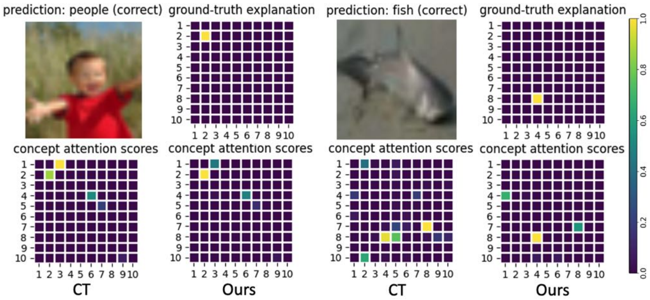


Figure 3. Comparison of class predictions for CT [65] and CCT ("Ours") in examples where both make correct CIFAR100 super-class predictions. The 100 classes are indexed from 1 (top left in 10x10 grid) to 100 (bottom right in 10x10 grid). (**Left**) Ground-truth class label is *boy* (12), but CT mispredicted as *baby* (3), whereas CCT's prediction is correct. (**Right**) Ground-truth label is *shark* (84), but CT incorrectly selects *ray* (78) whereas CCT again makes a correct class prediction.

5.2. Evaluation on CUB-200-2011

The CUB-200-2011 [83] dataset comprises 11,788 bird images categorized into 200 species. Each image is annotated with various discrete concepts, e.g., the shape of the beak, or the color of the body, aiding species identification. The dataset involves 312 concepts distributed unevenly across images, so we utilize a pre-processing method from [65] to address this. Additional results and details are described in Appendix C.5.

With Concept Explanations. We consider a real-world scenario where many-to-many and non-deterministic relationships between concepts and outputs exists, along with a mix of *global* and *spatially localized* concepts. We use CSA and CA modules within CCT to handle both global and spatial concepts, averaging their logits for interpretability. We use various backbones, including ViT [22], Swin Transformer (SwinT) [55], and ConvNeXt [56] for this dataset. We use the embeddings of the tokenized image patches, while as input to the CCT in charge of the concepts, we use the embedding of the CLS token.

In Table 2, we compare our CCT with other methods based on *Multi-stage* (i.e., complex training) and *End-to-end* (i.e., training with backpropagation) training. All of the configurations of our CCT achieve over 87% classification accuracy, clearly outperforming other approaches. This confirms that the overall configuration of CCT enhances classification performance. Notably, our model surpasses non-interpretable baselines (B-CNN) and methods requiring complex training.

Figure 4 highlights the distinction between CT and our

[4] For this experiment, Swin Transformer and ConvNeXt were not used as a backbone for our model because the number of parameters of two models (both SwinT-T and ConvNeXt-T are 28M) is larger than one of ViT-T.

[5] For global concepts, we use the embedding of the CLS token as inputs.

Method	Test Accuracy (%)		
Multi-stage	Part R-CNN: 76.4 SPDA-CNN: 85.1 2-level attn.: 77.9 ProtoPNet: 84.8	PS-CNN: 76.2 PA-CNN: 82.8 FCAN: 82.0 Deform-ProtoPNet : 86.5	PN-CNN: 85.4 MG-CNN: 83.0 Neural const.: 81.0 PIP-net : $84.3_{\pm 0.2}$
End-to-end	B-CNN: 85.1 ST-CNN: 84.1 CEM: 77.1	CAM: 70.5 MA-CNN: 86.5 ProtoPFormer: 84.9	DeepLAC: 80.3 RA-CNN: 85.3 CT (w/w.o): $86.4_{\pm 0.2}/75.4_{\pm 0.3}$
CCT (ours)	ViT-L+SA: $90.0_{\pm 0.3}$ SwinT-L+SA: $90.7_{\pm 0.02}$ ConvNeXt-L+SA: $87.8_{\pm 0.3}$	ViT-L+I-SA: $90.3_{\pm 0.3}$ SwinT-L+I-SA : $90.9_{\pm 0.4}$ ConvNeXt-L+I-SA: $89.3_{\pm 0.6}$	ViT-L+BO-QSA: $90.3_{\pm 0.1}$ SwinT-L+BO-QSA (w/w.o): $\mathbf{91.2}_{\pm 0.2}/90.9_{\pm 0.4}$ ConvNeXt-L+BO-QSA : $89.4_{\pm 0.4}$

Table 2. Performance comparison on CUB-200-2011. For B-CNN [53], Part R-CNN [96], PS-CNN [37], PN-CNN [7], SPDA-CNN [95], PA-CNN [46], MG-CNN [85], 2-level attn. [88], FCAN [54], Neural const. [76], ProtoPNet [12], CAM [98], DeepLAC [52], ST-CNN [38], MA-CNN [97], and RA-CNN [24], performance from [65]. For ProtoPFormer [89] and CEM [94], best performance directly from their works. For CT [65] and CCT, results from our evaluation. The number of parameters of the backbone we use is: 307M (ViT-L), 197M (SwinT-L), and 197M (ConvNeXt-L). (w/w.o) indicates the performance with/without ground-truth concept explanations.

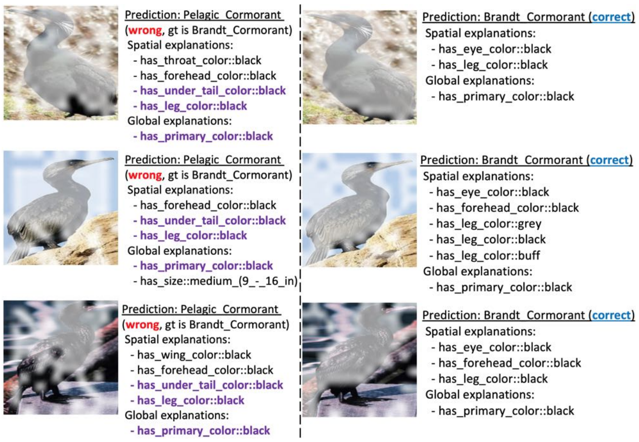

Prediction: Pelagic_Cormorant
(wrong, gt is Brandt_Cormorant)
Spatial explanations:
- has_throat_color::black
- has_forehead_color::black
- has_under_tail_color::black
- has_leg_color::black
Global explanations:
- has_primary_color::black
Prediction: Brandt_Cormorant (correct)
Spatial explanations:
- has_eye_color::black
- has_leg_color::black
Global explanations:
- has_primary_color::black
Prediction: Pelagic_Cormorant
(wrong, gt is Brandt_Cormorant)
Spatial explanations:
- has_forehead_color::black
- has_under_tail_color::black
- has_leg_color::black
Global explanations:
- has_primary_color::black
- has_size::medium_(9_-_16_in)
Prediction: Brandt_Cormorant (correct)
Spatial explanations:
- has_eye_color::black
- has_forehead_color::black
- has_leg_color::grey
- has_leg_color::black
- has_leg_color::buff
Global explanations:
- has_primary_color::black
Prediction: Pelagic_Cormorant
(wrong, gt is Brandt_Cormorant)
Spatial explanations:
- has_wing_color::black
- has_forehead_color::black
- has_under_tail_color::black
- has_leg_color::black
Global explanations:
- has_primary_color::black
Prediction: Brandt_Cormorant (correct)
Spatial explanations:
- has_eye_color::black
- has_forehead_color::black
- has_leg_color::black
Global explanations:
- has_primary_color::black

Figure 4. Prediction comparison between CT and our CCT with ground-truth explanation. (**Left**) All CT's predictions are incorrect. The highlighted explanations in purple are key attributes typically present when correctly classifying `Pelagic_Cormorant`, the incorrect label. (**Right**) Our CCT's predictions are correct.

CCT, illustrating that CT lacks learning global concepts. We emphasize concepts with the highest attention scores in all images. The figure shows a case where the CT's predictions were completely incorrect by converging to the single wrong label. Although the ground-truth class of the images was `Brandt_Cormorant`, all CT's predictions were `Pelagic_Cormorant`. Furthermore, the attributes CT used to make the incorrect classification included a subset of attributes (purple-colored attributes in Fig. 4) typically associated with correct predictions of the incorrect `Pelagic_Cormorant` label. Thus, the image-patch-centric CT hallucinates key aspects associated with incorrect labels, whereas our CCT shows more robust concept explanations of correct classifications.

Without Concept Explanations. Although the domain expert's knowledge is the most effective tool for guiding a model's explainability, the pre-precessing step to define the visual concepts for tasks may require time-consuming labeling and rely on human judgment. Importantly, we demonstrate that our CCT also works effectively without concept explanations, a capability not shared by other models. We can easily set up our loss with $\lambda_{expl} = 0$ (Final Loss defined in section 4.3) and evaluate our model as the same hyperparameter setups of the experiment with explanation using only classification loss and sparsity loss.

In Table 2, our CCT's best configuration (Swin-L+BO-QSA) without explanation achieved 90% test accuracy, which is very marginal compared to the one with explanation. In contrast, CT can also be trained without explanation, but its performance is starkly degraded.

Figure 5 visualizes the activation of the learned latent concepts in our model from the dataset. Following [84], we additionally trained our CCT by setting the number of class labels (n) to 50 and the number of latent concepts (k) to 20. The figure showcases five latent concepts—7, 3, 19, 5, and 4. In our experimental results, we identified key concepts that CCT focuses on when classifying bird images. Concept 7 focuses on the beak and upper torso, while Concept 3 considers multiple features like the beak, eyes, and tail. Concept 19 captures unique head and feather structures, Concept 5 outlines the bird's entire body, and Concept 4 isolates birds from complex backgrounds. These demonstrate the model's nuanced focus and classification skills, even in an unsupervised environment where we don't have access to its concept explanation.

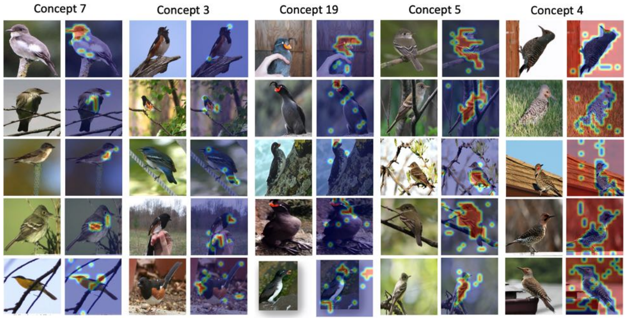


Figure 5. The activation of latent concepts learned from CUB-200-2011 and examples showing each latent concept. Further results can be found in Appendix C.5.

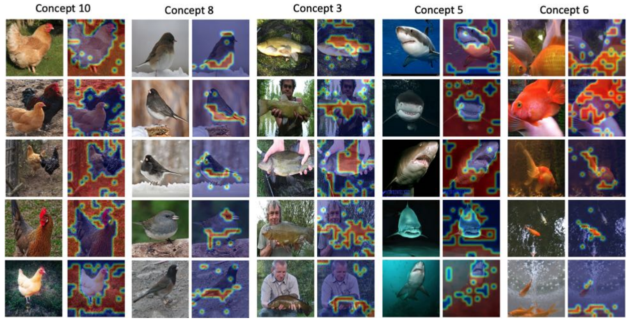


Figure 6. The activation of latent concepts learned from ImageNet and examples showing each latent concept. Further results can be found in Appendix C.6.

Model	Test Acc. (%)
Vanilla ViT-S	$83.3_{\pm 0.2}$
ProtoPFormer(Deit-B) [89]	$83.4_{\pm 2.2}$
ProtoPool [68] (ResNet-101)	$76.5_{\pm 0.8}$
ProtoPNet [11] (ResNet-101)	$77.7_{\pm 0.3}$
Deform-ProtoPNet [21] (ResNet-101)	$76.1_{\pm 0.3}$
CT [65] (ViT-S)	$27.0_{\pm 0.2}$
BotCL† [84] (ResNet-101)	83.0
CCT: ViT-S+SA	$76.3_{\pm 0.2}$
CCT: ViT-S+I-SA	$83.6_{\pm 0.2}$
CCT: ViT-S+BO-QSA	$\mathbf{83.7}_{\pm 0.2}$

Table 3. Test accuracy on ImageNet. We used the first 200 classes following [84]. † indicates the best result from [84]. The number of parameters is: 22M (ViT-S), 45M (ResNet-101), 50M (ConvNeXt-S), and 86M (Deit-B).

5.3. Evaluations on ImageNet

Finally, to validate that our model can learn latent concepts without explanations, we tested CCT on ImageNet following the approach in [84]. We used the relatively small ViT (ViT-S) as the backbone[6]. Additional results and details are in Appendix C.6.

Table 3 shows ImageNet classification performance. The combination of ViT-S with BO-QSA in the experiment has the best performance, achieving 83.7% test accuracy despite using a small-sized backbone, which meets or surpasses the performance of other SOTA methods—including some concept-based approaches that were not applicable or achieved very poor performance.

In addition, we trained a simple CCT model by setting the number of classes (n) to 20 and the number of latent concepts (k) to 10 as in [84] to visualize the consistency of the learned concepts. Figure 6 depicts five concepts we selected and five pairs of representative images for each latent concept learned from ImageNet. Our CCT shows that it excels in isolating and emphasizing specific object features across various concepts. Concept 10 isolates hen contours by highlighting the background, while Concept 8 focuses on birds' ventral regions. Concept 3 skillfully segregates fish components, Concept 5 captures shark oral regions, and Concept 6 outlines goldfish shapes. CCT's capabilities in contouring and highlighting semantically meaningful areas surpass those of existing models. It also excels in unsupervised image retrieval, clustering semantically similar images together as evident in Fig. 6. This showcases the strength of our concept-centric approach in generating semantically coherent results.

6. Conclusions

We proposed Concept-Centric Transformers, an intrinsically interpretable model via a Shared Global Workspace, allowing for achieving better interpretability, performance, and versatility for when expert knowledge is available or not. A natural future research direction is to extend the concept extraction module to acquire composable "pieces" of knowledge [77] and then learn the underlying composition rules or mechanisms relating the acquired pieces to each other. Those rules can be captured formally, as with first-order logic [6]. It is also promising to investigate other applications, such as model debugging and medical-image diagnosis.

Acknowledgments: Supported in part by NSF award 2223839 and USACE ERDC award W912HZ-21-2-0040.

References

[1] David Alvarez Melis and Tommi Jaakkola. Towards robust interpretability with self-explaining neural networks. *Advances in Neural Information Processing Systems*, 31, 2018. 1, 2, 4

[6] We avoided using a backbone larger than ResNet-101 (45M), which excluded SwinT-S (50M) and ConvNeXt-S (50M).

[2] Bernard J Baars. *A cognitive theory of consciousness*. Cambridge University Press, 1993. 2

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014. 1

[4] Carliss Young Baldwin and Kim B Clark. *Design rules: The power of modularity*, volume 1. MIT press, 2000. 1

[5] Dana H Ballard. Cortical connections and parallel processing: Structure and function. *Behavioral and brain sciences*, 9(1):67–90, 1986. 1

[6] Pietro Barbiero, Gabriele Ciravegna, Francesco Giannini, Pietro Lió, Marco Gori, and Stefano Melacci. Entropy-based logic explanations of neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 6046–6054, 2022. 1, 2, 8

[7] Steve Branson, Grant Van Horn, Serge Belongie, and Pietro Perona. Bird species categorization using pose normalized deep convolutional nets. *arXiv preprint arXiv:1406.2952*, 2014. 7

[8] Rodney A Brooks. Intelligence without representation. *Artificial intelligence*, 47(1-3):139–159, 1991. 1

[9] Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8):832, 2019. 1, 5

[10] Michael Chang, Tom Griffiths, and Sergey Levine. Object representations as fixed points: Training iterative refinement algorithms with implicit differentiation. *Advances in Neural Information Processing Systems*, 35:32694–32708, 2022. 3, A

[11] Chaofan Chen, Oscar Li, Alina Barnett, Jonathan Su, and Cynthia Rudin. This looks like that: deep learning for interpretable image recognition. *CoRR*, abs/1806.10574, 2018. 5, 8

[12] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. *Advances in Neural Information Processing Systems*, 32, 2019. 1, 2, 7

[13] Zhi Chen, Yijie Bei, and Cynthia Rudin. Concept whitening for interpretable image recognition. *Nature Machine Intelligence*, 2(12):772–782, 2020. 1, 2

[14] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, 2014. 3

[15] Tirtharaj Dash, Sharad Chitlangia, Aditya Ahuja, and Ashwin Srinivasan. A review of some techniques for inclusion of domain-knowledge into deep neural networks. *Scientific Reports*, 12(1):1040, 2022. 6

[16] Stanislas Dehaene and Jean-Pierre Changeux. Experimental and theoretical approaches to conscious processing. *Neuron*, 70(2):200–227, 2011. 2

[17] Stanislas Dehaene, Michel Kerszberg, and Jean-Pierre Changeux. A neuronal model of a global workspace in effortful cognitive tasks. *Proceedings of the national Academy of Sciences*, 95(24):14529–14534, 1998. 2

[18] Stanislas Dehaene, Hakwan Lau, and Sid Kouider. What is consciousness, and could machines have it? *Robotics, AI, and Humanity: Science, Ethics, and Policy*, pages 43–56, 2021. 2

[19] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2, 5

[20] Ameet Deshpande and Karthik Narasimhan. Guiding attention for self-supervised learning with transformers. *arXiv preprint arXiv:2010.02399*, 2020. 5

[21] Jon Donnelly, Alina Jade Barnett, and Chaofan Chen. Deformable protopnet: An interpretable image classifier using deformable prototypes. *CoRR*, abs/2111.15000, 2021. 5, 8

[22] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 6, B, C

[23] Marc Fischer, Mislav Balunovic, Dana Drachsler-Cohen, Timon Gehr, Ce Zhang, and Martin Vechev. DL2: training and querying neural networks with logic. In *International Conference on Machine Learning*, pages 1931–1941. PMLR, 2019. 2, 5, 6, B, C

[24] Jianlong Fu, Heliang Zheng, and Tao Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4438–4446, 2017. 2, 7

[25] Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. *Advances in Neural Information Processing Systems*, 32, 2019. 1, 2

[26] Ross Girshick. Fast R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1440–1448, 2015. 2

[27] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014. 2

[28] Eleonora Giunchiglia, Mihaela Catalina Stoian, and Thomas Lukasiewicz. Deep learning with logical constraints. *arXiv preprint arXiv:2205.00523*, 2022. 6

[29] Anirudh Goyal and Yoshua Bengio. Inductive biases for deep learning of higher-level cognition. *Proceedings of the Royal Society A*, 478(2266):20210068, 2022. 2

[30] Anirudh Goyal, Aniket Rajiv Didolkar, Alex Lamb, Kartikeya Badola, Nan Rosemary Ke, Nasim Rahaman, Jonathan Binas, Charles Blundell, Michael Curtis Mozer, and Yoshua Bengio. Coordination among neural modules through a shared global workspace. In *International Conference on Learning Representations*, 2021. 1, 2, 3, 5

[31] Anirudh Goyal, Alex Lamb, Jordan Hoffmann, Shagun Sodhani, Sergey Levine, Yoshua Bengio, and Bernhard Schölkopf. Recurrent independent mechanisms. *arXiv preprint arXiv:1909.10893*, 2019. 2

[32] Yash Goyal, Amir Feder, Uri Shalit, and Been Kim. Explaining classifiers with causal concept effect (CaCE). *arXiv preprint arXiv:1907.07165*, 2019. 1, 2

[33] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)*, 51(5):1–42, 2018. 4, 5

[34] Nick Hoernle, Rafael Michael Karampatsis, Vaishak Belle, and Kobi Gal. MultiplexNet: Towards fully satisfied logical constraints in neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 5700–5709, 2022. 5, 6, B

[35] Jinyung Hong and Ted Pavlic. Representing prior knowledge using randomly, weighted feature networks for visual relationship detection. In *Combining Learning and Reasoning: Programming Languages, Formalisms, and Representations*, 2022. 2

[36] Jinyung Hong and Theodore P. Pavlic. An insect-inspired randomly, weighted neural network with random fourier features for neuro-symbolic relational learning. In *Proceedings of the 15th International Workshop on Neural-Symbolic Learning and Reasoning (Ne/Sy 2022)*, October 25–27 2021. 2

[37] Shaoli Huang, Zhe Xu, Dacheng Tao, and Ya Zhang. Part-stacked CNN for fine-grained visual categorization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1173–1182, 2016. 2, 7

[38] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in Neural Information Processing Systems*, 28, 2015. 7

[39] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021. 2

[40] Sarthak Jain and Byron C Wallace. Attention is not explanation. *arXiv preprint arXiv:1902.10186*, 2019. 1

[41] Baoxiong Jia, Yu Liu, and Siyuan Huang. Improving object-centric learning with query optimization. In *The Eleventh International Conference on Learning Representations*, 2022. 3, A

[42] Dmitry Kazhdan, Botty Dimanov, Mateja Jamnik, Pietro Liò, and Adrian Weller. Now you see me (CME): concept-based model extraction. *arXiv preprint arXiv:2010.13233*, 2020. 1, 2

[43] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *International Conference on Machine Learning*, pages 2668–2677. PMLR, 2018. 1, 2

[44] Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (un) reliability of saliency methods. *Explainable AI: Interpreting, explaining and visualizing deep learning*, pages 267–280, 2019. 1

[45] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International Conference on Machine Learning*, pages 5338–5348. PMLR, 2020. 1, 2

[46] Jonathan Krause, Hailin Jin, Jianchao Yang, and Li Fei-Fei. Fine-grained recognition without part annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5546–5555, 2015. 7

[47] Alex Krizhevsky. Learning multiple layers of features from tiny images. Master's thesis, University of Toronto, Toronto, Canada, 2009. 5

[48] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. Faithful and customizable explanations of black box models. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 131–138, 2019. 4

[49] Liangzhi Li, Bowen Wang, Manisha Verma, Yuta Nakashima, Ryo Kawasaki, and Hajime Nagahara. Scouter: Slot attention-based classifier for explainable image recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1046–1055, 2021. 3

[50] Oscar Li, Hao Liu, Chaofan Chen, and Cynthia Rudin. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 1, 2

[51] Zenan Li, Zehua Liu, Yuan Yao, Jingwei Xu, Taolue Chen, Xiaoxing Ma, L Jian, et al. Learning with logical constraints but without shortcut satisfaction. In *The Eleventh International Conference on Learning Representations*, 2023. B

[52] Di Lin, Xiaoyong Shen, Cewu Lu, and Jiaya Jia. Deep lac: Deep localization, alignment and classification for fine-grained recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1666–1674, 2015. 7

[53] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear CNN models for fine-grained visual recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1449–1457, 2015. 7

[54] Xiao Liu, Tian Xia, Jiang Wang, Yi Yang, Feng Zhou, and Yuanqing Lin. Fully convolutional attention networks for fine-grained recognition. *arXiv preprint arXiv:1603.06765*, 2016. 7

[55] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2, 6, B, C

[56] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 2, 6, B, C

[57] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *Advances in Neural Information Processing Systems*, 33:11525–11538, 2020. 2, 3, A

[58] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 2017. 1, 2

[59] Marvin Minsky. *Society of mind*. Simon and Schuster, 1988. 2

[60] Tsendsuren Munkhdalai, Alessandro Sordoni, Tong Wang, and Adam Trischler. Metalearned neural memory. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[61] Meike Nauta, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. Pip-net: Patch-based intuitive prototypes for interpretable image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2744–2753, June 2023. 5

[62] Anh Nguyen, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox, and Jeff Clune. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. *Advances in Neural Information Processing Systems*, 29, 2016. 2

[63] Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, et al. Hopfield networks is all you need. *arXiv preprint arXiv:2008.02217*, 2020. 2

[64] Marco Tulio Ribeiro, Sameer Singh, and Carlos 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*, pages 1135–1144, 2016. 1, 2

[65] Mattia Rigotti, Christoph Miksovic, Ioana Giurgiu, Thomas Gschwind, and Paolo Scotton. Attention-based interpretability with concept transformers. In *International Conference on Learning Representations*, 2021. 1, 2, 4, 5, 6, 7, 8, A, B

[66] Philip Robbins. Modularity of mind. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2017 edition, 2017. 1, 2

[67] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019. 1

[68] Dawid Rymarczyk, Lukasz Struski, Michal Górszczak, Koryna Lewandowska, Jacek Tabor, and Bartosz Zielinski. Interpretable image classification with differentiable prototypes assignment. *CoRR*, abs/2112.02902, 2021. 5, 8

[69] Adam Santoro, Ryan Faulkner, David Raposo, Jack Rae, Mike Chrzanowski, Theophane Weber, Daan Wierstra, Oriol Vinyals, Razvan Pascanu, and Timothy Lillicrap. Relational recurrent neural networks. *Advances in neural information processing systems*, 31, 2018. 2

[70] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017. 1, 2

[71] Sofia Serrano and Noah A Smith. Is attention interpretable? *arXiv preprint arXiv:1906.03731*, 2019. 1

[72] Murray Shanahan. A cognitive architecture that combines internal simulation with a global workspace. *Consciousness and cognition*, 15(2):433–449, 2006. 2

[73] Murray Shanahan. *Embodiment and the inner life: Cognition and Consciousness in the Space of Possible Minds*. Oxford University Press, 2010. 2

[74] Murray Shanahan. The brain's connective core and its role in animal cognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1603):2704–2714, 2012. 2

[75] Murray Shanahan and Bernard Baars. Applying global workspace theory to the frame problem. *Cognition*, 98(2):157–176, 2005. 2

[76] Marcel Simon and Erik Rodner. Neural activation constellations: Unsupervised part model discovery with convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1143–1151, 2015. 7

[77] Gautam Singh, Yeongbin Kim, and Sungjin Ahn. Neural systematic binder. In *The Eleventh International Conference on Learning Representations*, 2023. 8

[78] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017. 1, 2

[79] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5):828–841, 2019. 1

[80] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328. PMLR, 2017. 2

[81] Aäron Van Den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *International Conference on Machine Learning*, pages 1747–1756. PMLR, 2016. 2

[82] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017. 4

[83] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. Caltech–UCSD Birds-200-2011 (CUB-200-2011) dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 2, 5, 6

[84] Bowen Wang, Liangzhi Li, Yuta Nakashima, and Hajime Nagahara. Learning bottleneck concepts in image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10962–10971, 2023. 3, 5, 7, 8, B, C, D, F, G, H, I

[85] Dequan Wang, Zhiqiang Shen, Jie Shao, Wei Zhang, Xiangyang Xue, and Zheng Zhang. Multiple granularity descriptors for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2399–2406, 2015. 7

[86] Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. *arXiv preprint arXiv:1908.04626*, 2019. 1

[87] Chao-Yuan Wu, Yanghao Li, Karttikeya Mangalam, Haoqi Fan, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Memvit: Memory-augmented multiscale vision transformer

for efficient long-term video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13587–13597, 2022. 2

[88] Tianjun Xiao, Yichong Xu, Kuiyuan Yang, Jiaxing Zhang, Yuxin Peng, and Zheng Zhang. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 842–850, 2015. 7

[89] Mengqi Xue, Qihan Huang, Haofei Zhang, Lechao Cheng, Jie Song, Minghui Wu, and Mingli Song. ProtoPFormer: Concentrating on prototypical parts in vision transformers for interpretable image recognition. *arXiv preprint arXiv:2208.10431*, 2022. 2, 5, 7, 8

[90] Zhun Yang, Adam Ishay, and Joohyung Lee. Learning to solve constraint satisfaction problems with recurrent transformer. In *Proceedings of the Eleventh International Conference on Learning Representations*, 2023. K

[91] Chih-Kuan Yeh, Been Kim, Sercan Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. On completeness-aware concept-based explanations in deep neural networks. *Advances in Neural Information Processing Systems*, 33:20554–20565, 2020. 1, 2

[92] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*, 2015. 2

[93] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016. 6

[94] Mateo Espinosa Zarlenga, Pietro Barbiero, Gabriele Ciravegna, Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, Frederic Precioso, Stefano Melacci, Adrian Weller, Pietro Lio, et al. Concept embedding models. In *NeurIPS 2022-36th Conference on Neural Information Processing Systems*, 2022. 1, 2, 7

[95] Han Zhang, Tao Xu, Mohamed Elhoseiny, Xiaolei Huang, Shaoting Zhang, Ahmed Elgammal, and Dimitris Metaxas. Spda-cnn: Unifying semantic part detection and abstraction for fine-grained recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1143–1152, 2016. 7

[96] Ning Zhang, Jeff Donahue, Ross Girshick, and Trevor Darrell. Part-based R-CNNs for fine-grained category detection. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 834–849. Springer, 2014. 2, 7

[97] Heliang Zheng, Jianlong Fu, Tao Mei, and Jiebo Luo. Learning multi-attention convolutional neural network for fine-grained image recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5209–5217, 2017. 2, 7

[98] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016. 2, 7

[99] Bolei Zhou, Yiyou Sun, David Bau, and Antonio Torralba. Interpretable basis decomposition for visual explanation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 119–134, 2018. 2