TOWARDS A UNIFIED NEURAL ARCHITECTURE FOR VISUAL RECOGNITION AND REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Recognition and reasoning are two pillars of visual understanding. However, these tasks have an imbalance in focus; whereas recent advances in neural networks have shown strong empirical performance in visual recognition, there has been comparably much less success in solving visual reasoning. Intuitively, unifying these two tasks under a singular framework is desirable, as they are mutually dependent and beneficial. Motivated by the recent success of multi-task transformers for visual recognition and language understanding, we propose a unified neural architecture for visual recognition and reasoning tasks with a generic interface (e.g., tokens) for all tasks. Our framework enables the principled investigation of how different visual recognition tasks, datasets, and inductive biases can help enable spatiotemporal reasoning capabilities. Noticeably, we find that object detection, which requires spatial localization of individual objects, is the most beneficial recognition task for reasoning. We further demonstrate via probing that implicit object-centric representations emerge automatically inside our framework. We also discover that visual reasoning and object detection respond to drastically different model components; certain architectural choices such as the backbone model of the visual encoder have a significant impact on visual reasoning, but little on object detection. Given the results of our experiments, we believe that a fruitful direction forward is to consider visual reasoning a first-class citizen alongside visual recognition, as they are strongly correlated but benefit from potentially different design choices.

1 INTRODUCTION

Modern advances in convolutional neural networks (LeCun et al., 1995) have demonstrated great proficiency in visual recognition tasks, such as image classification (Krizhevsky et al., 2012), object detection (Ren et al., 2015), and instance segmentation (He et al., 2017). More recently, the Transformer architecture (Vaswani et al., 2017), initially designed for language understanding tasks such as machine translation, has demonstrated competitive performance on image (Dosovitskiy et al., 2020) and video (Arnab et al., 2021) recognition tasks. Although the performance gains of the Transformer over alternative architectures for visual perception is still under debate (Tolstikhin et al., 2021; Liu et al., 2022), one concrete benefit of the Transformer is its versatility in modeling diverse input modalities (Sun et al., 2019; Jaegle et al., 2021), and its flexibility in unifying a wide range of perception and reasoning tasks as sequence prediction (Raffel et al., 2019; Chen et al., 2021). Indeed, scaling Transformers massively in size, compute, and data has enabled many exciting recent developments such as multi-modal "foundation models" (Bommasani et al., 2021) and multi-task "generalist agents" (Reed et al., 2022). Motivated by their success, our paper investigates how this paradigm, where a neural network is employed agnostic of task-specific inductive biases, can be used to build a unified model that can perform both visual recognition and visual reasoning tasks.

Visual perception and reasoning are the two pillars of visual understanding. Despite the rapid progress on visual question answering (Antol et al., 2015; Hudson & Manning, 2019b), primarily due to Transformer networks and large-scale pre-training (Alayrac et al., 2022), it has been observed that state-of-the-art neural networks designed for visual recognition have much less success in solving even the most basic reasoning tasks that require causal inference (Yi et al., 2019; Zhang et al., 2021) or the notion of object permanence (Girdhar & Ramanan, 2019). Our study focuses on these visual reasoning tasks, which require understanding object attributes, relationships, and dynamics.

Specifically, we use two diagnostic benchmarks: CATER (Girdhar & Ramanan, 2019), where a model must learn object permanence to track occluded objects in videos, and ACRE (Zhang et al., 2021), which requires a model to perform causal inference in a setup inspired by the famous Blicket experiment (Gopnik & Sobel, 2000).

Researchers (Greff et al., 2020; Santoro et al., 2021) have advocated that an object-centric, symbol-like representation is essential for compositional generalization required by reasoning tasks. It was also commonly believed that although neural networks may excel at extracting objects and their attributes given a pre-defined vocabulary, the reasoning module should be based on symbolic approaches (Mao et al., 2019). However, the seminal work ALOE by Ding et al. (2021) demonstrated that a Transformer neural network can not only perform reasoning, but also even significantly outperform its neuro-symbolic counterparts at times. Similar to neuro-symbolic approaches, ALOE still requires pre-computed object-centric representations (Burgess et al., 2019), where visual recognition is a separate component and arguably requires task-specific knowledge to design.

In this work, we propose a unified framework for recognition and reasoning, where visual representations can be extracted, organized, and routed dynamically to solve both sets of tasks in parallel. Under this framework, we are particularly interested in understanding whether visual representations are organized in an *object*-centric fashion in the intermediate layers, which conceptually would facilitate compositional generalization. We follow the notions by Greff et al. (2020) and refer to symbol-like entities as objects. In the context of our work, they can correspond to objects, object parts, or spatiotemporal object trajectories. We hypothesize that the choice of recognition task can help promote the emergence of such implicit object-centric representations; we further conjecture that tasks involving spatiotemporal understanding such as object detection are particularly helpful.

Under our proposed unified framework, we are able to achieve competitive performance on CATER for object tracking with occlusion, and reasonable performance on ACRE for causal inference. Furthermore, our framework enables the principled comparison between the effects of different design decisions on reasoning capabilities. When ablating over different visual recognition tasks, we find that the object detection task is critical for better reasoning capability of the model; we also qualitatively visualize via probing that object-centric representations emerge in the middle of the network. We believe these results show encouraging signals towards building a unified and generic multi-task framework for recognition and reasoning. Our ablation study also reveals a surprising observation that while the object detection performance is robust across different choices of neural architectures, such as the use of ViT (Dosovitskiy et al., 2020) or a ResNet (He et al., 2016) visual encoder, different inductive biases can lead to significant performance gaps on the model's reasoning performance. As most of the existing visual multi-task models (Alayrac et al., 2022) or benchmarks (Zhai et al., 2019) focus on visual recognition, we hope our findings can bring awareness to treating visual reasoning as a first-class citizen when designing neural architectures and benchmarking model performances.

Overall, our main contributions are as follows:

- We propose a unified end-to-end architecture that performs reasoning alongside object detection, eliminating the need for pipeline-style approaches.
- We utilize our architecture to investigate different considerations, such as inductive biases or visual recognition objectives, when designing solutions for spatiotemporal reasoning.
- We observe that the object detection task leads to the emergence of object-centric representation, which we hypothesize improves reasoning. We visualize and investigate these representations through probing.

2 Related Work

Visual reasoning. The huge success of deep learning for image understanding tasks have prompted researchers to tackle visual reasoning tasks, which are more challenging, in the form of visual question answering (Antol et al., 2015; Lu et al., 2016; Hudson & Manning, 2019b), visual commonsense extraction (Yatskar et al., 2016) and reasoning (Zellers et al., 2019), or visual dialog (Das et al., 2017). While neural networks have achieved remarkable success in these benchmarks that require reasoning, it has also been observed that dataset bias and language bias (Goyal et al., 2017) make it harder to concretely measure progress. This has lead to a suite of synthetic, diagnostic

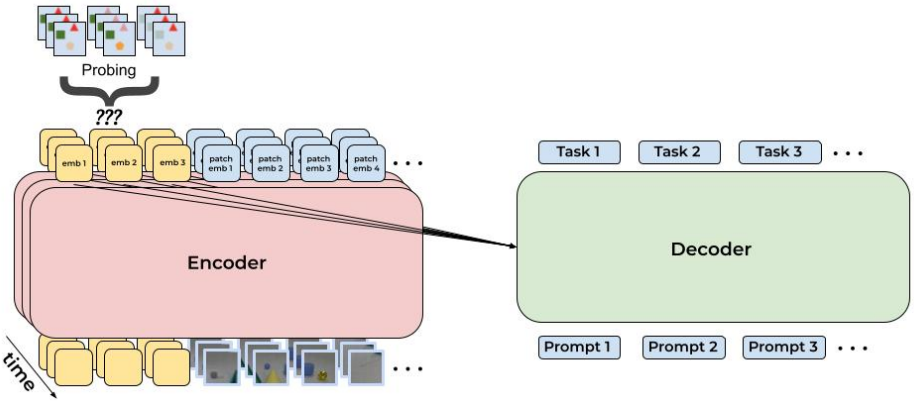

Figure 1: An illustration of our proposed unified neural architecture for visual recognition and reasoning. As in Pix2Seq (Chen et al., 2021), each image or its ResNet feature map is first broken into patches and fed to a Transformer encoder. We employ a few "slot" tokens (yellow) which condense the visual information and pass them to the decoder. The slot tokens from different frames of a video are concatenated over time. The Transformer decoder then autoregressively predicts task-specific output sequences according to their corresponding "prompts", such as object labels, (quantized) object locations, or answers to a reasoning task. To understand if the slot tokens encode object-centric representations we train a probing classifier, also modeled as an autoregressive decoder under our framework, and ask it to predict object locations given a randomly selected slot token. An object-centric token is expected to successfully detect only one or few objects.

datasets (Johnson et al., 2017; Yi et al., 2019; Girdhar & Ramanan, 2019; Zhang et al., 2021) to benchmark visual reasoning. These benchmarks highlighted the limitations of end-to-end trained neural networks, and often advocated for neural-symbolic methods (Mao et al., 2019; Hudson & Manning, 2019a). However, such neuro-symbolic approaches often require task-specific knowledge to define the "symbols" and the "programs", making them more challenging to generalize across tasks. Recently, Ding et al. (2021) demonstrated that an end-to-end trained Transformer network can indeed perform reasoning, given that its inputs are "symbol-like" object segments. Our paper aims to take one step further and investigate if it is possible to build a unified framework for visual recognition and reasoning without committing to a task-specific inductive bias at the model's inputs, which would make the framework more general.

Object-centric representation is an essential component for both neuro-symbolic (Mao et al., 2019) and Transformer-based (Ding et al., 2021) frameworks for visual reasoning. Greff et al. (2020) argue that a reasoning framework consists of three stages, namely *segregation*, *representation*, and *composition*. The *segregation* stage is responsible for extracting objects (or other symbol-like entities), and can be implemented with a supervised object detector (Mao et al., 2019), such as Mask R-CNN (He et al., 2017). There is also a line of research to learn low-level features correlated with objects, such as textures (Geirhos et al., 2018; Hermann et al., 2020; Olah et al., 2017), or even object detectors (Burgess et al., 2019; Locatello et al., 2020; Caron et al., 2021) with self-supervised objectives. Researchers working on model interpretability also attempt to visualize the internal activations of a trained neural network, and inspect if they correspond to objects or object parts (Bau et al., 2017; 2018). Our work takes a supervised framework (Chen et al., 2021) for its general-purpose transformer-based architecture, which can be easily generalized to reasoning tasks.

Multi-task Transformers have achieved tremendous success recently (Alayrac et al., 2022; Bommasani et al., 2021; Reed et al., 2022) on tasks that require visual perception and control. Particularly appealing is the Transformer's versatility in incorporating multi-modal inputs with minimal modality-specific assumptions (Akbari et al., 2021; Sun et al., 2019; Jaegle et al., 2021), and its flexibility to express the training objectives of diverse tasks all as sequence prediction (Raffel et al., 2019; Chen et al., 2021). Large-scale pre-training, both for the number of model parameters, and for the amount of (unlabeled) training data, is also found to be crucial (Bommasani et al., 2021). Our generalized framework, which unifies the object detection and visual reasoning tasks, belongs to the model family of multi-task Transformers.

3 METHOD

To create a unified model, we propose building off of the Pix2Seq framework (Chen et al., 2021). In its default formulation, Pix2Seq processes a single image as a sequence-to-sequence task; visual patches are fed in as inputs to an encoder, and bounding box predictions are generated as an output sequence via an autoregressive decoder. The encoder can be a Vision Transformer, and input patch embeddings can be implemented as linear projections (Dosovitskiy et al., 2020), or as the intermediate features of a ResNet (He et al., 2016). The autoregressive decoder conditions on the output embeddings of the encoder. Furthermore, it takes in a prompt, which can be used to help the model specify and differentiate between tasks, or to include task-relevant information to be conditioned on. This framework is general and flexible; there are no assumptions about the visual environment Pix2Seq can be applied to, nor does it have any requirements on preprocessing its inputs, and the output sequence is not limited to object detection as a task. Indeed, the outputs of other tasks can be structured as sequences as well, such as captioning, or the answers to reasoning questions.

Architecture. We therefore extend Pix2Seq to solve spatio-temporal reasoning problems using three main modifications. Firstly, we utilize a cross-attention bottleneck to force the representation of the input to be encoded in the form of a few rich tokens. Denoted by the yellow tokens in Figure 1, only these tokens are passed forward to the decoder for conditioning. We interpret these bottleneck tokens as slots, and hypothesize that they bind to object-centric information. We later make this explicit via our probing experiments. Secondly, we adapt Pix2Seq to the video domain by simply "stacking" the encoder structure; we reuse the same weights, but handle each frame independently. The output bottleneck tokens of each frame are then concatenated together, preserving temporal order, and positional encodings are added before they are passed to the decoder for conditioning.

Multi-task training. Lastly, we extend Pix2Seq to handle multiple tasks simultaneously. There are two approaches that can accomplish this: joint training, and alternating optimization. We first detail how to perform joint training; recall that the decoder portion of the Pix2Seq framework conditions on the output of the encoder, as well as a task-specific prompt. By feeding in multiple prompts over one step, a Pix2Seq implementation can jointly output the predictions for multiple tasks; the loss can then be computed to optimize the model with respect to several objectives at once. In the alternating optimization approach, a Pix2Seq model iteratively switches between optimizing for different tasks at fixed interval steps. Intuitively, since an ordering to tasks might potentially be useful, this may result in better performing unified models. For example, if we expect that the representations learned by the perceptual tasks might be subsequently useful for visual reasoning, we might prefer to first train on the perceptual task for a number of iterations before alternating. Switching back from optimizing for the visual reasoning task to the perceptual task can also be useful; it allows for the implicit representations to be supervised with reasoning-specific signals. In the extreme case, where we only perform one switch between tasks, this is equivalent to pre-training our extended Pix2Seq model on a perceptual task and finetuning it on a reasoning task. We explore different multi-task training strategies in our experiments and find that for us, a single switch from object detection to reasoning yields the best performance on the reasoning tasks.

It is noteworthy that this approach is general, and fully supports more tasks than two at a time, as illustrated by Figure 1. Different combinations of tasks, trained jointly or iteratively, may induce the emergence strong implicit representations - we leave such experimentation as interesting future work. Indeed, the properties of such representations are worth investigating, as they are optimized to be useful for a variety of applications, and can potentially be leveraged to even solve previously unseen tasks. In this work, we perform an initial analysis to understand the representations output by the encoder via probing techniques described below.

Representation Probing. We are interested in investigating the encoder's bottleneck embeddings to determine if they exhibit some object-centric or localized properties; as the decoder directly references them to solve tasks, their representational quality directly impacts performance. One way to evaluate the information contained within the embeddings is through probing; after designating a set of embeddings of interest (or combinations thereof, in the general case), we can probe if they contain object-centric information such as bounding box coordinates or class labels. The object detection training procedure can be reused under our framework with two changes: we randomly sample embeddings from the investigation set at each training iteration to attempt the object detection task,

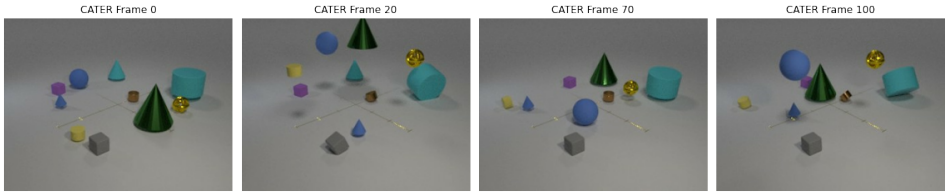

Figure 2: Example frames from a CATER video. The task is to determine the final location of the golden snitch, which may be occluded or covered throughout the video, as one of 36 locations.

and we freeze the encoder so that the representations are purely evaluated and are not modified or supervised further. After our decoder is trained to perform probing in such a fashion, we can apply it during inference to evaluate and visualize every embedding of interest. Intuitively, if the encoder chooses to organize visual information in an object-centric fashion, it is expected that every probed embedding can successfully detect only one or few objects in an image, as information about other objects and their corresponding attributes are *routed* to other embeddings.

We note that since our framework also models probing as sequence prediction, it can be easily extended to incorporate arbitrary available ground-truth probing information available in a given domain, such as color and material, through constructing an appropriate target prediction sequence. Overall, our unified framework naturally enables flexible probing functionality, thus enabling users to interpret and analyze the internal representations of the model with minimal adjustment.

4 EXPERIMENTS

We showcase our model on two spatiotemporal reasoning tasks, CATER (Girdhar & Ramanan, 2019) and ACRE (Zhang et al., 2021), and perform ablations over them. The encoder and decoder of the model is also initialized from a pre-trained ImageNet classification checkpoint, on top of which all visual and reasoning tasks are further optimized on. All reported experiments were performed on 32 Dragonfish TPUs per run.

4.1 CATER

We first evaluate our model on the CATER task. CATER is composed of videos of CLEVR (Johnson et al., 2017) objects moving and, importantly, occluding each other. Conceptually, it is meant to simulate the shell game, where the objective is to locate the final position of a ball that is hidden under a cup and shuffled with other cups. In the CATER task, larger objects can occlude smaller objects visually, with respect to the camera viewpoint. Furthermore, objects can also completely cover up smaller objects and move around with them, effectively moving multiple objects at once while occluding them entirely from view. As in the shell game, the CATER task is to locate the final position of a unique golden object, named the "snitch", where locations are represented as units on a 6x6 grid. Therefore, the problem can be interpreted as a 36-way classification problem. Whereas the CATER dataset features an additional split where the camera viewpoint can also move freely over the course of the video, we focus on the static camera setting in our experiments. A visual demonstration of a sample CATER task is provided in Figure 2.

By default, each CATER video contains 301 frames. We follow the experimental setup in ALOE, where we resize each image to 64x64 resolution, and sample 80 frames uniformly at random from a full video for each training example, preserving temporal order. We also perform evaluation on 80 frame subsequences, but space their indices as evenly as possible. We utilize a batch size of 256, and find convergence in 7,000 steps. We find that using 3500 warmup steps, a learning rate of $3e^{-4}$ with a linear scheduler, and the Adam optimizer (Kingma & Ba, 2014) with weight decay factor 0.05, generally produces the best results. For other hyperparameter settings, we either match the configurations used by ALOE, or selected them from validation set performance. When adapting our framework on CATER, we concatenate only the first token from the encoder output of each frame and expose it to the decoder; we also provide ablations over the number of tokens to use per frame in Table 1.

Table 1: Ablations on the CATER benchmark.

Encoder Backbone	Recognition Task	Tokens per frame	CATER Top 1 (Static)
ResNet + Transformer	None (Ground-Up)	1 (slot)	18.10%
Vision Transformer	None (Ground-Up)	1 (slot)	4.36%
ResNet + Transformer	LA-CATER ObjDet	1 (slot)	**74.09%**
Vision Transformer	LA-CATER ObjDet	1 (slot)	56.64%
ResNet + Transformer	LA-CATER ObjDet	10 (slots)	71.68%
ResNet + Transformer	MS-COCO ObjDet	1 (slot)	69.79%

As a default baseline, we first attempt to train our proposed Pix2Seq video model to solve the CATER task alone, without the addition of any other tasks. In Table 1, we discover that it struggles to perform reasoning; our model achieves 18.10% test accuracy when using a ResNet Transformer encoder backbone, and a mere 4.36% when using a pure Vision Transformer encoding backbone. For reference, a random guess on the location would be accurate 2.78% of the time.

We then evaluate our multi-task model formulation, by training on the object-detection task before learning how to perform reasoning. Firstly, however, to even train on object detection we require ground truth bounding box supervision. As these labels are not present in the default CATER videos, we use the annotations from the LA-CATER dataset (Shamsian et al., 2020). The LA-CATER dataset is generated to be identical to the CATER domain; it features the same camera configuration, the same usage of CLEVR objects, and the same number of possible objects per scene. We discover that our end-to-end model is able to perform the CATER task with 74.09% test accuracy using a ResNet Transformer encoder and 56.64% using a Vision Transformer backbone. Our end-to-end multitask model is therefore able to outperform the default ALOE model, which achieves 70.6% test accuracy. Its performance is also comparable with the ALOE model that uses a task-specific L1 loss, which achieves $74.0 \pm 0.3\%$ (See Appendix) . In contrast, we do not use any extra auxiliary task-specific losses in our setup. We therefore conclude that the addition of the general object-detection task helps the model perform spatio-temporal reasoning in an end-to-end unified model. Given ALOE's finding that object-centric abstractions of the input are important for reasoning, which they leverage a pre-trained MONET model to generate, we take this as a preliminary sign that the model forms strong object-centric representations implicitly.

We also explore whether object-detection as an objective is generally useful, or if the benefits are only limited to the visual domain of the reasoning task. In other words, we investigate if there are inherent properties of object detection that could potentially enable the learning of object-centric abstractions that can extend beyond visual domains, and still be useful in enabling reasoning capabilities. We test this in our model by evaluating CATER performance with object-detection pre-training on the MS-COCO dataset. Composed of natural images, the MS-COCO dataset is visually dissimilar from the synthetic CATER environment, and the two datasets share no overlapping object classes. However, we find that our method still achieves 69.79% on test performance using a ResNet Transformer encoder. Therefore, even though it is trained for object detection on a completely different visual domain, we see that our method is still able to exhibit reasoning capabilities, even beating the default ALOE performance. Whereas ALOE requires pretraining a separate model on the same visual demonstrations from the same domain in a pipeline-style approach, we demonstrate that our model is still able to perform reasoning in an end-to-end way, leveraging object-detection capabilities on out-of-domain data. This is a surprising and powerful result that further diminishes the need for task-specific considerations or pipeline-style approaches, and hints at the efficacy of object detection as a general task.

Furthermore, we ablate over the number of tokens provided by the encoder to the decoder, and find that using only one token per frame works the best. This is potentially due to the singular token containing enough information to solve the task of locating the snitch. To provide intuition on this, we visualize if the token contains an accurate belief over the location of the snitch at arbitrary frames in our probing experiments below.

Lastly, we note that in all of our experiments, the ResNet Transformer encoder backbone is much more performant than the Vision Transformer encoder backbone. The performance disparity when

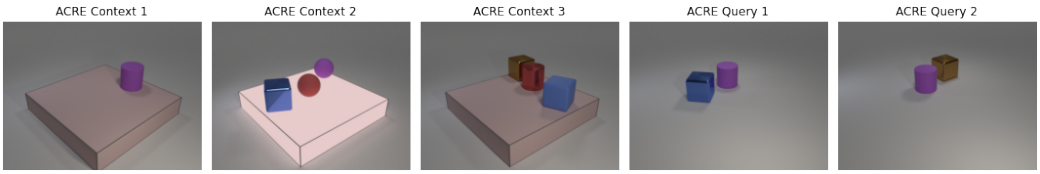

Figure 3: An example of the ACRE task. The first three frames represent context panels, which expose the state of the underlying platform to the model. The model must predict the platform state for the two query configurations in Frames 4 and 5. The result of Frame 4 is undetermined; although we know the purple cylinder does not light up the platform, we do not have enough information regarding the blue metal cube, as the signal in Frame 2 may or may not be the result of the other objects present. That Frame 5 would not light up the platform can be deduced from the first and third context frame. For ACRE, six context frames are provided, with four associated queries.

Table 2: Comparisons across visual perception objectives and datasets.

Dataset	Visual Recognition Task	ACRE (Comp)	CATER Top 1 (Static)
LA-CATER	Object Detection	67.27%	**74.09%**
LA-CATER	Visible Object Detection	81.65%	73.44%
LA-CATER	Count All Objects	44.04%	68.36%
LA-CATER	Count Unique Objects	41.48%	64.78%
LA-CATER	Snitch Detection	79.70%	72.66%
MS-COCO	Object Detection	**83.81%**	69.79%

applied to a reasoning task is a surprising result, since both architectures achieved similarly high AP50 performance on LA-CATER (81.53 for the ResNet Transformer, and 82.44 for the Vision Transformer). We therefore believe that convolutional inductive bias is still important. In particular, as ALOE has demonstrated the importance of object-centric abstractions of input, perhaps the convolutional structure is more adept at isolating and encoding object-centric information than the linear operation employed by the vanilla Vision Transformer encoding scheme. We believe this is a signal that architectural considerations will be important when designing solutions to reasoning problems, and that simply reusing settings that work on pure visual recognition tasks can potentially inhibit performance.

4.2 ACRE

We also evaluate our model on the ACRE task, which tests causal inference capabilities. ACRE is based off the Blicket experiment from developmental psychology (Gopnik & Sobel, 2000). In the original formulation, a machine lights up when certain objects, called "Blickets" are placed on it. Preschool-aged children are tasked with determining which objects are Blickets, given demonstrations of certain object configurations and their resulting effects on the machine. In the ACRE task, CLEVR objects are placed on a fixed platform, which glows or remains dim depending on the "Blicketness" properties of the objects. The model is provided with six context frames demonstrating different combinations of objects and their corresponding platform state, as well as a query frame containing an object combination that the model is expected to predict the result of.

There are four types of reasoning tasks the model is tested on, categorized by question types: direct, indirect, screened-off, and backward-blocking. The potential answers for each of these questions is either that the platform's state is on, off, or unable to be determined. In a direct question, the query combination is previously observed during one of the provided context trials; the model must be able to recognize it from the context and reference it to retrieve the answer. In an indirect question, the query combination is novel, and the result must be deduced from several frames. In screen-off questions, the model must learn the dynamic that as long as one Blicket object is present, the entire combination would turn the machine on. Lastly, in the backward-blocking questions, the model must correctly deduce that the query combination cannot be ascertained from the available contexts.

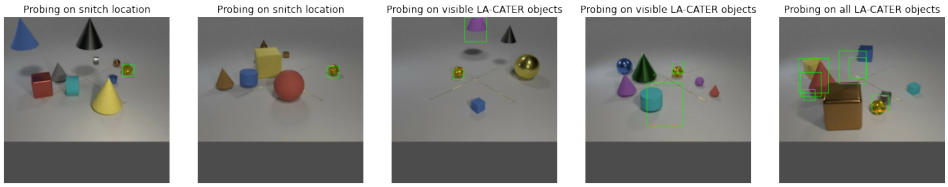

Figure 4: Visualizing the predicted location of the snitch from a frozen, learned CATER embedding.

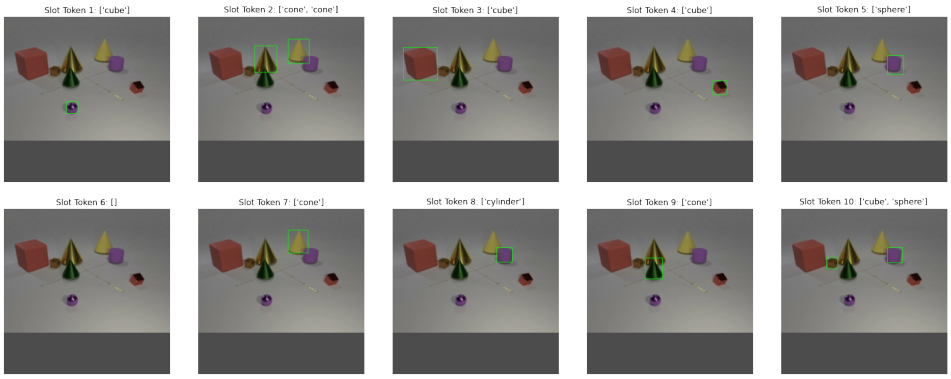

Figure 5: We visualize the true positive bounding box and shape predictions for each of the ten bottleneck embeddings in a frozen ResNet + Transformer encoder using probing. Note that each token seems to be adept at encoding one or a few objects well.

As in the CATER experiments, we utilize a model trained on the object detection task using the LA-CATER dataset. We reuse many of the same hyperparameter configurations from the CATER setup, except with 3000 warmup steps and 50,000 total training steps.

We first investigate how different types of visual recognition tasks can enable reasoning capabilities in Table 2. We therefore ablate over a variety of different tasks. In LA-CATER Object Detection, the visual recognition is tasked with predicting the bounding boxes and shapes of all objects in the scene, hidden or visible; for MS-COCO the model predicts the bounding boxes and class labels. In Visible Object Detection, the model is only asked to predict the bounding boxes and shape labels for all visible objects from the camera perspective. In Count All Objects and Count Unique Objects, the model is tasked with predicting the number of total objects in the scene, and the number of unique shapes in the scene, respectively; notably, these objectives do not require the model to learn strong spatial localization. Lastly, the Snitch Detection objective encourages the model to predict the bounding box location of the golden snitch. We find that the best performance on ACRE is achieved through visual objectives that involve spatial localization. Indeed, the top two performances on both ACRE and CATER utilize the Object Detection task: on MS-COCO and visible objects from LA-CATER for ACRE, and LA-CATER and visible LA-CATER for CATER. Learning to predict hidden or occluded objects, as in the complete LA-CATER Object Detection, may not be helpful for the ACRE task since there are no fully occluded objects in the environment. We also notice that the visual recognition tasks that do not require explicit spatial localization, such as Count All Objects or Count Unique Objects, perform worst on their respective reasoning environments. Furthermore, we find that the snitch detection performs well; on CATER, this is to be expected since it is directly useful for the final reasoning objective. Overall, these results support the hypothesis that perceptual tasks that learn spatial localization of objects are beneficial for unlocking reasoning capabilities.

4.3 PROBING

We perform probing under the setup described in the Methods section, to understand the information encoded in the learned embeddings under our unified architecture. We begin by investigating the singular bottleneck embedding learned by a CATER model that utilized an LA-CATER Object Detection task with a ResNet + Transformer backbone. We load and freeze the encoder of the

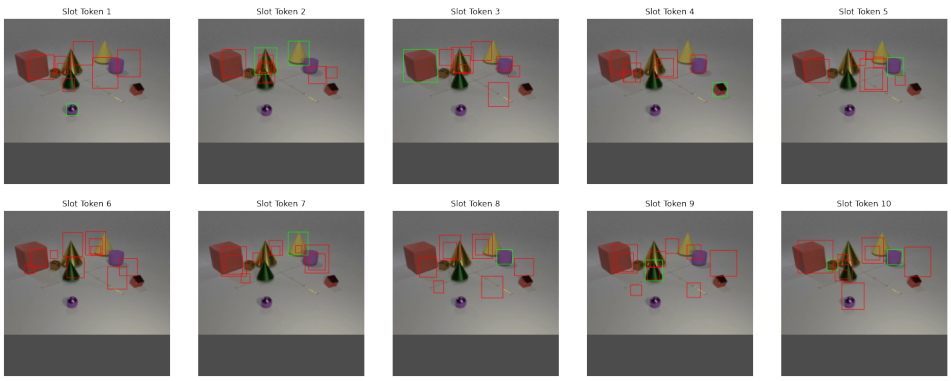

Figure 6: We visualize all bounding box predictions for each token in a frozen ResNet + Transformer encoder using probing. We highlight the true positive bounding box predictions in green.

learned CATER model, and proceed to train an autoregressive decoder on the snitch detection task, LA-CATER Visible Object Detection, and LA-CATER Object Detection. We then visualize our predictions in Figure 4 for randomly selected frames, and demonstrate qualitatively that the learned embedding consistently encodes an accurate prediction of where the snitch is at any given time, regardless of the probing objective.

We also perform probing experiments on a ResNet Transformer encoder that was previously trained on LA-CATER for the object detection task using ten bottleneck embeddings to understand how the reasoning models for CATER and ACRE are initialized. As we utilize this object detection pretraining procedure before each of the two reasoning tasks we evaluate on, the embedding outputs of this encoder are especially interesting to investigate in order to understand how they can potentially enable reasoning capabilities. We freeze the encoder, and proceed to learn an autoregressive decoder on the LA-CATER object detection task, but selecting one token at random each iteration. After convergence, we use the trained decoder to evaluate each individual bottleneck token of the encoder by qualitatively visualizing their predicted bounding boxes on a random LA-CATER test set frame.

In Figure 6 we display all the predicted bounding boxes with a confidence score above 90% for each token. We notice three things - firstly, no token is encoding every object, which is an encouraging result. In fact, there exists a token, token number 6, that does not appear to encode any object strongly. This is not of particular issue, as there are ten total "slot" tokens learned, but only eight objects in the scene. Secondly, we discover that with the exception of number 6, the tokens are able to predict one or a few objects exceptionally accurately. We visualize these clean predictions in Figure 5, and display their corresponding predicted shapes as well. These results demonstrate that individual tokens are able to not only understanding where certain objects are but their properties as well. Lastly, we observe that despite each token representing one or a few objects well, all of the objects in the scene are accounted for and there are few overlaps. Therefore we discover that in combination, the tokens learned under our framework are able to collectively represent an entire scene in terms of its objects well, while encoding object-centric information individually.

5 Conclusion and Future Work

We attempt to build a unified framework for visual recognition and reasoning with general-purpose Transformers. We hypothesize that object detection motivates the network to learn object-centric representations which are beneficial for visual reasoning. We generalize Pix2Seq, a transformer-based sequence prediction framework for object detection, to jointly tackle detection and reasoning on two diagnostic datasets, CATER and ACRE. Our quantitative and qualitative results show encouraging signs that object detection indeed helps visual reasoning, and that object-centric representations seem to emerge from object detection pretraining. Interestingly, our experiments also reveal that although different inductive biases may have little impact on object detection performance, their relative gaps on the reasoning benchmarks are nontrivial. We hope that our findings can bring awareness and consideration to reasoning performance when designing network architectures, as well as motivate further explorations on building unified multi-task models for perception and reasoning. In the future, we would like to explore other recognition tasks and joint training strategies.

REFERENCES

Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, 34, 2021.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6836–6846, 2021.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6541–6549, 2017.

David Bau, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou, Joshua B Tenenbaum, William T Freeman, and Antonio Torralba. Gan dissection: Visualizing and understanding generative adversarial networks. *arXiv preprint arXiv:1811.10597*, 2018.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Christopher P Burgess, Loic Matthey, Nicholas Watters, Rishabh Kabra, Irina Higgins, Matt Botvinick, and Alexander Lerchner. Monet: Unsupervised scene decomposition and representation. *arXiv preprint arXiv:1901.11390*, 2019.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021.

Ting Chen, Saurabh Saxena, Lala Li, David J Fleet, and Geoffrey Hinton. Pix2seq: A language modeling framework for object detection. *arXiv preprint arXiv:2109.10852*, 2021.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 326–335, 2017.

David Ding, Felix Hill, Adam Santoro, Malcolm Reynolds, and Matt Botvinick. Attention over learned object embeddings enables complex visual reasoning. *Advances in Neural Information Processing Systems*, 34, 2021.

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.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

Rohit Girdhar and Deva Ramanan. Cater: A diagnostic dataset for compositional actions and temporal reasoning. *arXiv preprint arXiv:1910.04744*, 2019.

Alison Gopnik and David M Sobel. Detecting blickets: How young children use information about novel causal powers in categorization and induction. *Child development*, 71(5):1205–1222, 2000.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.

Klaus Greff, Sjoerd Van Steenkiste, and Jürgen Schmidhuber. On the binding problem in artificial neural networks. *arXiv preprint arXiv:2012.05208*, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

Katherine Hermann, Ting Chen, and Simon Kornblith. The origins and prevalence of texture bias in convolutional neural networks. *Advances in Neural Information Processing Systems*, 33:19000–19015, 2020.

Drew Hudson and Christopher D Manning. Learning by abstraction: The neural state machine. *Advances in Neural Information Processing Systems*, 32, 2019a.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6700–6709, 2019b.

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*, pp. 4651–4664. PMLR, 2021.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2901–2910, 2017.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *arXiv preprint arXiv:2201.03545*, 2022.

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.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. *Advances in neural information processing systems*, 29, 2016.

Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B Tenenbaum, and Jiajun Wu. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. *arXiv preprint arXiv:1904.12584*, 2019.

Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2017. doi: 10.23915/distill.00007. https://distill.pub/2017/feature-visualization.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, et al. A generalist agent. *arXiv preprint arXiv:2205.06175*, 2022.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

Adam Santoro, Andrew Lampinen, Kory Mathewson, Timothy Lillicrap, and David Raposo. Symbolic behaviour in artificial intelligence. *arXiv preprint arXiv:2102.03406*, 2021.

Aviv Shamsian, Ofri Kleinfeld, Amir Globerson, and Gal Chechik. Learning object permanence from video. In *European Conference on Computer Vision*, pp. 35–50. Springer, 2020.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7464–7473, 2019.

Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in Neural Information Processing Systems*, 34, 2021.

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.

Mark Yatskar, Vicente Ordonez, and Ali Farhadi. Stating the obvious: Extracting visual common sense knowledge. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 193–198, 2016.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. Clevrer: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442*, 2019.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6720–6731, 2019.

Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, et al. The visual task adaptation benchmark. 2019.

Chi Zhang, Baoxiong Jia, Mark Edmonds, Song-Chun Zhu, and Yixin Zhu. Acre: Abstract causal reasoning beyond covariation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10643–10653, 2021.