Transfer Learning in Visual and Relational Reasoning

Anonymous CVPR submission

Paper ID 7813

Abstract

Transfer learning has become the de facto standard in computer vision and natural language processing, especially where labeled data is scarce. Accuracy can be significantly improved by using pre-trained models and subsequent fine-tuning. In visual reasoning tasks, such as image question answering, transfer learning is more complex. In addition to transferring the capability to recognize visual features, we also expect to transfer the system's ability to reason. Moreover, for video data, temporal reasoning adds another dimension. In this work, we formalize these unique aspects of transfer learning and propose a theoretical framework for visual reasoning, exemplified by the well-established CLEVR and COG datasets. Furthermore, we introduce a new, end-to-end differentiable recurrent model (SAMNet), which shows state-of-the-art accuracy and better performance in transfer learning on both datasets. The improved performance of SAMNet stems from its capability to decouple the abstract multi-step reasoning from the length of the sequence and its selective attention enabling to store only the question-relevant objects in the external memory.

1. Introduction

In recent years, neural networks, being at the epicenter of the Deep Learning [?] revolution, became the dominant solutions across many domains, from Speech Recognition [?], Image Classification [?], Object Detection [?], to Question Answering [?] and Machine Translation [?] among others. At their core, being statistical models [?, ?], neural networks rely on the assumption that training and testing samples are independent and identically distributed (*iid*); i.e. sampled from a common input space under similar data distribution characteristics. However, in many real-world scenarios, this assumption does not hold. Moreover, as modern neural models often have millions of trainable parameters, training them requires vast amounts of data, which for some domains (e.g., medical) can be very expensive and/or extremely difficult to collect. One of the widely used solutions for the above mentioned problems is Transfer Learning [?, ?], a technique which enhances model performance by transferring *information* from one domain to another.

TK: Two-three sentences about TL in CV and NLP - short!!

TK: The following needs to be rewritten - stronger!

In this work we focus on transfer learning in multimodal tasks combining vision and language [?]. More precisely, we narrow the scope to transfer learning between visual reasoning tasks that have a "nice" logical structure, e.g., [?, ?, ?]. While models such as BERT and ResNet can be transferred efficiently in the same modality they were pretrained on, challenges arise once the modalities have been fused. For example, the CoGenT (Constrained Generalization Test) variant of the CLEVR dataset [?] contains two sets with similar questions, but differing on combinations of object-attribute values in images (??). In this case, training on the first variant might yield entangled feature representations that may fail reasoning tasks on the second one. In video reasoning, an additional challenge in the temporal dimension is whether a model trained on shorter video sequences will transfer over to longer ones, e.g., the Canonical and the Hard variants of the COG dataset [?] (Section 5.4). To address these challenges, mechanisms such as attention [?] and external memory [?, ?, ?] which facilitate higher-level abstractions, seem more promising.

Motivated by these considerations:

1. We propose a new model, called SAMNet (Selective Attention Memory Network), which achieves state-of-the-art results on COG [?], a Video QA reasoning dataset.

2. We propose a taxonomy of transfer learning, inspired from [?], applied to the domain of visual reasoning. Articulated around 3 main axes, we illustrate it through the COG dataset, as well as the CLEVR [?] diagnostic dataset for Image QA.

3. Subsequently, we measure the impact of transferring the whole pretrained SAMNet model in the 3 proposed transfer learning settings: feature transfer, temporal transfer and reasoning transfer. This analysis is supported by an extensive set of experiments using the COG

and CLEVR datasets, as well as their variants. Several of these experiments show significant transfer learning capabilities of SAMNet.

2. Related work

TK: That section will focus **only** on transfer learning

In Computer Vision, it is now standard practice to pretrain an image encoder (such as VGG [?] or ResNet [?]) on large-scale datasets (such as ImageNet [?]), and reuse the weights in unrelated domains and tasks, such as segmentation of cars [?] or Visual Question Answering (VQA) in a medical domain [?]. Such performance improvements are appealing, especially in cases where both the domain (natural vs. medical images) and the task (image classification vs. image segmentation vs VQA) change significantly.

Similar developments have emerged in the Natural Language Processing (NLP) community. Using shallow word embeddings, such as word2vec [?] or GloVe [?], pretrained on large corpuses from e.g. Wikipedia or Twitter, has become a standard procedure when working with different NLP domains and tasks. Recently, there is a clear, growing trend of utilization of deep contextualized word representations such as ELMo [?] (based on bidirectional LSTMs [?]) or BERT [?] (based on the Transformer [?] architecture), where entire deep networks (not just the input layer) are pretrained on very large corporas. In analogy to pretrained image encoders, the NLP community has also started to create model repositories, some with dozens of pretrained models ready to be downloaded and used. HuggingFace [?] is one of the most notable examples.

The success of transfer learning raises several research questions, such as the characteristics which make a dataset more favorable to be used in pretraining (notably ImageNet [?]), or regarding the observed performance correlation of models with different architectures between the source and target domains [?]. One of the most systematic works in this area is the computational taxonomic map for task transfer learning [?], which aimed at discovering the dependencies between twenty-six 2D, 2.5D, 3D, and semantic computer vision tasks.

TK: missing review: TL in visual reasoning, refer to CoGenT and COG

TK: after reading that section the reader should end up with a conclusion that: there are no good models for TL in VR and, moreover, the datasets are randomly testing this or that, there is no theoretical framework showing that do they mean/bigger picture is missing

3. Selective Attention Memory (SAM) Network

TK: Description of the SAMNet

4. Transfer Learning

TK: description of the theoretical framework being the foundation driving our experiments

4.1. Feature transfer

4.2. Temporal transfer

4.3. Reasoning transfer

TK: experiments should follow that order AND should be consistent with the used terminology

5. Experiments

TK: order: CLEVR + baselines without TL, COG + baseline without TL, Feature Transfer-CLEVR/CoGenT, Temporal Transfer - COG, Reasoning Transfer - CLEVR/CoGenT, Reasoning Transfer-COG

5.1. CLEVR/CoGenT dataset: baseline comparison

TK: description of CLEVR and CoGenT datasets, task families/groups

Dataset	Cubes	Cylinders	Spheres
CoGenT A	Family A	Family B	Any color
CoGenT B	Family B	Family A	Any color

Table 1: Restrictions on feature combinations in A & B conditions of the CoGenT variant of the CLEVR dataset.

TK: comparison of our model with selected baselines - pure CLEVR? (or CoGenT?), no transfer learning

TK: figure(s) with accuracy on CLEVR/CoGenT is/are missing!

5.2. COG dataset: baseline comparison

TK: description of dataset, task families/groups TK: figure 5 from the orig paper

For groups at the lowest level, we chose the following task classes to be placed in those groups. Below, substitute each of *Shape* and *Color* for X to obtain the task class.

Basic: *ExistX*, *GetX* and *Exist*;
Obj-Attr: *SimpleCompareX* and *AndSimpleCompareX*;
Compare: *CompareX*, *AndCompareX* & *ExistXOf*;
Spatial: *ExistSpace*, *ExistXSpace*, and *GetXSpace*;
Cognitive: *ExistLastColorSameShape*, *ExistLastShapeSameColor* and *ExistLastObjectSameObject*

TK: comparison of our model with baseline - pure COG, no transfer learning TK: figure 3 from the orig paper

baseline model [?]

We trained SAMNet using 8 reasoning steps (k=8) and external memory of 8 address locations, each storing an array of 128 floats. We compared our results with the baseline model introduced in the same paper as the COG dataset [?]. The most important results are highlighted in Figure 1; full comparison can be found in the supplementary material.

TK: comparison with results from softpahts?

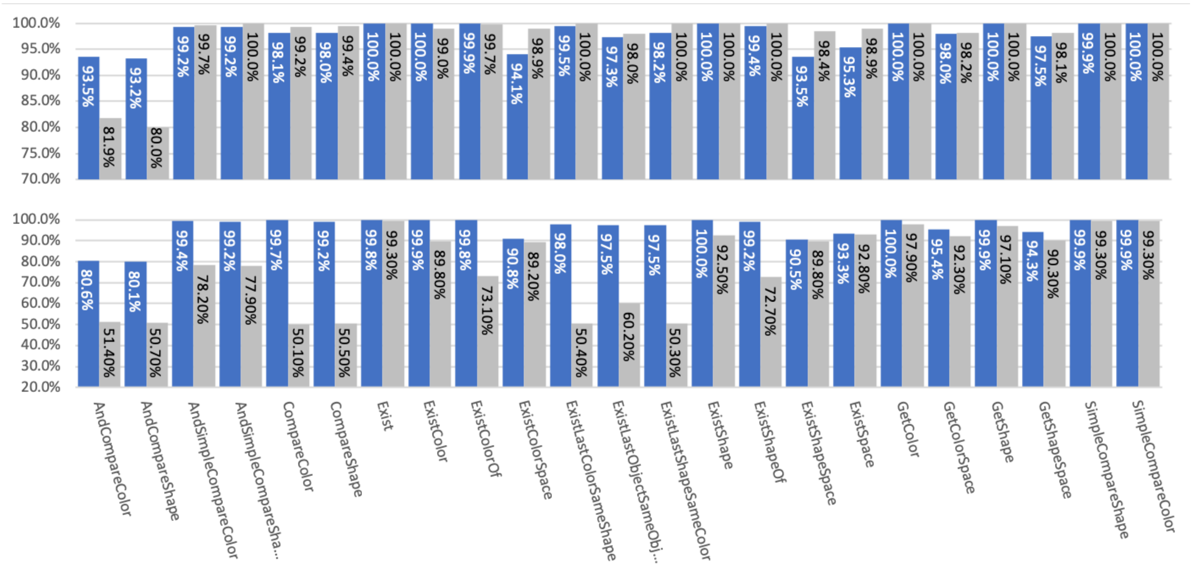

Figure 1: Comparison of test set accuracies of SAMNet (blue) with original results achieved by the baseline model [?] (gray) on Canonical (top) and Hard (bottom) variants of the COG dataset.

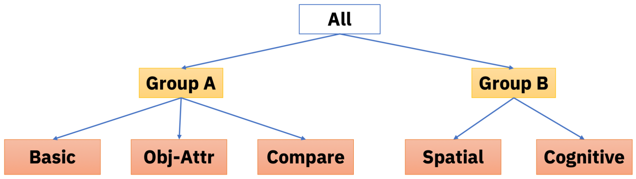

Figure 2: Hierarchy of Task Groups.

5.3. Feature transfer on CLEVR-CoGenT

TK: comparison of our model with baselines - CoGenT TK: figure 3 from the orig paper

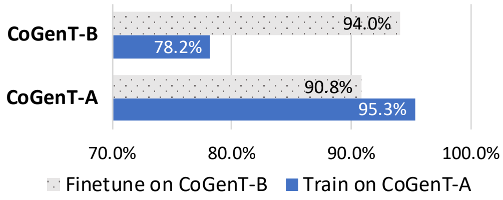

Figure 3: Test accuracy on CoGenT-A & -B when training on CoGenT-A (blue) and fine-tuning on CoGenT-B (gray).

5.4. Temporal transfer in COG

The goal here is to test the transfer learning ability concerning the frame sequence length, frame history required for reasoning, and the number of object distractors. For that purpose, we compare both models when trained on the Canonical variant but tested on the Hard variant (Figure 4). The light gray color indicates original accuracies of the baseline model from paper, whereas dark gray indicates accuracies of the baseline model obtained by running the original code provided by the authors [?].

The first column displays the scores of the traditional ML setup when training and testing on the Canonical variant. The observed close scores in light and dark gray underscore the baseline model reproducibility. For both cases of zero-shot learning (second column–91.6% vs 65.9%) and fine-tuning using a single epoch (third column–96.7% vs. 78.1%), SAMNET outperforms the baseline model significantly. Interestingly, this fine-tuning yields a mild boost of +0.6% on the earlier reported accuracy in **??** (fourth column). These results suggest that it suffices to train SAMNet on simpler videos to enable learning of good memory usage and attention on relevant entities in order to achieve comparable, if not better, performance on longer video frames with more complex scenes.

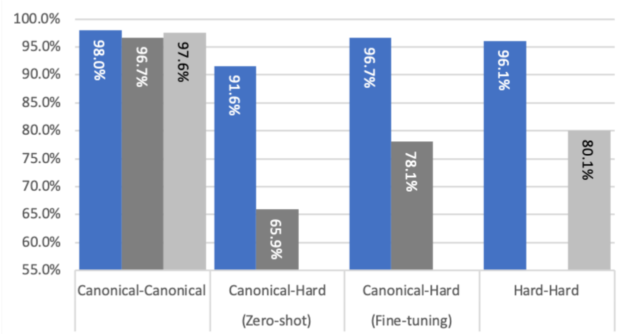

Figure 4: Total accuracies of SAMNet (blue) and baseline models (light/dark gray) when testing generalization from Canonical to Hard variants of the dataset.

5.5. Reasoning transfer on CLEVR-CoGenT

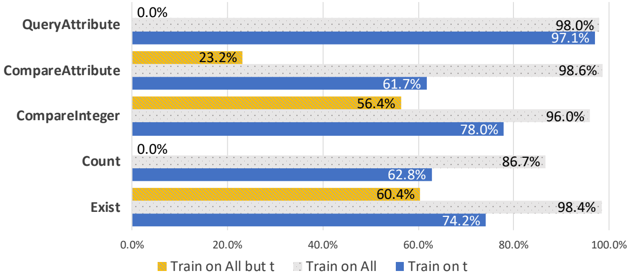

Figure 5: CLEVR-CoGenT accuracies for all tasks t when training on t only, training on all tasks jointly and training on all tasks but t.

5.5.1 Reasoning transfer on COG

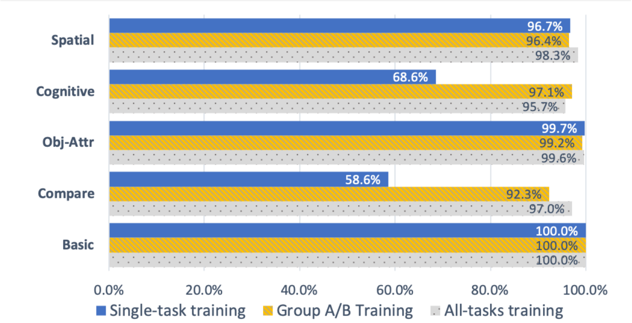

Figure 6: COG accuracies for all task groups t when training on t only; training on Group A or B; and on all tasks.

6. Summary