Graph-Based Global Reasoning Networks

Abstract

1. Introduction

> Relational reasoning between distant regions of arbitrary shape is crucial for many computer vision tasks.
> Humans can easily understand the relations among different regions of an image/video. However, deep CNNs cannot capture such relations without stacking multiple convolution layers, since an individual layer can only capture information locally.
> This is very inefficient, since relations between distant regions of arbitrary shape on the feature map can only be captured by a near-top layer with a sufficiently large receptive field to cover all the regions of interest. **To solve this problem, we propose a unit to directly perform global relation reasoning by projecting features from regions of interest to an interaction space and the distribute back to the original coordinate space. In this way, relation reasoning can be performed in early stages of a CNN model.**
>
> **latent interaction space** where global reasoning can be performed directly. Within this interaction space, a set of regions that share similar semantics are represented by a single feature, instead of a set of scattered coordinate-specific features from the input. Reasoning the relations of multiple different regions is thus simplified to modeling those between the corresponding features in the intersection space.

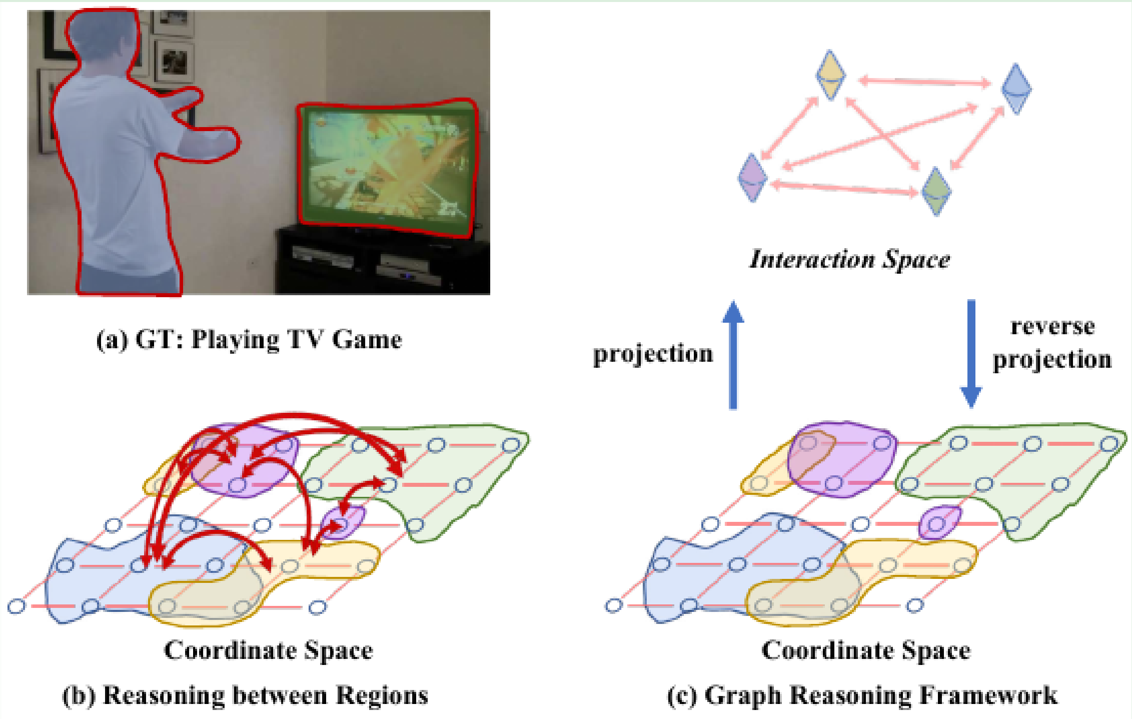


Figure 1: Illustration of our main idea. Aiming at capturing relations between arbitrary regions over the full input space (shown in different colors), we propose a novel approach for reasoning globally (shown in Fig. (c)). Features from the colored regions in coordinate space are projected into nodes in *interaction space*, forming a fully-connected graph. After reasoning over the graph, node features are projected back to the coordinate space.

We thus build a graph connecting these features within the interaction space ans perform relation reasoning over the graph.
After the reasoning, the updated information is then projected back to the original coordinate space for down-streaming tasks.
Accordingly, we devise a **Global Reasoning unit (GloRe)** to efficiently implement the coordinate-interaction space mapping process by weighted global pooling and weighted broadcasting, as well as the relation reasoning by graph convolution, which is differentiable and also end-to-end trainable.

Different from the recently proposed Non-local Neural Networks (NL-Nets) and Double Attention Networks which only focus on delivering information and rely on convolution layers for reasoning, our proposed model is able to directly reason on relations over regions.
Similarly, Squeeze-and-Excitation Networks (SE-Nets) only focus on incorporating image-level features via global average pooling, leading to an interaction graph containing only one node. It is not designed for regional reasoning as our proposed method.

> Contributions:
>
> - Reasoning globally. Mapping a set of features into an interaction space where reasoning can be computed efficiently and then mapping back to features.
>   Global Reasoning unit (GloRe) that implements the coordinate-interaction space mapping by weighted global pooling ans weighted broadcasting, and the relation reasoning via graph convolution in the interaction space.
>   We conduct extensive experiments on a number of datasets and show the Global Reasoning unit can bring consistent performance boost for a wide range of backbones including ResNet, ResNeXt, Se-Net and DPN for both 2D and 3D CNNs, on image classification, semantic segmentation and video action recognition task.

2. Related Work

> - **Deep Architecture Design**
>   Research on deep architecture on building more efficient convolution layer topologies, aiming at alleviating optimization difficulties or increasing efficiency of backbone architectures.
>   Residual Networks (ResNet), and DenseNet are proposed to alleviate the optimization difficulties of deep neural networks. DPN combines benefits of these two networks with further improved performance.
>   Xception, MobileNet, and ResNeXt use grouped or depth-wise convolutions to reduce the computational cost.
> - **Global Context Modeling**
>   Many efforts try to overcome the limitation of local convolution operators by introducing global contexts.
>   PSP-Net and DenseASPP combine multi-scale features to effectively enlarge the receptive field of the convolution layers for segmentation tasks.
> - **Graph-based Reasoning**
>   Graph-based methods have been very popular in recent years and shown to be an efficient way of relation reasoning.
>   We adopt the reasoning power of graph convolutions to build a generic, end-to-end trainable module for reasoning between disjoint and distant regions, regardless of their shape and without the need for object detectors or extra annotations.

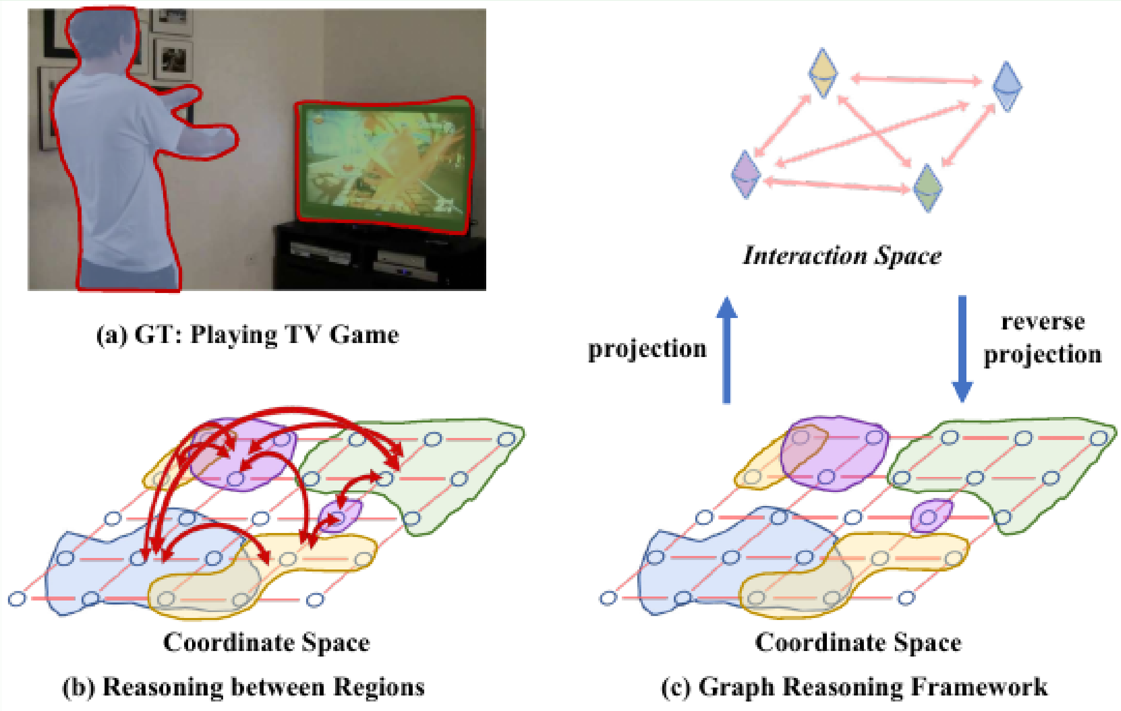
3. Graph-based Global Reasoning

> In this section, we first provide an overview of the proposed Global Reasoning unit, the core unit to our graph-based reasoning network, and introduce the motivation and rationale for its

> design. We then describe its architecture in details. Finally, we elaborate on how to apply it for several different computer vision tasks.

3.1 Overview

> Our proposed GloRe unit is motivated by overcoming the intrinsic limitation of convolution operations for modeling global relations.
> But capturing relations among disjoint and distant regions of arbitrary shape requires stacking multiple such convolution layers, which is highly inefficient.
> To solve this problem, we propose to first project the features X from the coordinate space Ω to the features V in a latent interaction space H, where each set of disjoint regions can be represented by a single feature instead of a bunch of features at different locations.
> Once we obtain the feature for each node of graph, we apply a general graph convolution to model and reason about the contextual relations between each pair of nodes.
> After that, we perform a reverse projection to transform the resulting features back to the original coordinate space, providing complementary features for the following layers to learn better task-specific representations. Figure 1(c)



(a) GT: Playing TV Game

(b) Reasoning between Regions

(c) Graph Reasoning Framework

Figure 1: Illustration of our main idea. Aiming at capturing relations between arbitrary regions over the full input space (shown in different colors), we propose a novel approach for reasoning globally (shown in Fig. (c)). Features from the colored regions in coordinate space are projected into nodes in *interaction space*, forming a fully-connected graph. After reasoning over the graph, node features are projected back to the coordinate space.

3.2 From Coordinate Space to Interaction Space

The first step is to find the projection function $f(*)$ that maps original features to the interaction space H. The new features in the interaction space are more friendly for global reasoning over disjoint and distant regions.

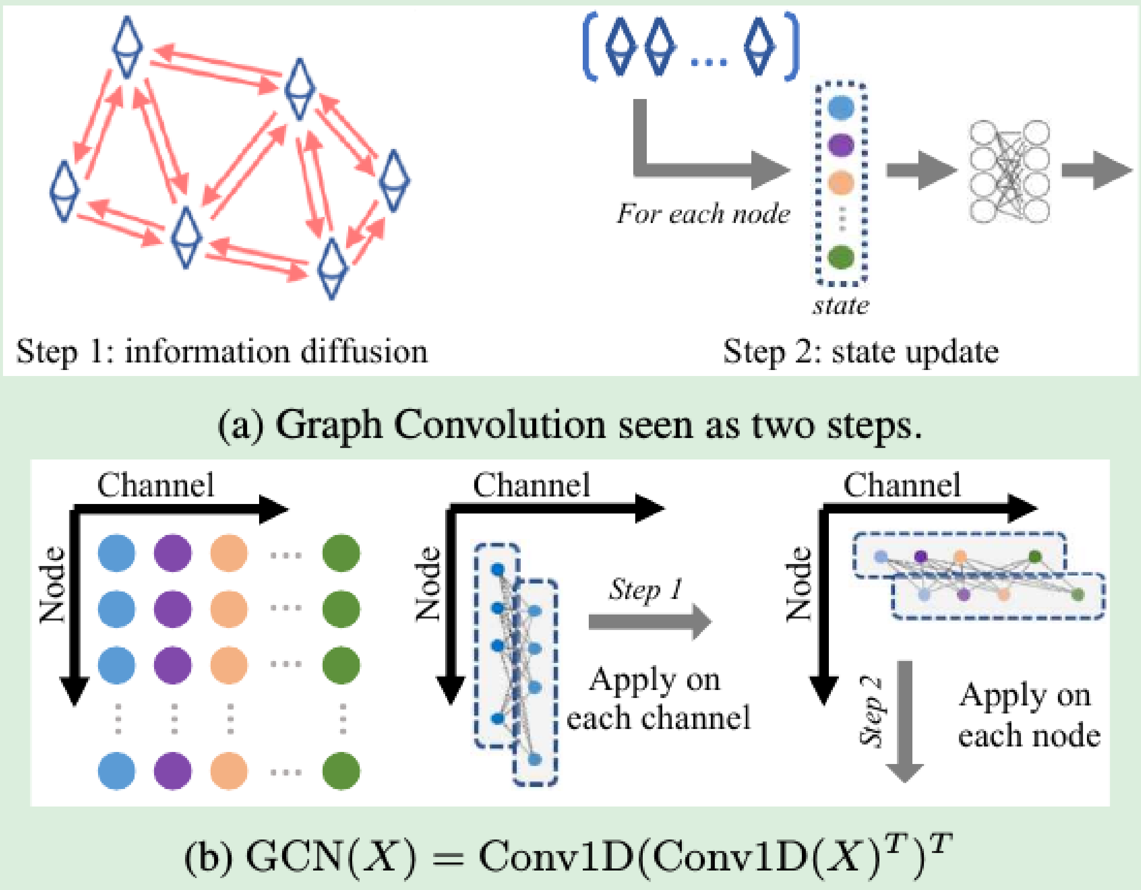
1. The convolution layer is end-to-end trainable.
2. Its training does not require any object bounding box.
3. It is simple to implement and faster in speed.
4. It is more generic since the convolution output can be both positive and negative, which linearly fuses the information in the coordination space.

3.3 Reasoning with Graph Convolution

> After projecting the features from coordinate space into the interaction space, we have graph where each node contains feature descriptor. Capturing relations between arbitrary regions in the input is now simplified to capturing interactions between the features of the corresponding nodes.
>
> **graph convolution**

3.4 From Interaction Space to Coordinate Space

> To make the above building block compatible with existing CNN architectures, the last step is to project the output features back to the original space the relation reasoning. In this way, the updated features from reasoning can be utilized by the following convolution layers to make better decisions. This reverse projection is very similar to the projection in the first step.



(a) Graph Convolution seen as two steps.

(b) $\mathrm{GCN}(X) = \mathrm{Conv1D}(\mathrm{Conv1D}(X)^T)^T$

> Figure 3: Relation reasoning through graph convolution. (a) An intuitive explanation of graph convolution. (b) Implementation of Graph Convolution using two-direction 1D convolutions.

3.5 Deploying the Global Reasoning Unit

> The core processing of the proposed Global Reasoning unit happens after flattening all dimensions referring to locations. It therefore straightforwardly applies to 3D or 1D features by adapting the dimensions that operate in the coordinate space and then flattening the corresponding dimensions.
> In practice, due to its residual nature, the proposed Global Reasoning unit can be easily incorporated into a large variety of existing backbone CNN architectures.
> **Graph-Based Global Reasoning Networks**.

4. Experiments

> We begin with image classification task on the large-scale ImageNet dataset for studying key properties of the proposed method, which servers as the main benchmark dataset.

4.1 Implementation Details

> - **Image Classification**
>   We first use ResNet-50 as a shallow CNN to conduct ablation studies and then use deeper CNNs to further exam the effectiveness of the proposed method.
> - **Semantic Image Segmentation**
> - **Video Action Recognition**

4.2 Results on ImageNet

> We first conduct ablation studies using ResNet-50 as the backbone architecture and considering two scenarios: 1) when only one extra block is added; 2) when multiple extra blocks are added.
>
> - **Ablation Study**

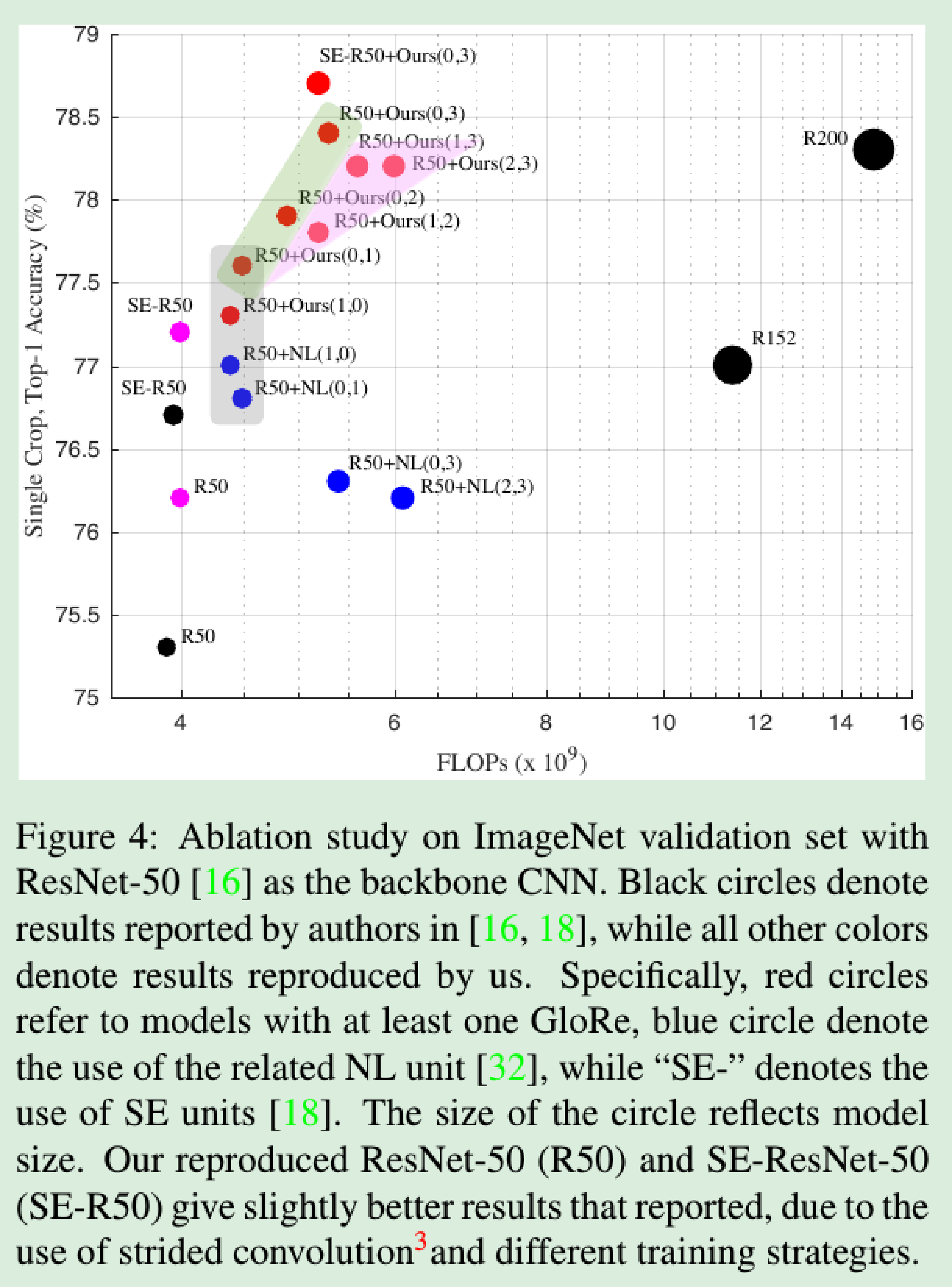
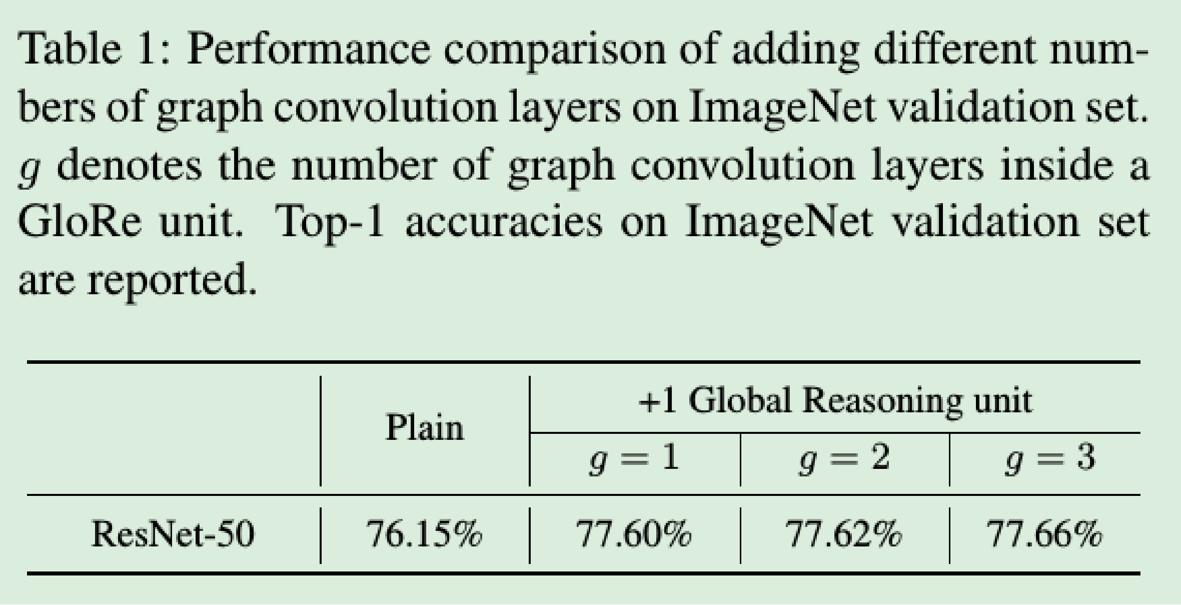


Figure 4: Ablation study on ImageNet validation set with ResNet-50 [16] as the backbone CNN. Black circles denote results reported by authors in [16, 18], while all other colors denote results reproduced by us. Specifically, red circles refer to models with at least one GloRe, blue circle denote the use of the related NL unit [32], while "SE-" denotes the use of SE units [18]. The size of the circle reflects model size. Our reproduced ResNet-50 (R50) and SE-ResNet-50 (SE-R50) give slightly better results that reported, due to the use of strided convolution[3] and different training strategies.



> Table 1: Performance comparison of adding different numbers of graph convolution layers on ImageNet validation set. g denotes the number of graph convolution layers inside a GloRe unit. Top-1 accuracies on ImageNet validation set are reported.
>
> | | Plain | +1 Global Reasoning unit | | |
> |---|---|---|---|---|
> | | | $g = 1$ | $g = 2$ | $g = 3$ |
> | ResNet-50 | 76.15% | 77.60% | 77.62% | 77.66% |

4.3 Results on Cityscapes

4.4 Results on Kinetics

4.5 Visualizing the GloRe Unit

5. Conclusion

> In this paper, we present a highly efficient approach for global reasoning that can be effectively implemented by project information from the coordinate space to nodes in an interaction space graph where we can directly reason over globally-aware discriminative features.
> The GloRe unit is an efficient instantiation of the proposed approach, where projection and reverse projection are implemented by weighted pooling and weighted broadcasting, respectively, and interactions over the graph are modeled via graph convolution.
> It is lightweight, easy to implement and optimize, while extensive experiments show that it can effectively learn features complementary to various popular CNNs and consistently boost their performance on both 2D and 3D tasks over a number of datasets.