Learning efficient representations for image and video understanding

a selection of recent works (*)

Yannis Kalantidis
Universidad Politécnica de Cataluña
Barcelona, 21 November 2019

(*) Research conducted as a research scientist at **Facebook AI**

Overview

Motivation

Challenges for Representation Learning

Reducing computation: **Octave Convolutions**

Non-local reasoning: **Global Reasoning Networks**

Higher-level understanding: **Grounded Video Description**

Summary

Overview

Motivation

Challenges for Representation Learning

Reducing computation: **Octave Convolutions**

Non-local reasoning: **Global Reasoning Networks**

Higher-level understanding: **Grounded Video Description**

Summary

Motivation

Visual data is ubiquitous

- Smartphones / Social media
  - ***Video*** the most popular form of communication online
- Self-driving cars
- Robotics & Drones
- Satellites data
- Medical devices

More people today have **access to mobile phones than to piped water supply**
[World Bank]

Satellite Data and Traditional Knowledge at Work for Kenya's Pastoralists [agrilinks.org]

Large online farming communities
https://www.facebook.com/africafarmersclub/

Visual representations

Compact representations of the visual inputs
that are more useful for solving vision tasks

“Hand-crafted” representations

- Utilizes domain knowledge
- Requires domain expertise
- Most common approach for decades

Learning representations

- Utilize large amounts of data
- Don’t design features; design models
- Use machine learning to optimize the parameters

Learning “better” representations

What is “better”?

- **Performance**

  Higher performance on end tasks
- **Generalization**

  Faster learning for new tasks
- **Efficiency**

  Smaller models, more expressive

  Faster testing & deployment

Overview

Motivation

Challenges for Representation Learning

Reducing computation: **Octave Convolutions**

Non-local reasoning: **Global Reasoning Networks**

Higher-level understanding: **Grounded Video Description**

Summary

Challenges for representation learning

Reducing computation

- Learning smaller but expressive models
- Learning in the compressed domain

Non-local & spatio-temporal understanding

- Long-term/range reasoning
- Reasoning on structured inputs

Higher-level & multi-modal reasoning

- Learning from diverse modalities and language
- Understanding human behaviour

Learning under more realistic scenarios

- Learning from fewer & imbalanced data
- Generalizing to novel classes (zero-/one-shot)

Recent publications on visual understanding

Exploiting parameter redundancy
Multi-Fiber Networks [ECCV 2018]

Reducing spatial redundancy
Octave Convolutions [ICCV 2019]

Fast discriminative motion cues
DMC-Net [CVPR 2019]

Compact spatio-temporal attention
A^2-Nets: Double Attention Networks [NeurIPS 2018]

Attend and reason globally
Graph-based Global Reasoning Networks [CVPR 2019]

Self-supervised learning of video highlights
Less is more [CVPR 2019]

Visual Grounding of natural language
Grounded video description [CVPR 2019]
Cyclical training [arXiv 2019]

Learning from long-tailed distributions
Large-Scale Visual Relationship Understanding [AAAI 2019]
Decoupling representation and classifier [arXiv 2019]

Reducing computation

Non-local & spatio-temporal understanding

Higher-level & multi-modal reasoning

Learning from fewer/imbalanced data

Recent publications on visual understanding

Exploiting parameter redundancy
Multi-Fiber Networks [ECCV 2018]

Reducing spatial redundancy
Octave Convolutions [ICCV 2019]

Fast discriminative motion cues
DMC-Net [CVPR 2019]

Compact spatio-temporal attention
A^2-Nets: Double Attention Networks [NeurIPS 2018]

Attend and reason globally
Graph-based Global Reasoning Networks [CVPR 2019]

Self-supervised learning of video highlights
Less is more [CVPR 2019]

Visual Grounding of natural language
Grounded video description [CVPR 2019]
Cyclical training [arXiv 2019]

Learning from long-tailed distributions
Large-Scale Visual Relationship Understanding [AAAI 2019]
Decoupling representation and classifier [arXiv 2019]

Reducing computation

Non-local & spatio-temporal understanding

Higher-level & multi-modal reasoning

Learning from fewer/imbalanced data

Overview

Motivation

Challenges for Representation Learning

Reducing computation: Octave Convolutions

Non-local reasoning: **Global Reasoning Networks**

Higher-level understanding: **Grounded Video Description**

Summary

A Vision For the Future

Drop an Octave: Reducing Spatial Redundancy in Convolutional Neural Networks with Octave Convolution

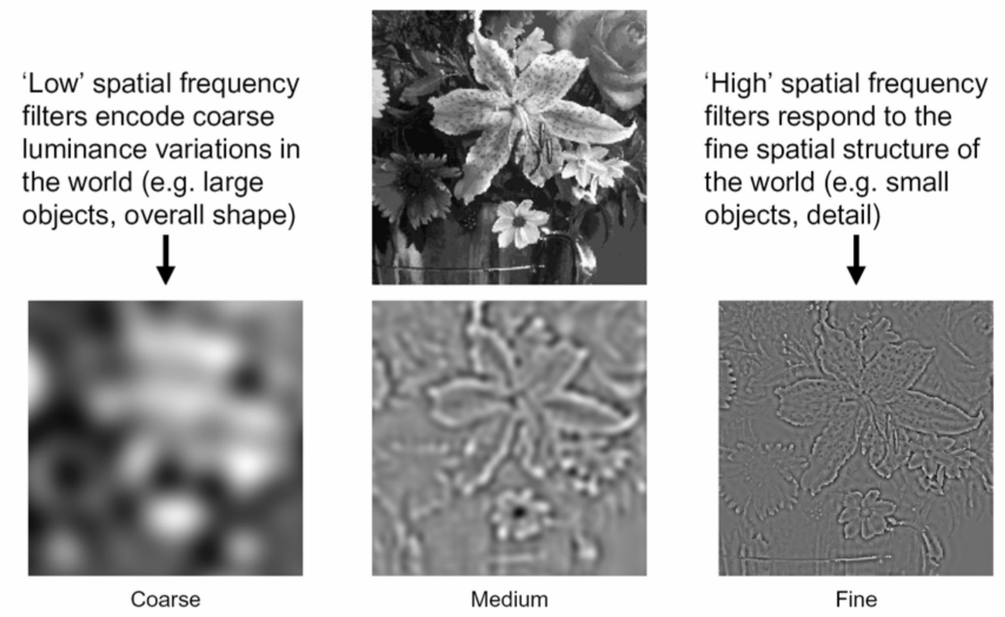
Yunpeng Chen[†‡], Haoqi Fan[†], Bing Xu[†], Zhicheng Yan[†], Yannis Kalantidis[†],
Marcus Rohrbach[†], Shuicheng Yan[‡♭], Jiashi Feng[‡]
[†]Facebook AI, [‡]National University of Singapore, [♭]Yitu Technology

[ICCV 2019]

Reducing Spatial Redundancy

Spatial-redundancy in feature maps

- Convolution kernels are highly local
- Not all convolutional kernels are high frequency filters
- Some feature maps must contain low frequency information that is smooth and slowly varying

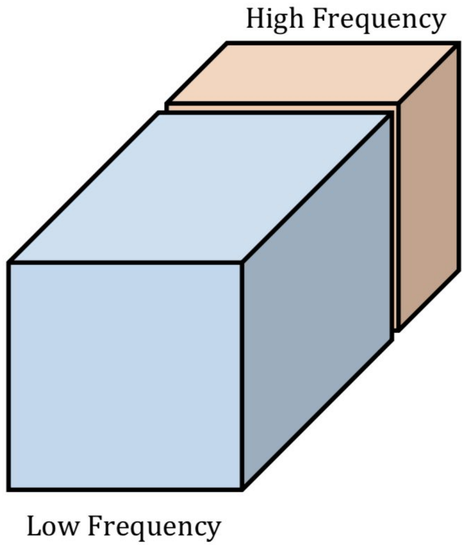


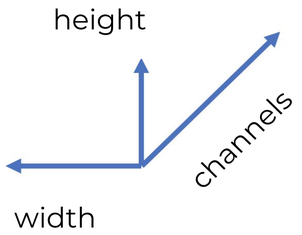
[FW Campbell and JG Robson. Application of fourier analysis to the visibility of gratings. *The Journal of physiology*, 1968]
[RL DeValois and KK DeValois. Spatial vision. Oxford psychology series, 1988]

Figure source: "Tutorial on Spatial Frequency Analysis", credits: Jon Morton

The multi-frequency feature representation

- Split convolutional feature maps into two groups at different spatial frequencies

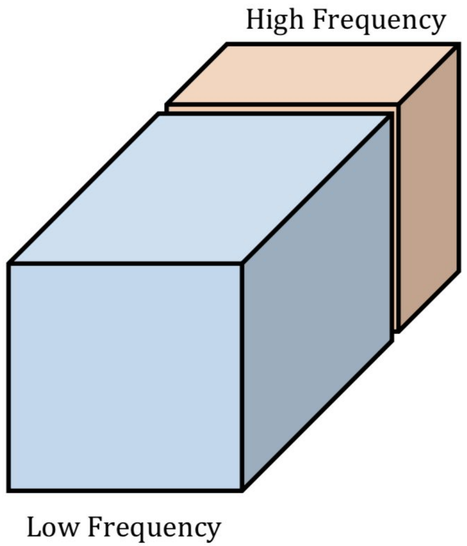


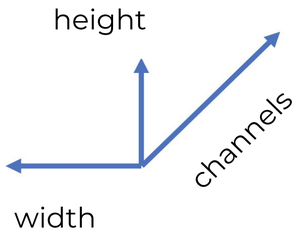


The multi-frequency feature representation

- Split convolutional feature maps into two groups at different spatial frequencies

How to exploit the spatial redundancy in the low frequency maps?



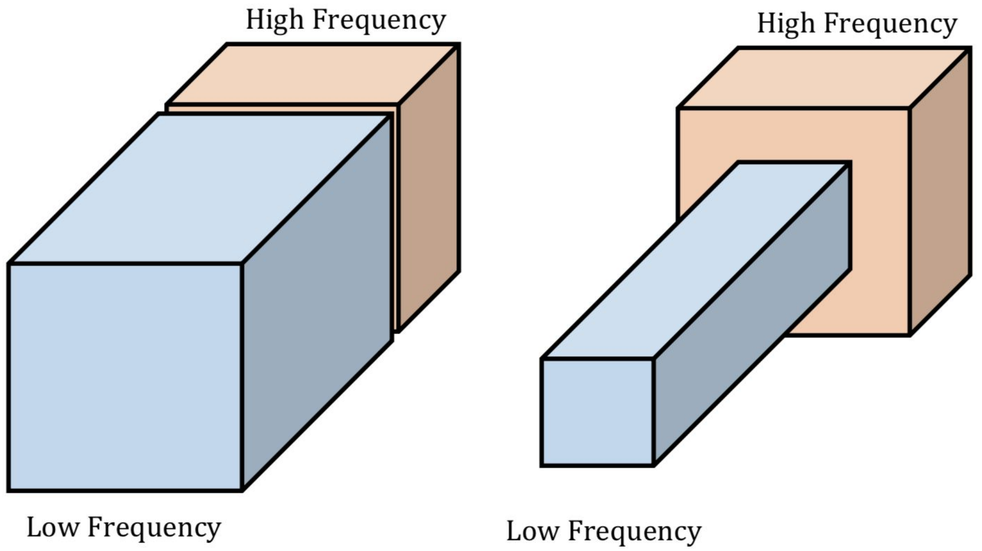


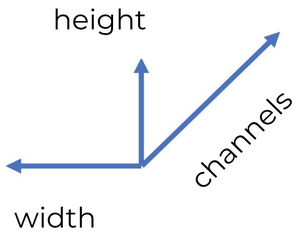
The multi-frequency feature representation

- Split convolutional feature maps into two groups at different spatial frequencies

How to exploit the spatial redundancy in the low frequency maps?

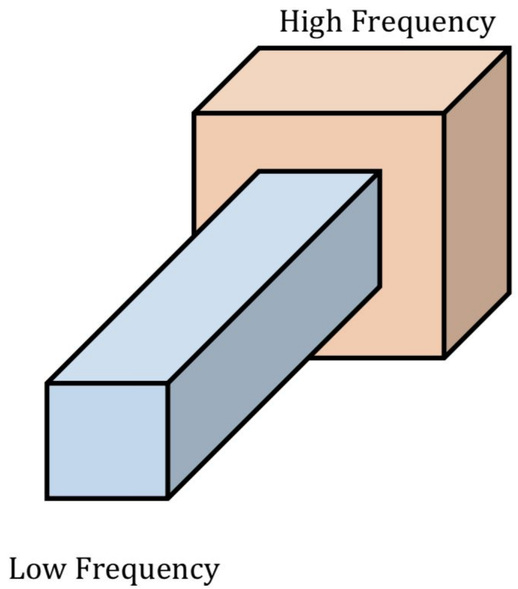
- Reduce their spatial resolution
- **Process them “one octave” lower**

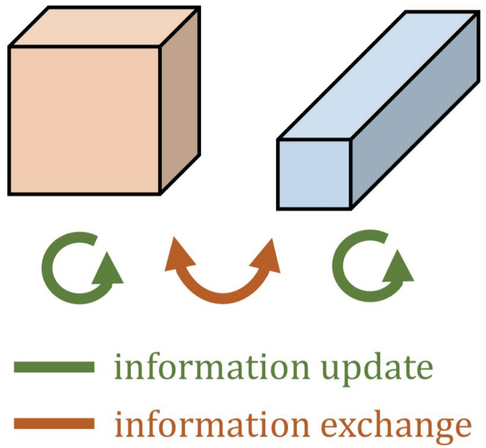




Octave Convolution

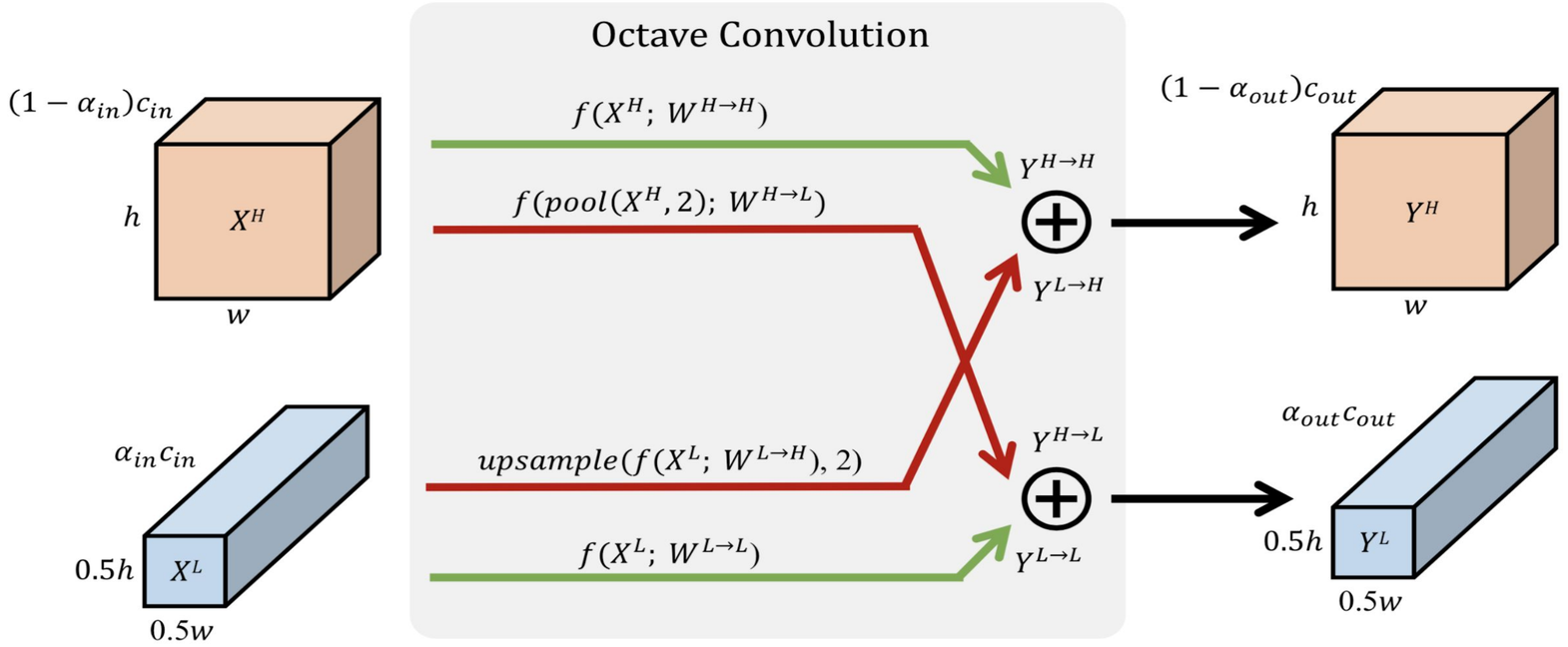
- A convolution operation that operates directly on our multi-frequency representation

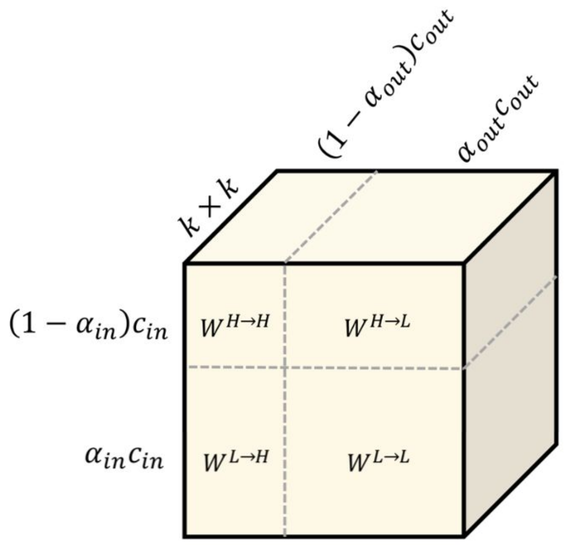




Architecture details - The OctConv kernel

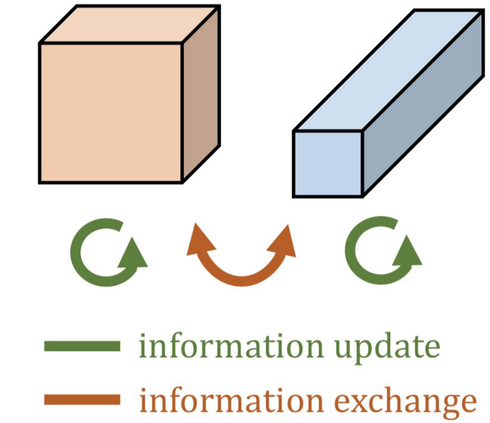
- Hyper-parameter α: The ratio of low / high maps

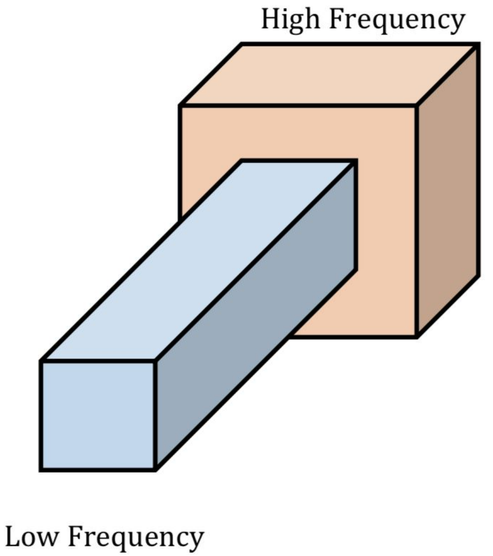




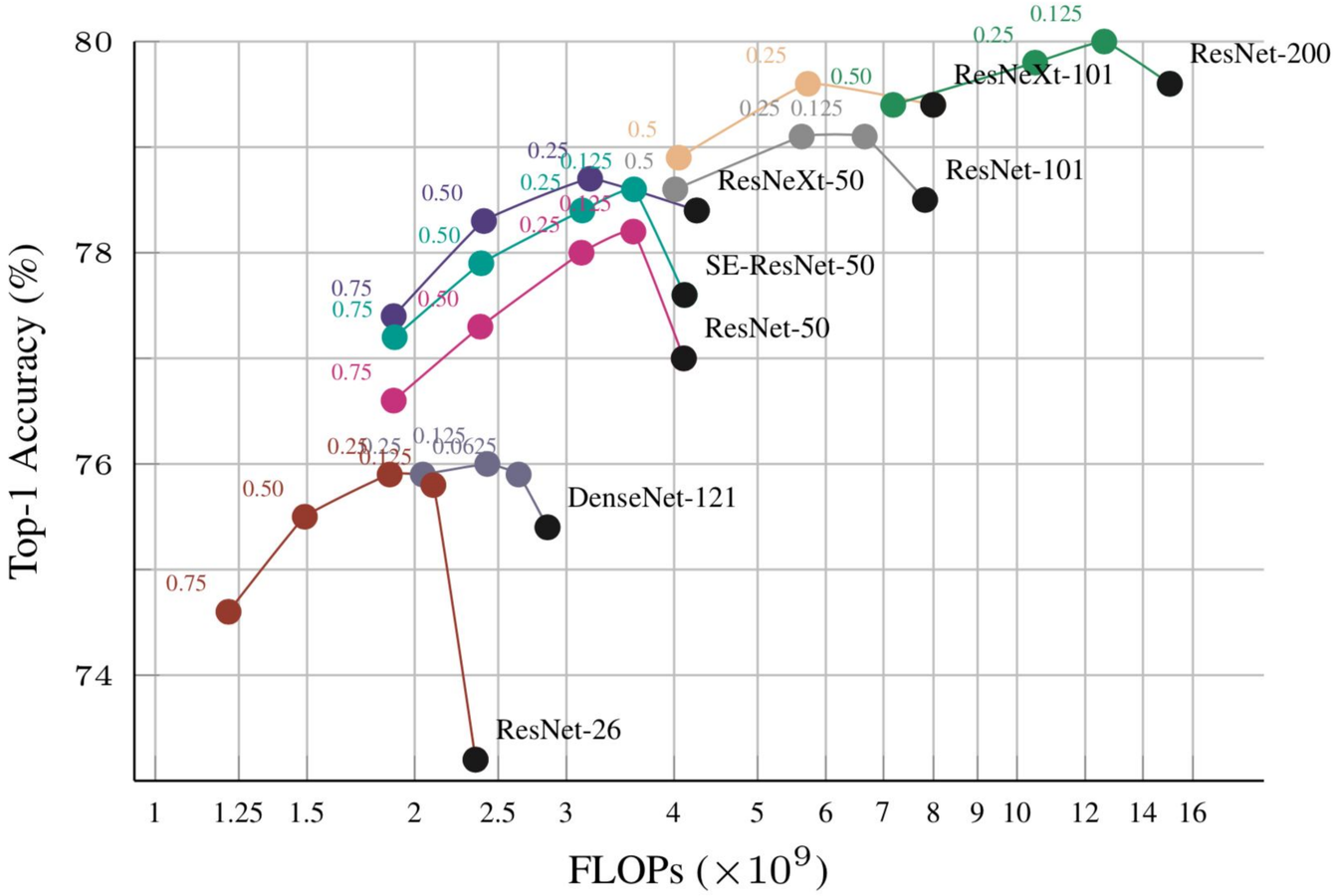
Advantages of the Octave Convolution

- **Multi-scale** processing with efficient **information exchange** between the low- and high-frequency maps
- Multi-scale processing **without altering the network architecture**
- Gains in terms of both computations & memory
- **Larger receptive field** for the low-frequency feature maps





Behaviour of OctConv for different ratios (α)



ImageNet for varying models and ratios

Results on Image Classification (ImageNet)

Small models (*)

- + 1.6% Acc for ShuffleNet
- + 2.2% Acc for MobileNet
- + 1% Acc for MobileNet (v2)
  - **73.0%** top-1 Acc, 295 MFLOPS

(*) for models with approx. the same GLOPS

Medium models (ResNe(X)t-50/101) (*)

- + 0.9% Acc vs Big-little Net [Chen *et al* 2019]
- + 1.9% Acc vs MG-Conv [Ke *et al* 2017]
- Oct-ResNeXt-101: **79.6 %** top-1 Acc with 44.2M params and 5.7 GFLOPS

[Ke *et al*. Multigrid Neural Architectures, CVPR 2017]
[Huang *et al*. Multi-Scale Dense Networks for Resource Efficient Image Classification, ICLR 2018]
[Chen *et al.* Big-Little Net: An Efficient Multi-Scale Feature Representation for Visual and Speech Recognition, ICLR 2019]

Results on Image Classification (ImageNet)

Method	#Params (M)	Training			Testing (320×320 / 331×331)		
		Input Size	Memory Cost (MB)	Speed (im/s)	#FLOPs (G)	Top-1 (%)	Top-5 (%)
NASNet-A (N=6, F=168) [48] ◇	88.9	331×331 / 320×320	$> 32,480$	43 ‡	23.8	82.7	96.2
AmoebaNet-A (N=6, F=190) [33] ◇	86.7		$> 32,480$	47 ‡	23.1	82.8	96.1
PNASNet-5 (N=4, F=216) [29] ◇	86.1		$> 32,480$	38 ‡	25.0	82.9	96.2
Squeeze-Excite-Net [21]	115.1		$> 32,480$	43 †	42.3	83.1	96.4
AmoebaNet-A (N=6, F=448) [33] ◇	469		$> 32,480$	15 §	104	83.9	96.6
Dual-Path-Net-131 [8]	79.5	224×224	31,844	83	32.0	81.5	95.8
SE-ShuffleNet v2-164 [32]	69.9		$> 32,480$	70 †	-	-	-
Squeeze-Excite-Net [21]	115.1		28,696	78	42.3	82.7	96.2
Oct-ResNet-152, $\alpha = 0.125$ (ours)	**60.2**		**15,566**	**162**	**22.2**	82.3	96.0
Oct-ResNet-152 + SE[4], $\alpha = 0.125$ (ours)	66.8		21,885	95	**22.2**	**82.9**	**96.3**

Results on large models

Results on Image Classification (ImageNet)

Method	#Params (M)	Training			Testing (320 × 320 / 331 × 331)		
		Input Size	Memory Cost (MB)	Speed (im/s)	#FLOPs (G)	Top-1 (%)	Top-5 (%)
NASNet-A (N=6, F=168) [48] ◇	88.9	331 × 331 / 320 × 320	> 32,480	43 ‡	23.8	82.7	96.2
AmoebaNet-A (N=6, F=190) [33] ◇	86.7		> 32,480	47 ‡	23.1	82.8	96.1
PNASNet-5 (N=4, F=216) [29] ◇	86.1		> 32,480	38 ‡	25.0	82.9	96.2
Squeeze-Excite-Net [21]	115.1		> 32,480	43 †	42.3	83.1	96.4
AmoebaNet-A (N=6, F=448) [33] ◇	469		> 32,480	15 §	104	83.9	96.6
Dual-Path-Net-131 [8]	79.5	224 × 224	31,844	83	32.0	81.5	95.8
SE-ShuffleNet v2-164 [32]	69.9		> 32,480	70 †	-	-	-
Squeeze-Excite-Net [21]	115.1		28,696	78	42.3	82.7	96.2
Oct-ResNet-152, $\alpha = 0.125$ (ours)	**60.2**		**15,566**	**162**	**22.2**	82.3	96.0
Oct-ResNet-152 + SE[4], $\alpha = 0.125$ (ours)	66.8		21,885	95	**22.2**	**82.9**	**96.3**

Results on large models

Results on Action Recognition (Kinetics)

Method	ImageNet Pretrain	#FLOPs (G)	Top-1 (%)	
I3D	✓	28.1	73.3	
Oct-I3D, α=0.1, (ours)	✓	**25.6**	**74.6**	**(+1.3)**
I3D + Non-local	✓	33.3	74.7	
Oct-I3D + Non-local, α=0.1, (ours)	✓	**28.9**	**75.7**	**(+1.0)**
SlowFast-R50 [12]		27.6 [5]	75.6	
Oct-SlowFast-R50, α=0.1, (ours)		**24.5**	**76.2**	**(+0.6)**
Oct-SlowFast-R50, α=0.2, (ours)		22.9	75.8	(+0.2)
(b) Kinetics-600 [2]				
I3D	✓	28.1	74.3	
Oct-I3D, α=0.1, (ours)	✓	**25.6**	**76.0**	**(+1.7)**

import OctConv as Conv

Octave Convolution

- An efficient, plug-and-play convolution operation that operates directly on our multi-frequency representation
- Better performance with less compute
- Consistent gains on different tasks and backbones
- Models and code are open-sourced (many 3rd party implementations)

... try it!

Overview

Motivation

Challenges for Representation Learning

Reducing computation: **Octave Convolutions**

Non-local reasoning: Global Reasoning Networks

Higher-level understanding: **Grounded Video Description**

Summary

A Vision For the Future

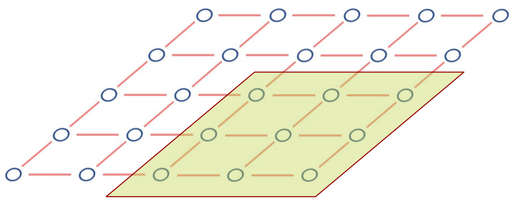
Graph-Based Global Reasoning Networks

Yunpeng Chen[†‡], Marcus Rohrbach[†], Zhicheng Yan[†], Shuicheng Yan[‡♭], Jiashi Feng[‡], Yannis Kalantidis[†]
[†]Facebook Research, [‡]National University of Singapore, [♭]Qihoo 360 AI Institute

[CVPR 2019]

Global Context Modeling

- Convolutions model **local** relations on the (spatio-temporal) coordinate space
- Require stacking multiple layers to capture relations between distant locations

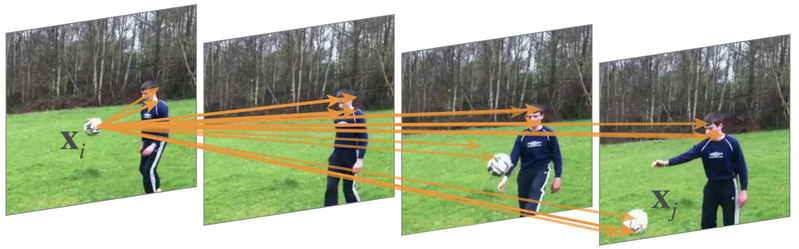


Incorporating non-local context

- Attention mechanisms

  [Vaswani *et al.* 2017, Wang *et al.* 2018, Chen *et al.* 2018]

- Interactions between all locations



[Wang *et al.* 2018]

[Vaswani *et al.* **Attention is all you need.** NIPS 2017]
[Wang *et al.* **Non-local Neural Networks**. CVPR, 2018]
[Chen, **Kalantidis**, *et al.* **A^2-Nets: Double Attention Networks.** NeurIPS 2018]

Motivation

Global context modeling is highly important

- Attention-like mechanisms becoming standard
- Doesn’t capture ***region*** interactions

GT: Playing TV Game

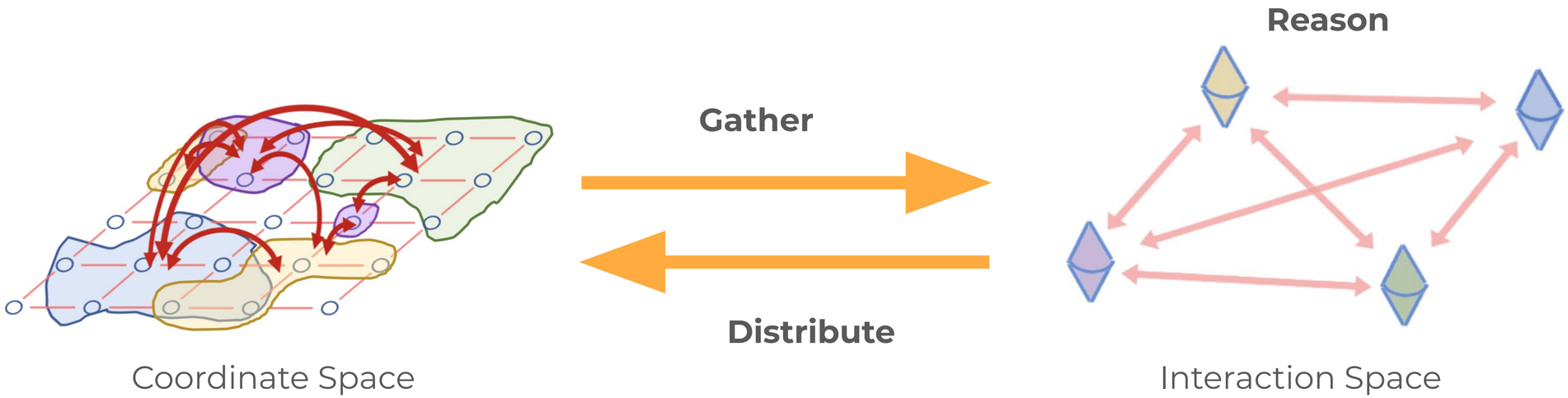
A limitation of current global context modeling

- Gather → Distribute: focus on delivering information
- Rely on convolutional layers for reasoning

*Figures depict the 2D (image) case for clarity

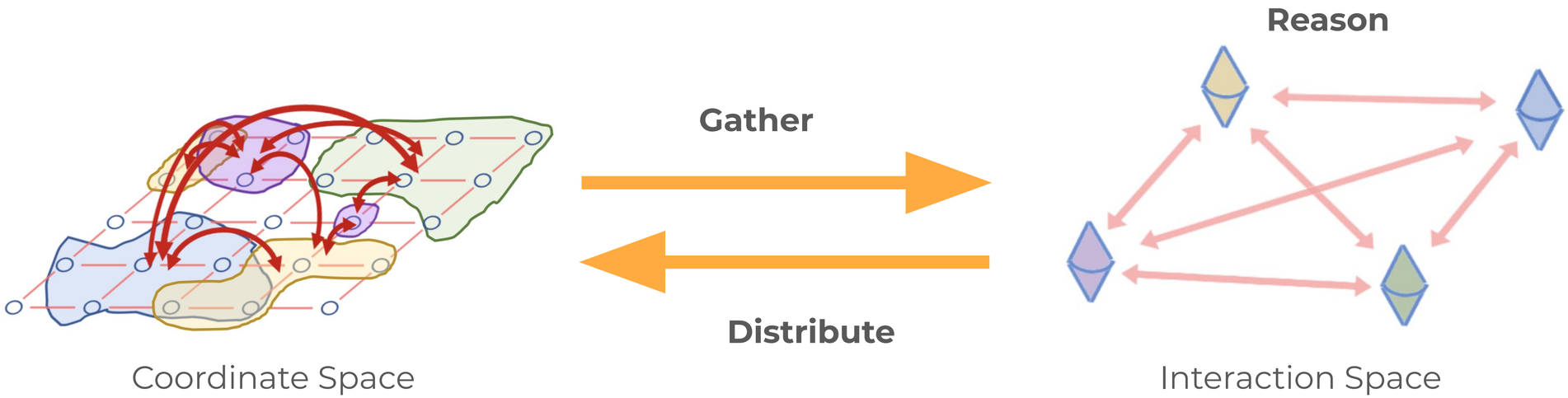
Global Reasoning Unit

- Beyond locations: capture (arbitraty) **region interactions**
- Gather → **Reason** → Distribute



Global Reasoning Unit

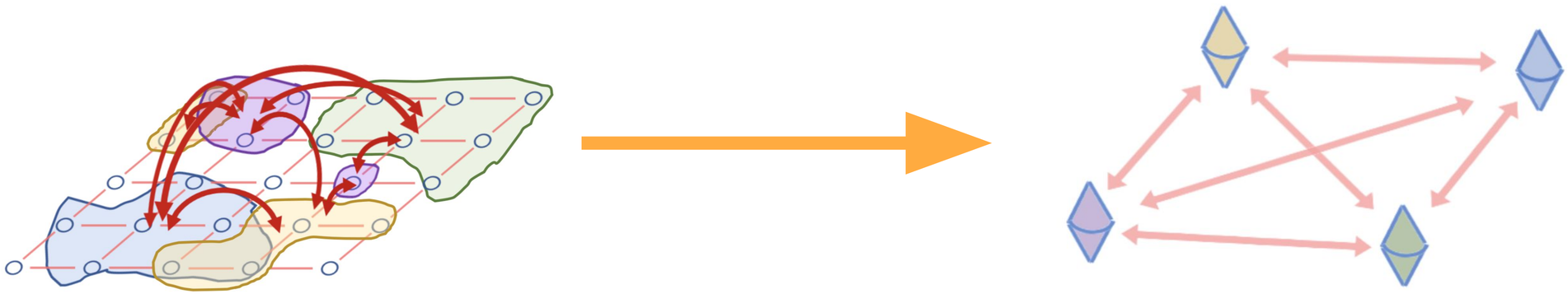
- Beyond locations: capture (arbitraty) **region interactions**
- Gather → **Reason** → Distribute



(arbitrary) region interactions in coordinate space reduce to much simpler, pairwise interactions in interaction space

Global Reasoning in three steps

1) From Coordinate Space to Interaction Space → Weighted projections

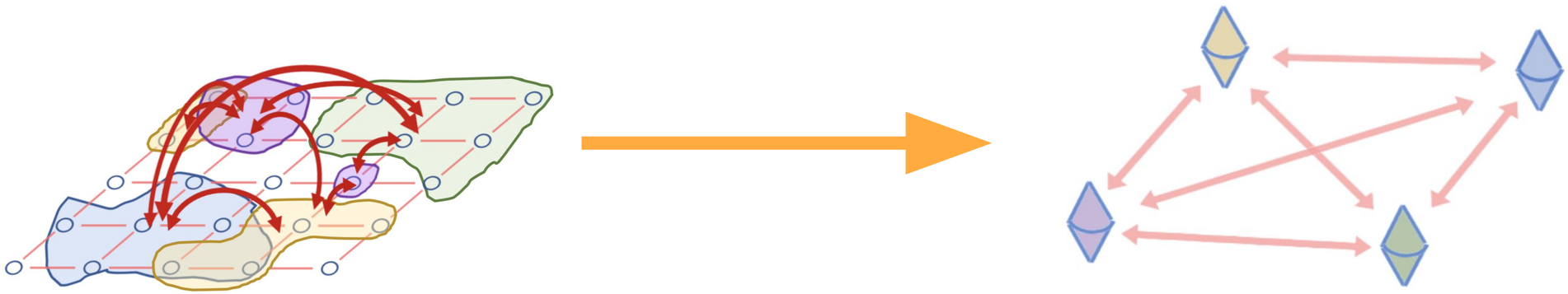


Coordinate Space

Interaction Space

Global Reasoning in three steps

1) From Coordinate Space to Interaction Space → Weighted projections
2) Reasoning in Interaction Space → Graph convolutions

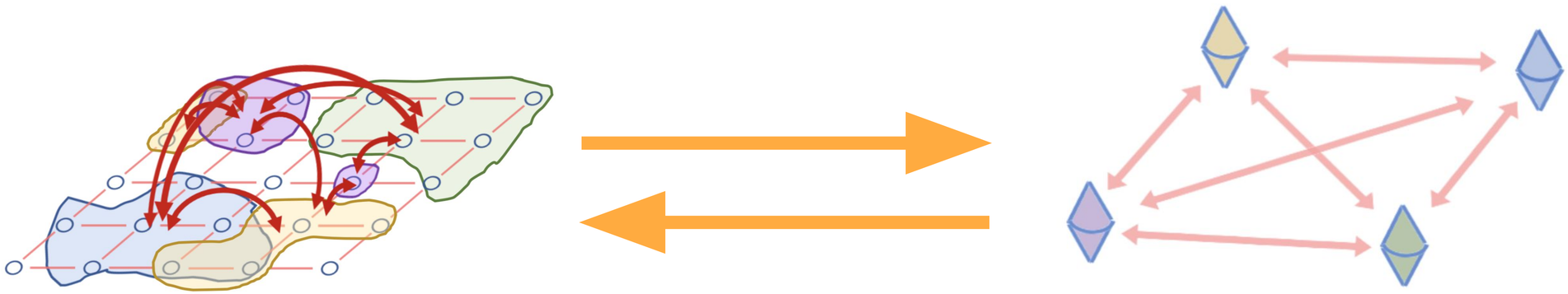


Coordinate Space

Interaction Space

Global Reasoning in three steps

1) From Coordinate Space to Interaction Space → Weighted projections
2) Reasoning in Interaction Space → Graph convolutions
3) From Interaction Space to Coordinate Space → Weighted broadcasting

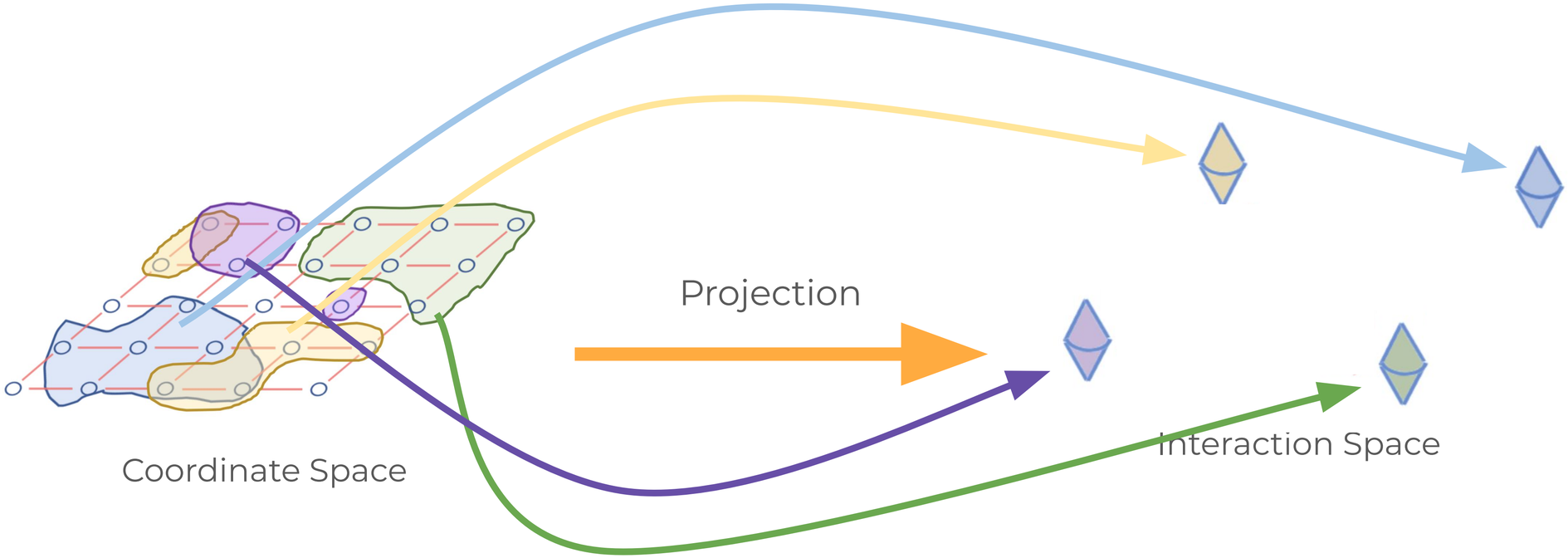


Coordinate Space

Interaction Space

From Coordinate Space to Interaction Space

- Learn a set of projections for (arbitrary) region features



From Coordinate Space to Interaction Space

Given a set of input features $X \in \mathbb{R}^{L \times C}$, compute projections $V \in \mathbb{R}^{N \times C}$

$$\mathbf{v}_i = \mathbf{b}_i X = \sum_{\forall j} b_{ij} \mathbf{x}_j$$

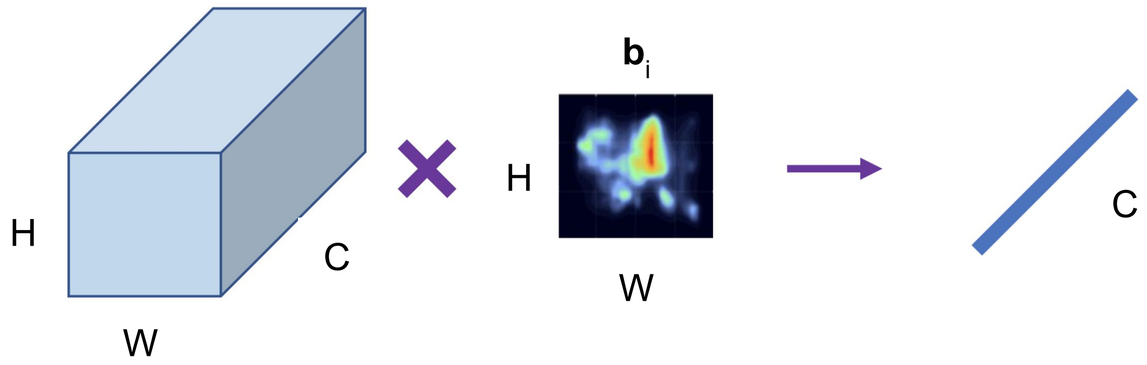
learnable projection weights

$$B = [\mathbf{b}_1, \cdots, \mathbf{b}_N] \in \mathbb{R}^{N \times L}$$

From Coordinate Space to Interaction Space

Given a set of input features $X \in \mathbb{R}^{L \times C}$, compute projections $V \in \mathbb{R}^{N \times C}$

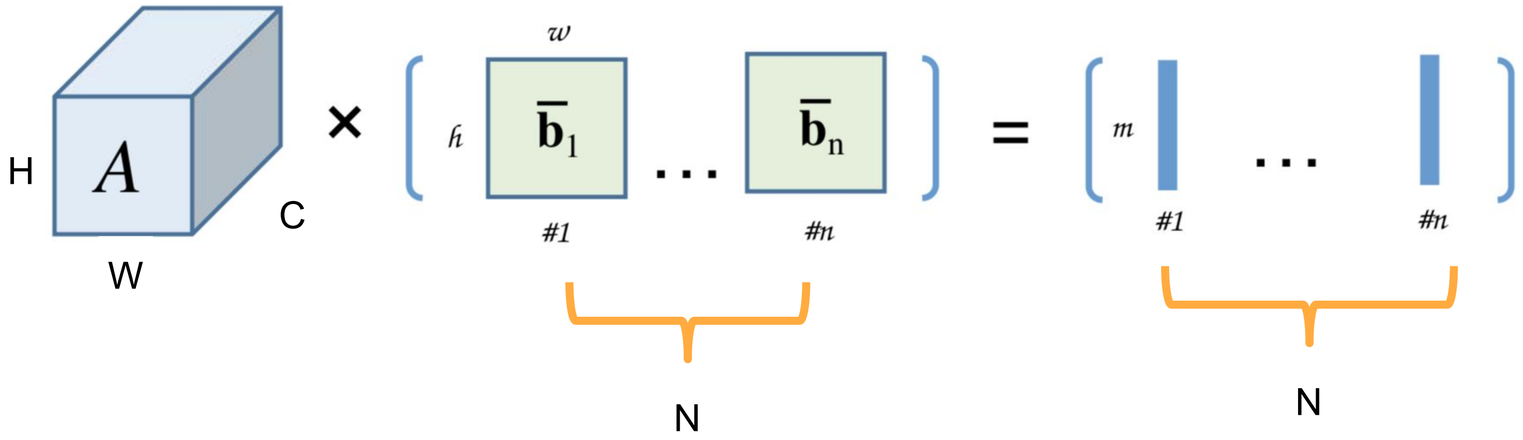
$$\mathbf{v}_i = \mathbf{b}_i X = \sum_{\forall j} b_{ij} \mathbf{x}_j$$



From Coordinate Space to Interaction Space

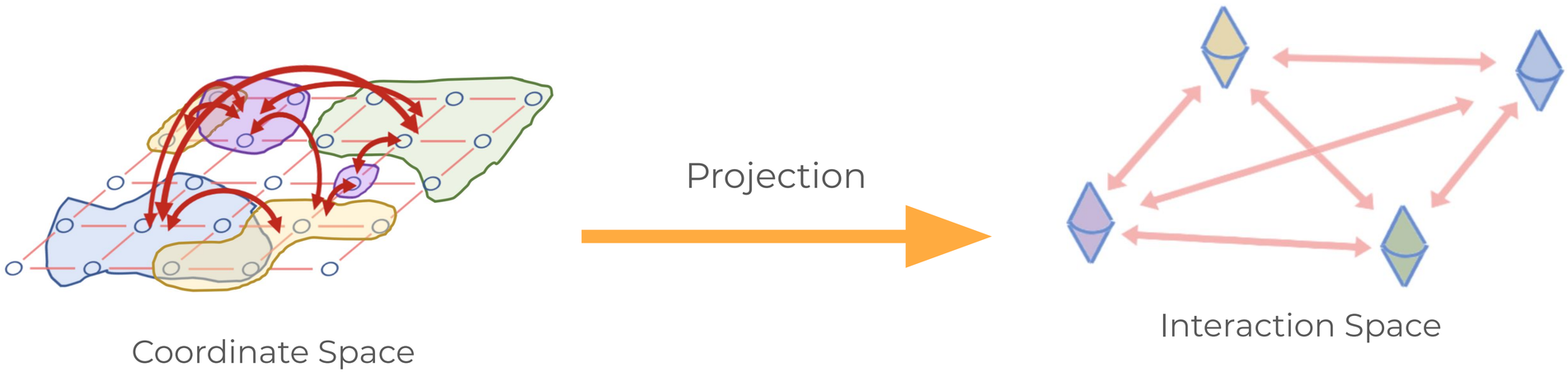
Given a set of input features $X \in \mathbb{R}^{L \times C}$, compute projections $V \in \mathbb{R}^{N \times C}$

$$\mathbf{v}_i = \mathbf{b}_i X = \sum_{\forall j} b_{ij} \mathbf{x}_j$$



From Coordinate Space to Interaction Space

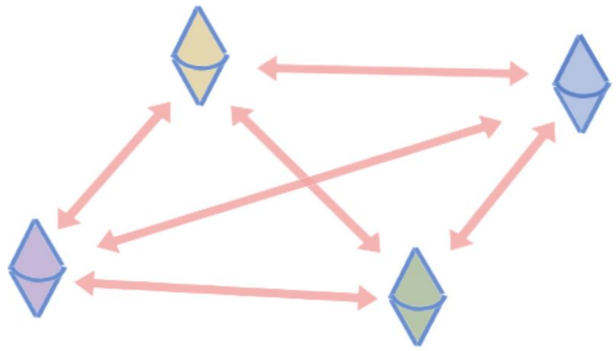
- After projection → N feature vectors $V \in \mathbb{R}^{N \times C}$
- *Relations between arbitrary regions → interactions between features*



Reasoning in Interaction Space

How to model such feature interactions?

- Treat each feature as a node in a fully-connected graph
- Learn the edge weights that correspond to interactions of features



Interaction Space

[Kipf & Welling. Semi-supervised classification with graph convolutional networks. ICLR, 2017]
[Li et al. Deeper Insights into Graph Convolutional Networks for Semi-Supervised Learning. AAAI 2018]

Reasoning in Interaction Space

How to model such feature interactions?

- Treat each feature as a node in a fully-connected graph
- Learn the edge weights that correspond to interactions of features
- Graph convolution formulation by [Kipf & Welling]:

$$\mathbf{Z} = GVW_g = ((I - A_g)V)W_g$$

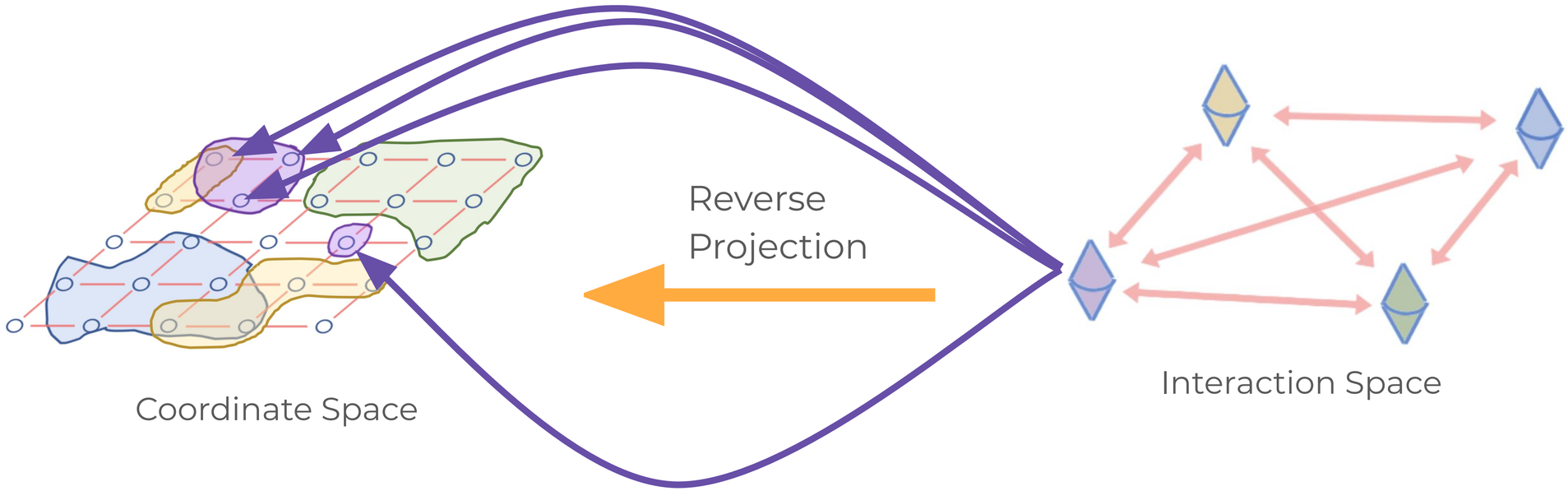
N x N (learnt)
adjacency matrix

state update

[Kipf & Welling. Semi-supervised classification with graph convolutional networks. ICLR, 2017]

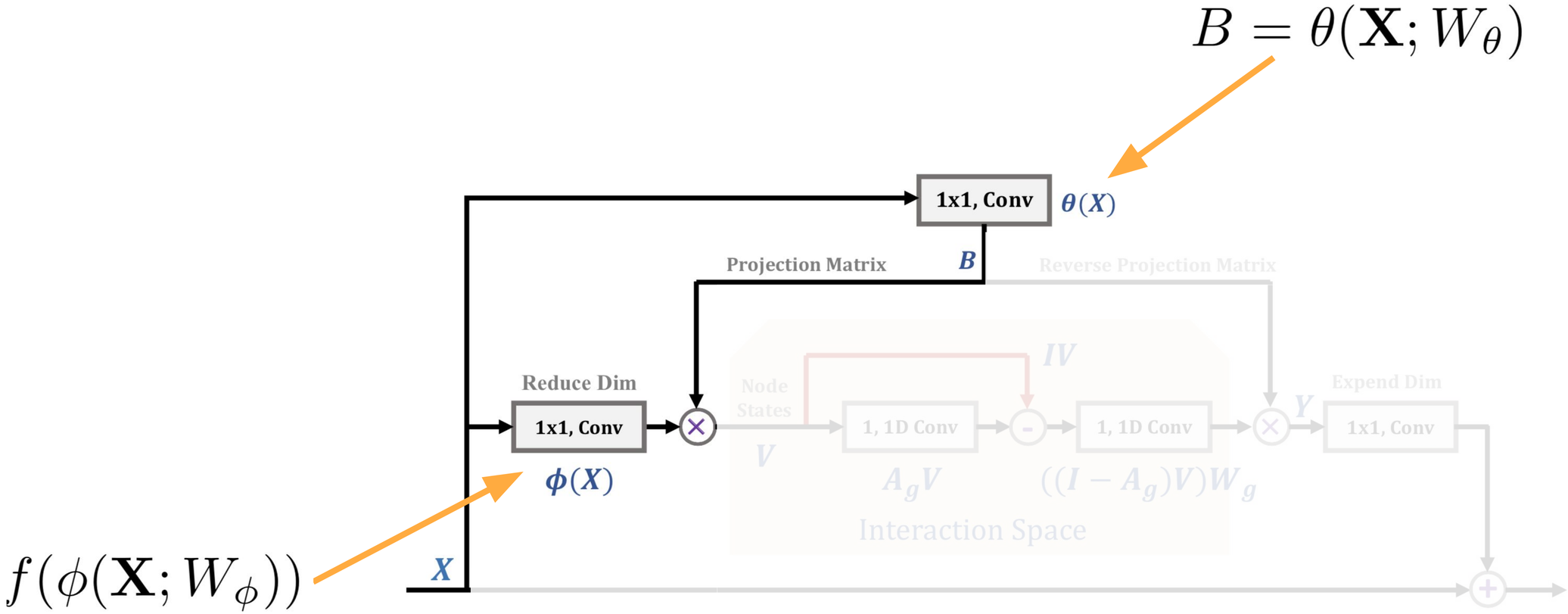
From Interaction Space to Coordinate Space

- Reverse projection: Distribute the updated states back
- Reuse projection weights $B^\top$



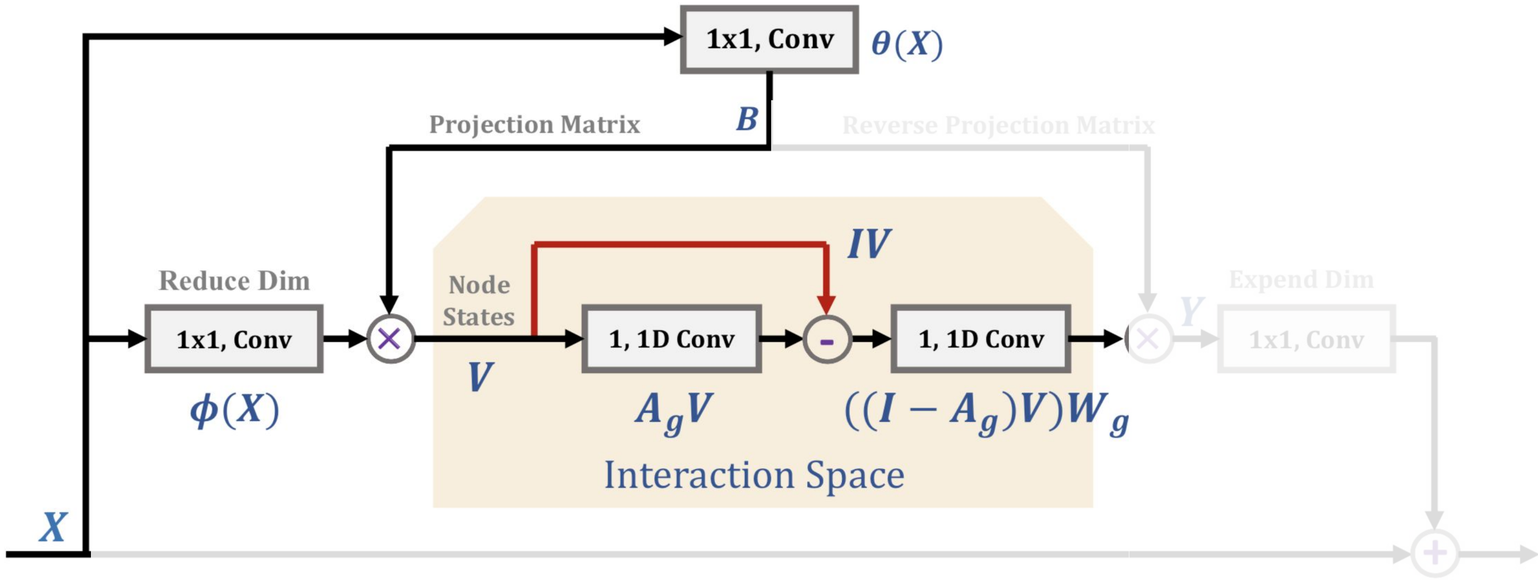
Global Reasoning (GloRe) Unit

- **Projection:** Weighted global pooling



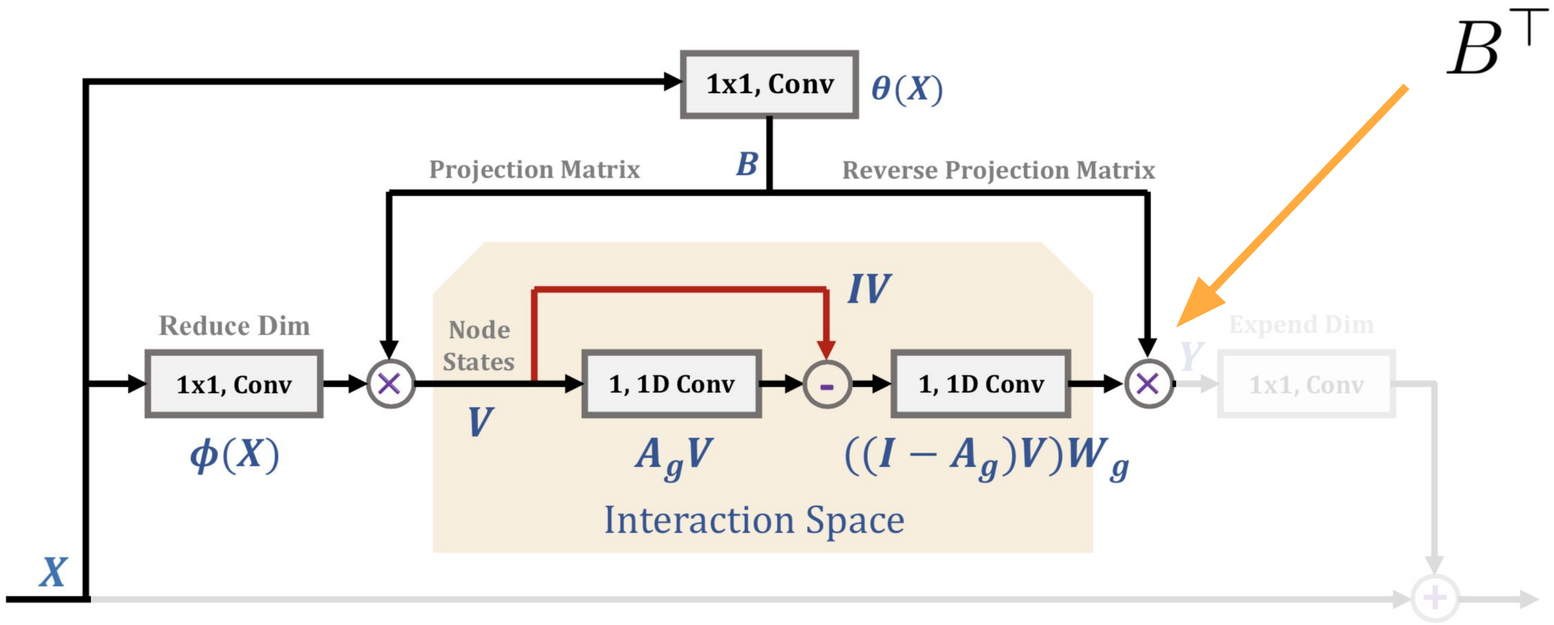
Global Reasoning (GloRe) Unit

- **Projection:** Weighted global pooling
- **Reasoning:** Graph Convolution



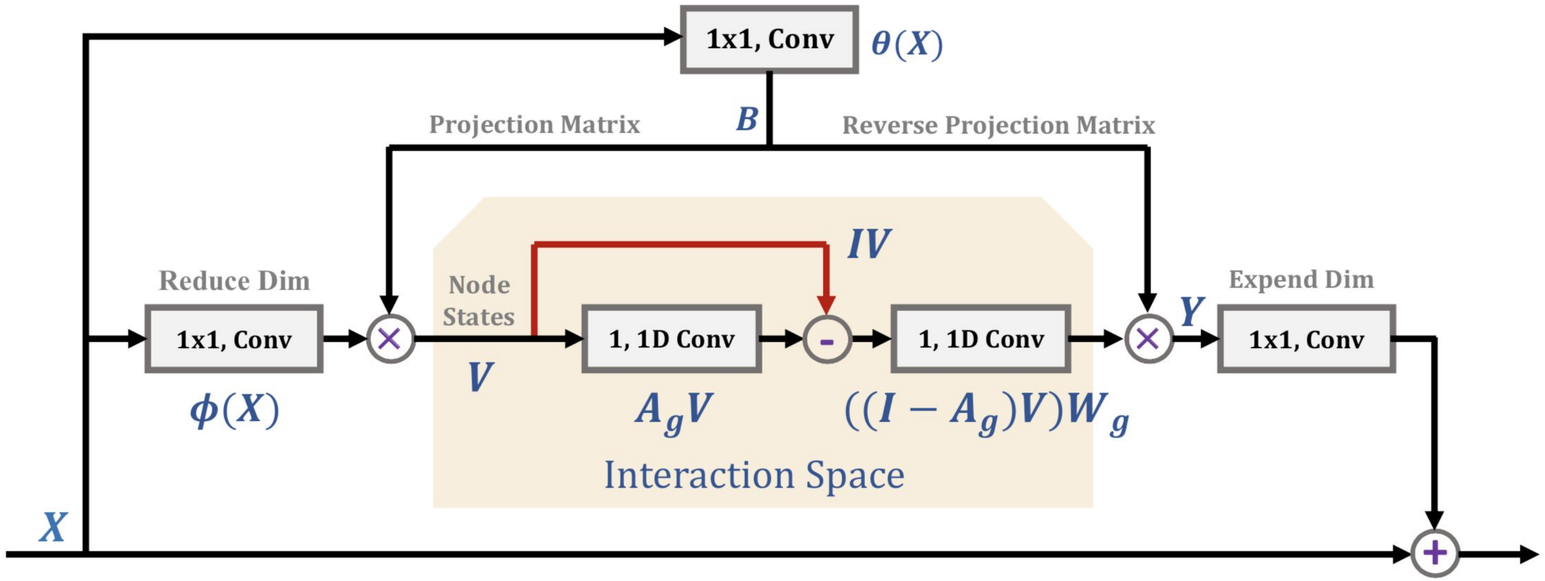
Global Reasoning (GloRe) Unit

- **Projection:** Weighted global pooling
- **Reasoning:** Graph Convolution
- **Reverse projection:** Weighted broadcasting



Global Reasoning (GloRe) Unit

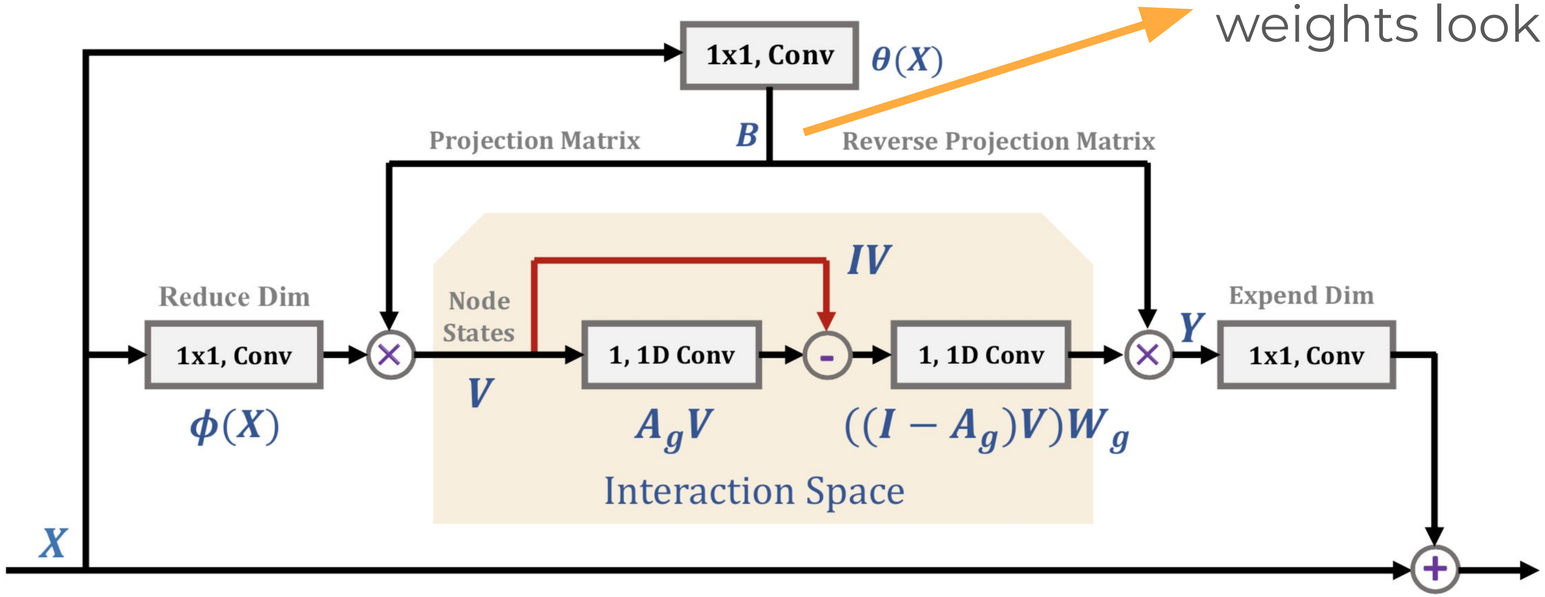
- **Projection:** Weighted global pooling
- **Reasoning:** Graph Convolution
- **Reverse projection:** Weighted broadcasting



Global Reasoning (GloRe) Unit

- **Projection:** Weighted global pooling
- **Reasoning:** Graph Convolution
- **Reverse projection:** Weighted broadcasting

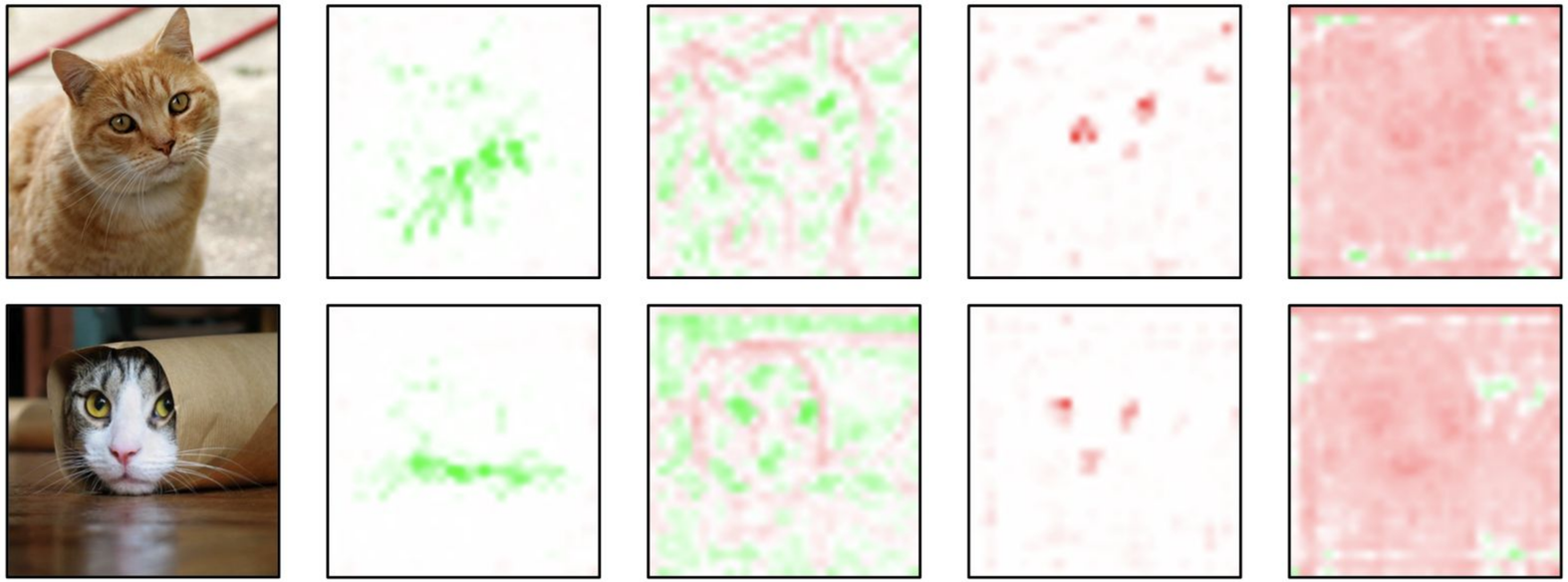
What do the learnt $\mathbb{R}^{N \times L}$ projection weights look like?



Global Reasoning (GloRe) Unit

Visualization of projection weights

What do the learnt projections $B = \theta(\mathbf{X}; W_\theta)$ look like?



Global Reasoning Networks

- The Global Reasoning (GloRe) unit is a plug-and-play residual unit that can be inserted in CNNs for different tasks

Image Classification & Action Recognition backbone CNNs

- Insert one or more units at different positions in the network

Semantic segmentation

- Insert before bottleneck

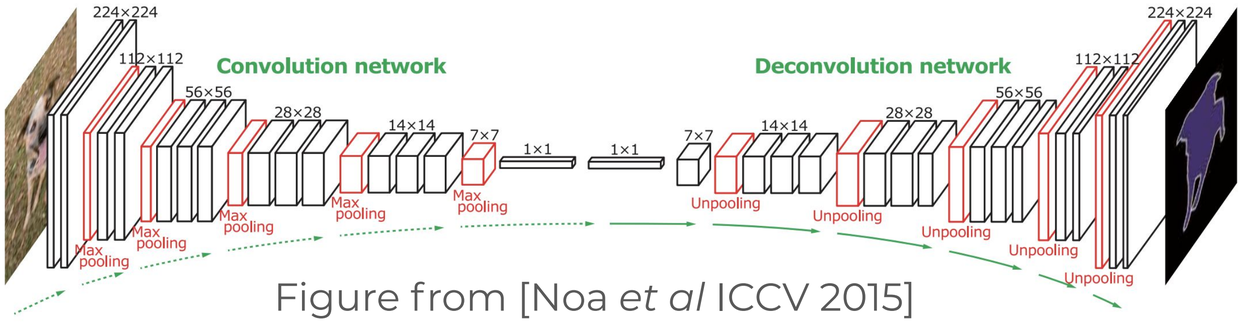
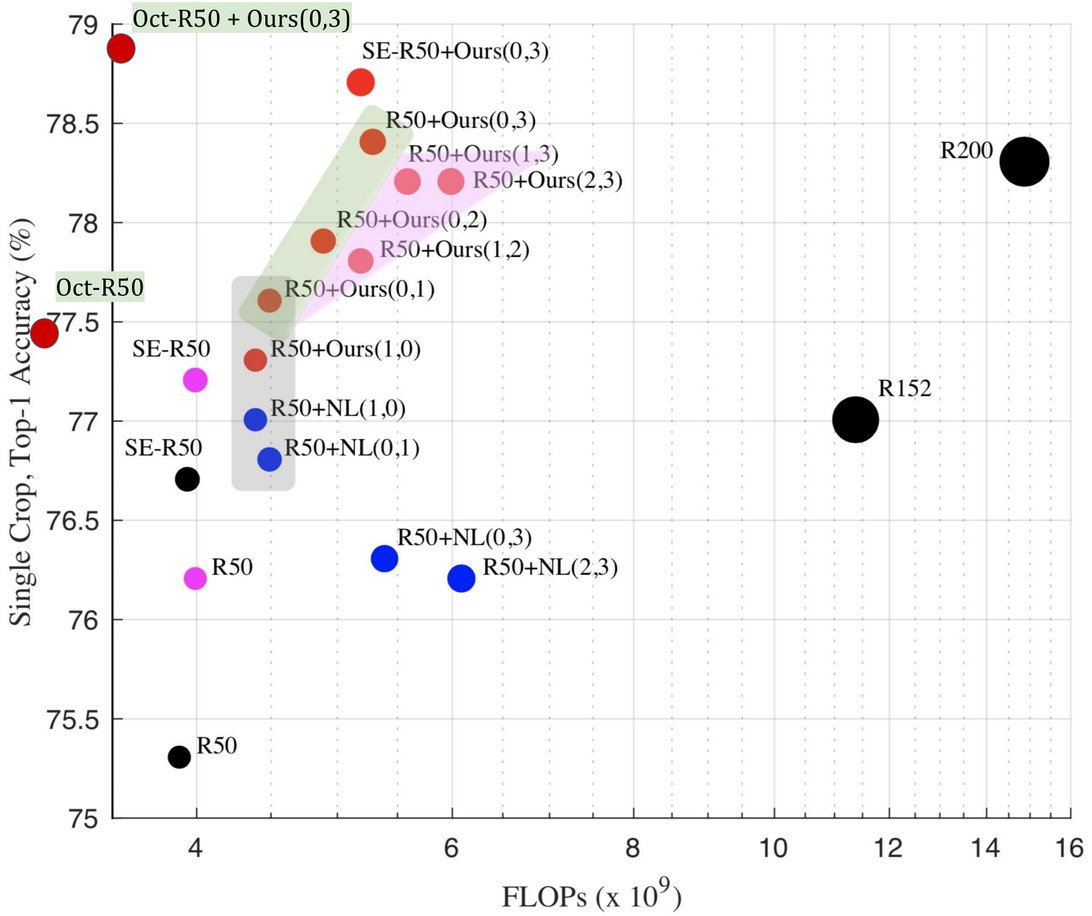


Figure from [Noa *et al* ICCV 2015]

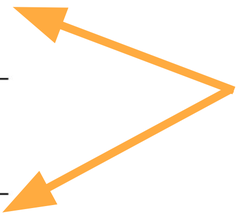
Ablations (ImageNet)

How many blocks to add and where?



Results on ImageNet

Method		Res3	Res4	GFLOPs	#Params	Top-1
ResNet50 [16]	Baseline			4.0	25.6M	76.2%
	GloRe (Ours)		+3	5.2	30.5M	**78.4%**
	GloRe (Ours)	+2	+3	6.0	31.4M	78.2%
SE-ResNet50 [18]	Baseline			4.0	28.1M	77.2%
	GloRe (Ours)		+3	5.2	33.0M	**78.7%**
ResNet200 [16]	Baseline			15.0	64.6M	78.3%
	GloRe (Ours)		+3	16.2	69.7M	79.4%
	GloRe (Ours)	+2	+3	16.9	70.6M	**79.7%**
ResNeXt101 [33] (32 × 4)	Baseline			8.0	44.3M	78.8%
	GloRe (Ours)	+2	+3	9.9	50.3M	**79.8%**
DPN-98 [9]	Baseline			11.7	61.7M	79.8%
	GloRe (Ours)	+2	+3	13.6	67.7M	**80.2%**
DPN-131 [9]	Baseline			16.0	79.5M	80.1%
	GloRe (Ours)	+2	+3	17.9	85.5M	**80.3%**

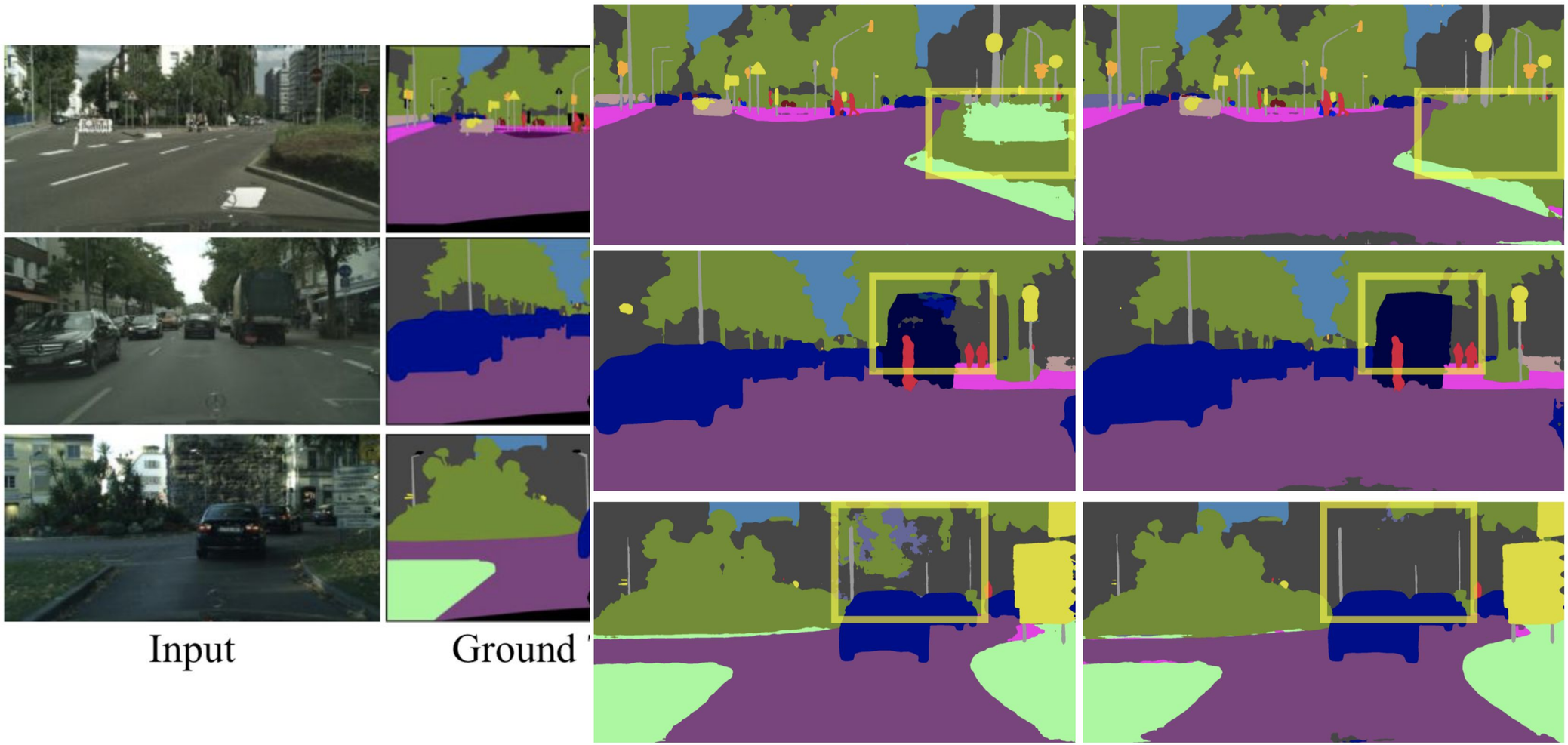


Results on Semantic Segmentation (Cityscapes)

Results on the test set, no “coarse” annotations used during training. FCN + multi-grid [Chen *et al* 2017]

Method	Backbone	IoU cla.	iIoU cla.	IoU cat.	iIoU cat.
DeepLab-v2 [4]	ResNet101	70.4%	42.6%	86.4%	67.7%
PSPNet [36]	ResNet101	78.4%	56.7%	90.6%	78.6%
PSANet [37]	ResNet101	80.1%			
DenseASPP [35]	ResNet101	80.6%			
FCN + 1 GloRe unit	ResNet50	79.5%	60.3%	91.3%	81.5%
FCN + 1 GloRe unit	ResNet101	**80.9%**	**62.2%**	**91.5%**	**82.1%**

Results on Semantic Segmentation (Cityscapes)



Results on Action Recognition (Kinetics)

Results on the validation set, RGB only (no Optical Flow)

Method	Backbone	Frames	FLOPs	Clip Top-1	Video Top-1
I3D-RGB [2]	Inception-v1	64	107.9 G	–	71.1%
R(2+1)D-RGB [29]	ResNet-xx	32	152.4 G	–	72.0%
MF-Net [8]	MF-Net	16	11.1 G	–	72.8%
S3D-G [34]	Inception-v1	64	71.4 G	–	74.7%
NL-Nets [31]	ResNet-50	8	30.5 G	67.12%	74.57%
GloRe (Ours)	ResNet-50	8	28.9 G	68.02%	**75.12%**
NL-Nets [31]	ResNet-101	8	56.1 G	68.48%	75.69 %
GloRe (Ours)	ResNet-101	8	54.5 G	68.78%	**76.09%**

Conclusions

The Global Reasoning (GloRe) unit

- A plug-and-play residual unit that enables efficient reasoning between arbitrary regions
- Is highly efficient (smaller computational cost than a self-attention)
- Consistent gains on multiple tasks
- Complementary to the Octave Convolution!
- Models and code are open-sourced

  ... try it!

Overview

Motivation

Challenges for Representation Learning

Reducing computation: **Octave Convolutions**

Non-local reasoning: **Global Reasoning Networks**

Higher-level understanding: Grounded Video Description

Summary

A Vision For the Future

Grounded Video Description

Luowei Zhou[1,2], Yannis Kalantidis[1], Xinlei Chen[1], Jason J. Corso[2], Marcus Rohrbach[1]
[1] Facebook AI, [2] University of Michigan
github.com/facebookresearch/grounded-video-description

[CVPR 2019]

Automatic Video Description

Machine: A man is seated on a bed.

Machine: A group of people are in a river.

Automatic Video Description

Machine: A man is seated on a bed.
Human: We see a man playing a **saxophone** in front of **microphones**.

Machine: A group of people are in a river.
Human: Several people are on a **raft** in the water.

Automatic Video Description

Machine: A man is seated on a **bed**.
Human: We see a man playing a **saxophone** in front of **microphones**.

Machine: A group of people are in a river.
Human: Several people are on a **raft** in the water.

Automatic Video Description

The machine generated descriptions are not always visually **grounded**

A man in a striped shirt is playing the piano while people watch him.

Automatic Video Description

The machine generated descriptions are not always visually **grounded**

Automatic Video Description

The machine generated descriptions are not always visually **grounded**

Why is this a problem?

- **hallucinations**
- reinforces **biases** of the training set

Automatic Video Description

The machine generated descriptions are not always visually **grounded**

Can explicit grounding benefit description generation?

Problem: No video dataset with descriptions and grounding

Video Description Datasets (w/ grounding)

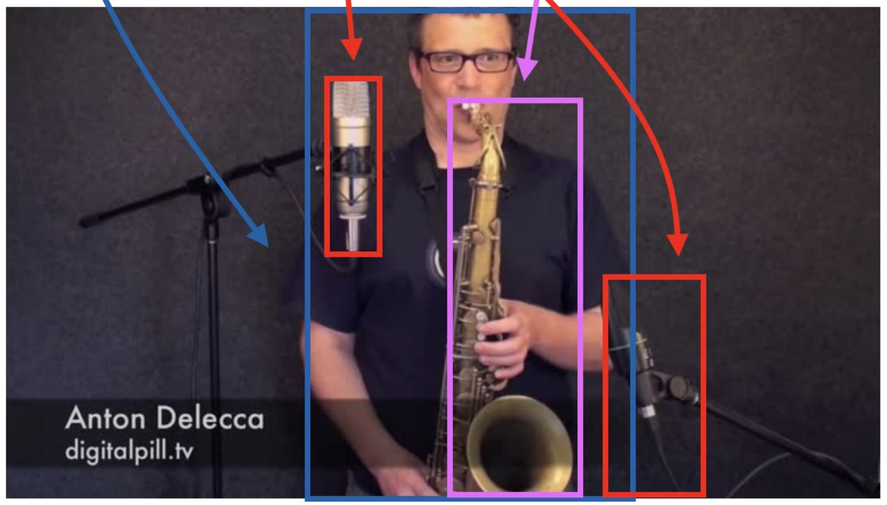
Dataset	# Video	# Sent	# Obj	# BBoxes
MPII-MD	1k	1k	4	2.6k
YouCook2	2k	15k	67	135k
ActivityNet Humans	5.3k	30k	1	63k
ActivityNet-Entities (ours)	**15k**	**52k**	**432**	**158k**

Based on ActivityNet Captions [Krishna *et al.* ICCV 2017] dataset

[Rohrbach et al. CVPR 2017, Zhou et al. BMVC 2018, Yamaguchi et al. ICCV 17, Krishna et al. ICCV 2017]

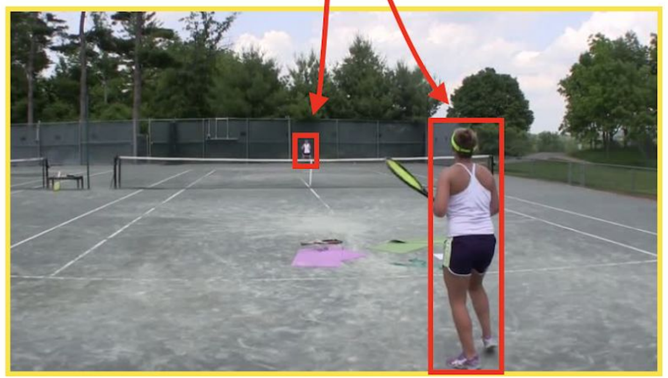
ActivityNet-Entities dataset

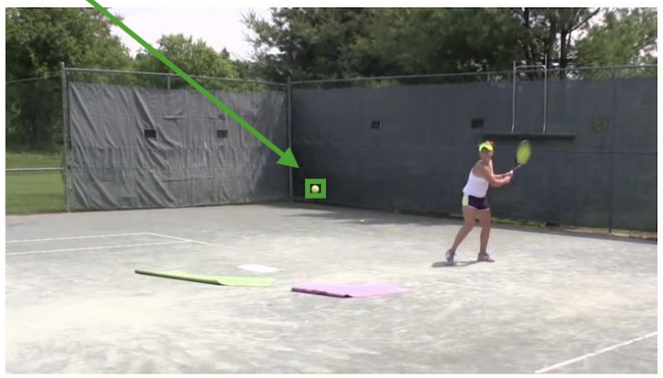
We see a man playing a saxophone in front of microphones.



ActivityNet-Entities dataset

Two women are on a tennis court, showing the technique to posing and hitting the ball.





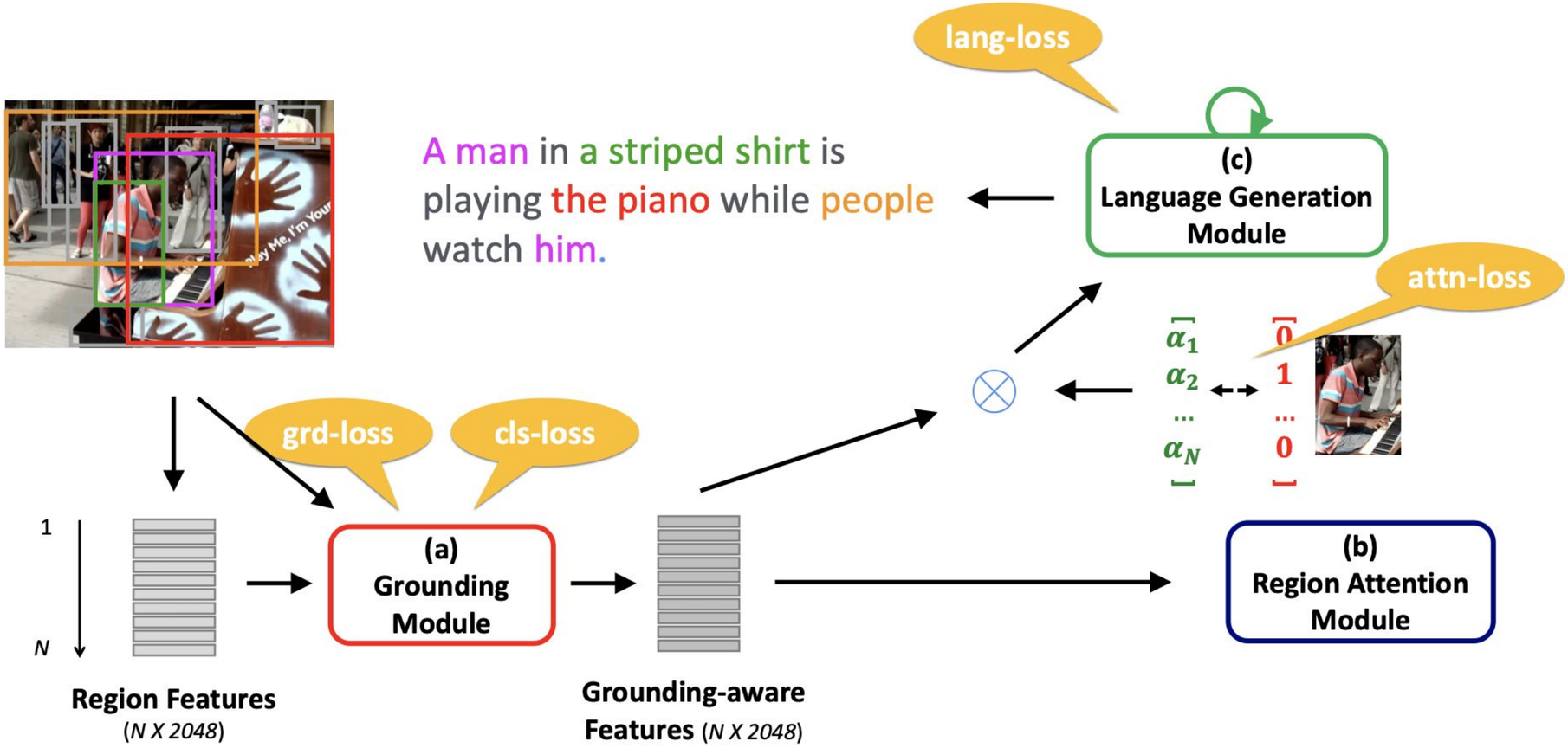
How can we exploit this information?

Grounded Video Description (GVD) model

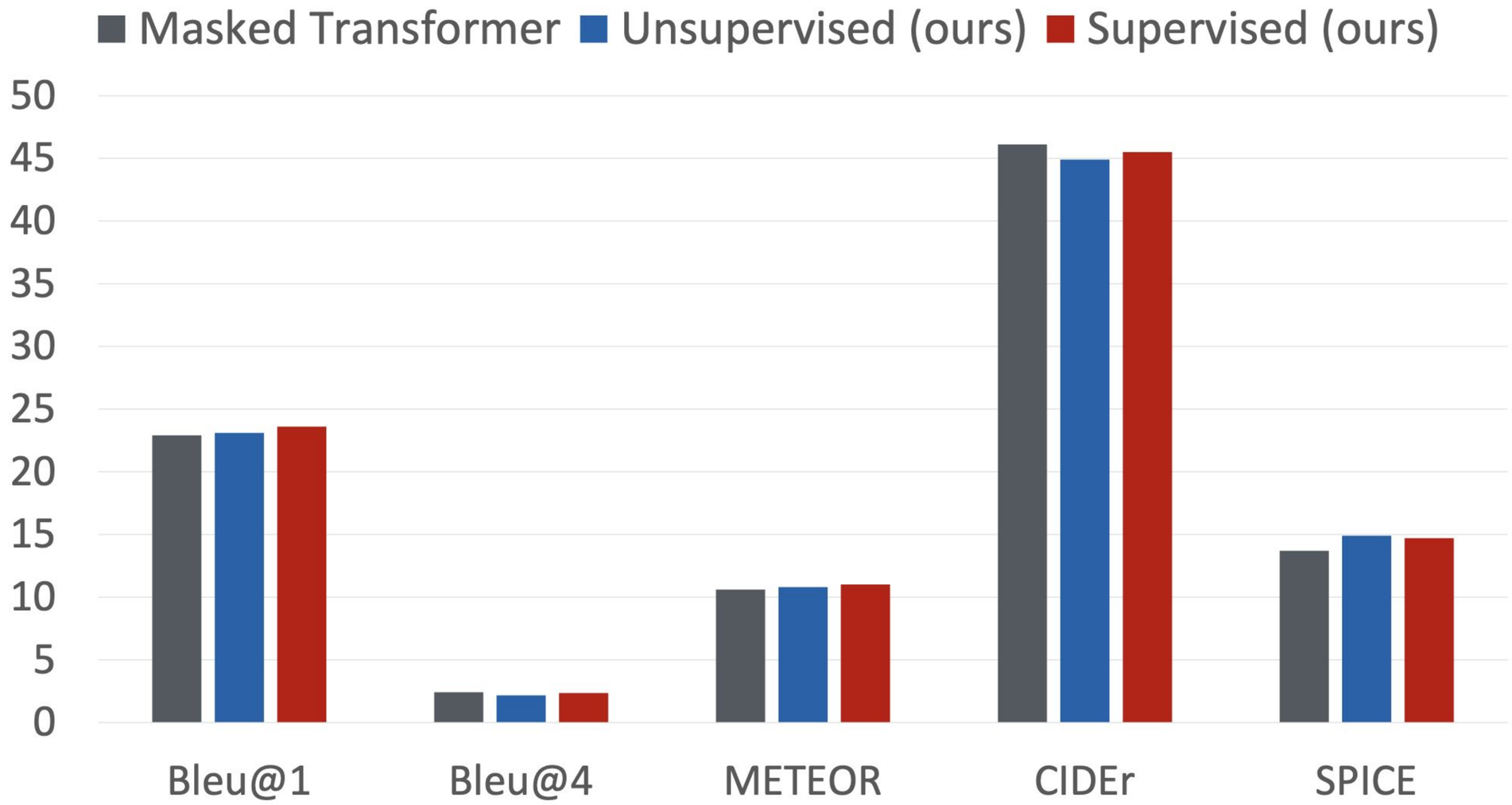
Regularize standard caption generation with three proxy tasks:

- Supervised attention
- Supervised grounding
- Region classification

Grounded Video Description (GVD) model



Results – Description Quality (ANet-Entities)



Results – Localization Accuracy (ANet-Entities)

Localize
object words
in the video

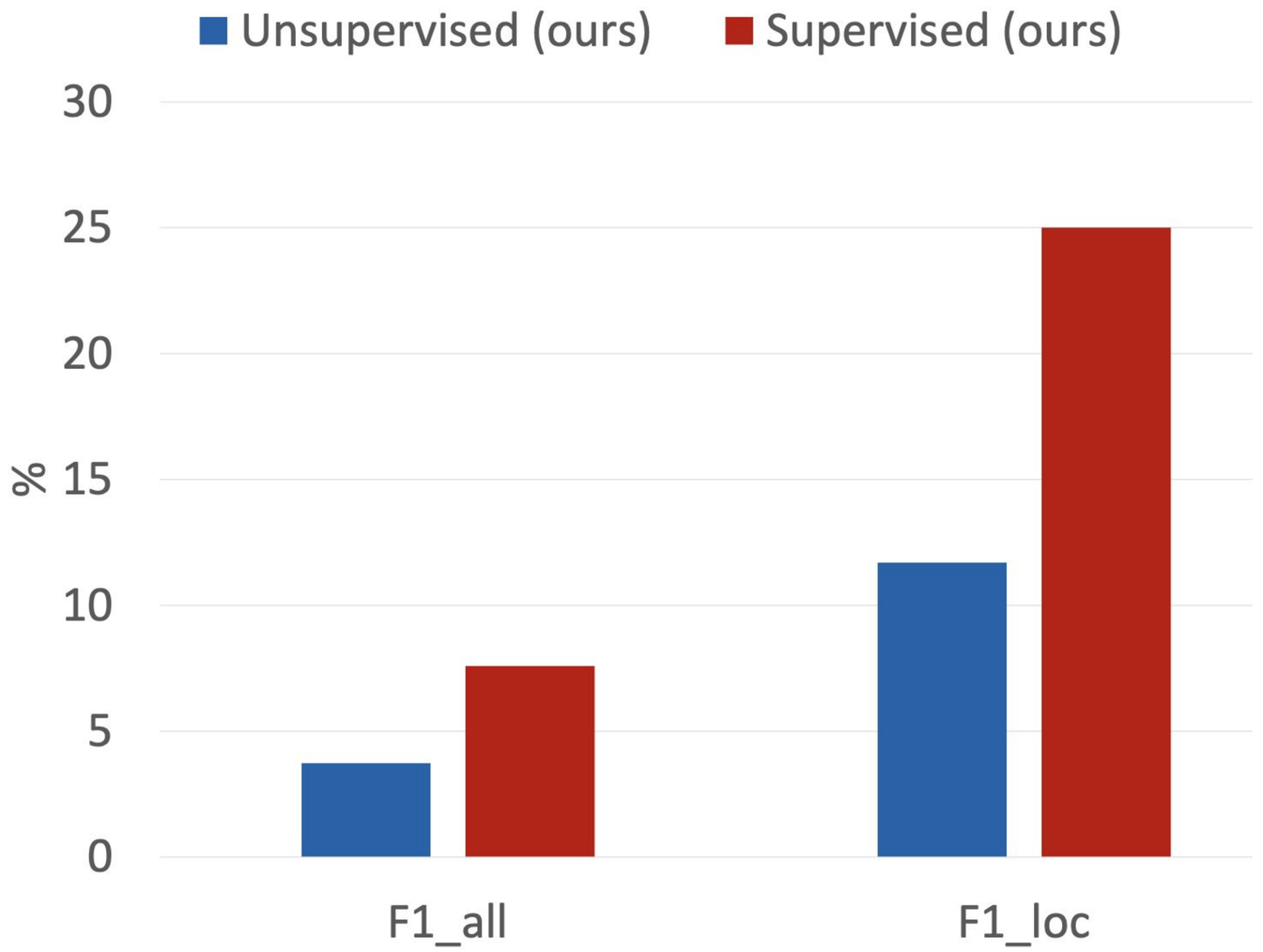
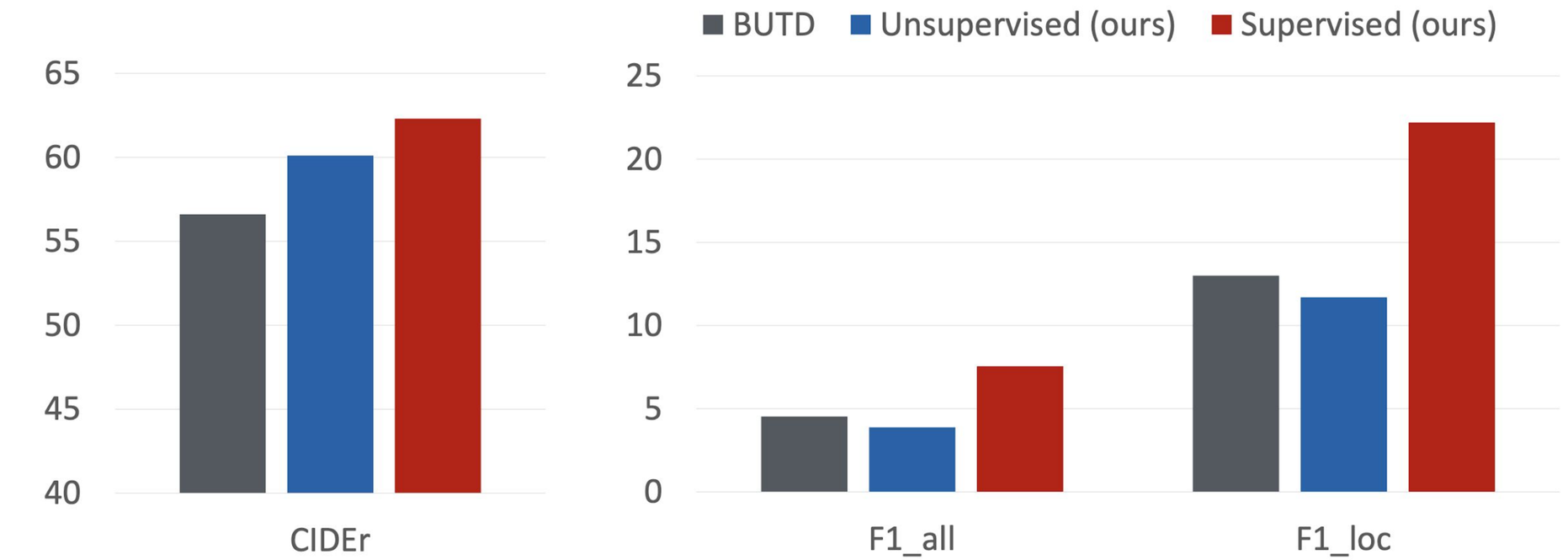


Image description (Flickr30k Entities)



BUTD: [Peter Anderson et al. CVPR 2018]

Qualitative Results

A group of people are in a raft down a river.

w/o grounding supervision: A group of people are in a river.

Qualitative Results

A man is seen standing in a room speaking to the camera while holding a bike .

w/o grounding supervision: A man is standing in a gym .

Conclusions

ActivityNet-Entities Dataset

- Grounding supervision for model training
- A dataset for evaluating grounding performance

Grounded Video Description model

- State of the art, and more interpretable descriptions

Code & Dataset Page: github.com/facebookresearch/ActivityNet-Entities

Overview

Motivation

Challenges for Representation Learning

Reducing computation: **Octave Convolutions**

Non-local reasoning: **Global Reasoning Networks**

Higher-level understanding: **Grounded Video Description**

Summary

Summary

Challenges for Representation Learning

Reducing computation

Non-local & spatio-temporal understanding

Higher-level & multi-modal reasoning

Octave Convolution

- PnP replacement for the conv
- import OctConv as Conv

Global Reasoning Networks

- Reasoning between arbitrary regions
- Complementary to the Octave Convolution

Grounded Video Descriptions

- ActivityNet-Entities: New dataset with grounding annotations
- GVD: New model with state-of-the-art performance and more grounded captions

Code on github

- Octave Convolutions

  facebookresearch/OctConv (++ 3rd party)

- Global Reasoning Networks

  facebookresearch/GloRe

- Grounded video description (dataset and code)

  facebookresearch/grounded-video-description

Thank you!

Questions?

References (Personal publications mentioned in these slides)

- B. Kang, S. Xie, M. Rohrbach, Z. Yan, A. Gordo, J. Feng, **Y. Kalantidis**. Decoupling Representation and Classifier for Long-Tailed Recognition. **arXiv:1910.09217, 2019.**
- C.Y. Ma, **Y. Kalantidis**, G. AlRegib, P. Vajda, M. Rohrbach, Z. Kira. Learning to Generate Grounded Image Captions without Localization Supervision. **arXiv:1906.00283, 2019**
- Y. Chen, H. Fan, B. Xu, Z. Yan, **Y. Kalantidis**, M. Rohrbach, Z. Yan, Y. Shuicheng, J. Feng. Drop an Octave: Reducing Spatial Redundancy in Convolutional Neural Networks with Octave Convolution. **ICCV, 2019**.
- Y. Chen, M. Rohrbach, Z. Yan, Y. Shuicheng, J. Feng, **Y. Kalantidis**. Graph-Based Global Reasoning Networks. **CVPR, 2019**.
- B. Xiong, **Y. Kalantidis**, D. Ghadiyaram, K. Grauman. Less is More: Learning Highlight Detection from Video Duration. **CVPR, 2019**.
- L. Zhou, **Y. Kalantidis**, X. Chen, J. Corso, M. Rohrbach. Grounded Video Description. **CVPR, 2019.**
- Z. Shou, Z. Yan, **Y. Kalantidis**, L. Sevilla-Lara, M. Rohrbach, SF. Chang. DMC-Net: Generating Discriminative Motion Cues for Compressed Video Action Recognition. **CVPR, 2019**.
- J. Zhang, **Y. Kalantidis**, M. Rohrbach, M. Paluri A. Elgammal, M. Elhoseiny. Large-Scale Visual Relationship Understanding. **AAAI, 2019.**
- Y. Chen, **Y. Kalantidis**, J. Li, Y. Shuicheng, J. Feng. Double Attention Networks. **NeurIPS, 2018.**
- Y. Chen, **Y. Kalantidis**, J. Li, Y. Shuicheng, J. Feng. Multi-Fiber Networks. **ECCV, 2018.**

Personal website: https://www.skamalas.com