ECCV'20 ONLINE
23-28 AUGUST 2020

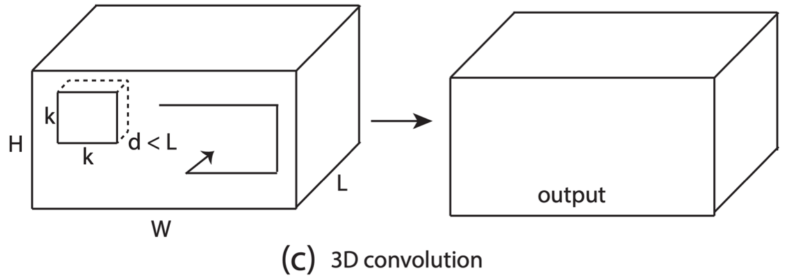
AttentionNAS: Spatiotemporal Attention Cell Search for Video Classification

Xiaofang Wang, Xuehan Xiong, Maxim Neumann, AJ Piergiovanni,
Michael S. Ryoo, Anelia Angelova, Kris M. Kitani, Wei Hua

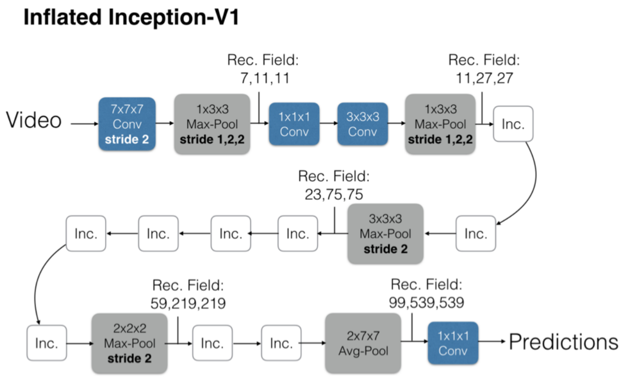
Google

Carnegie Mellon University

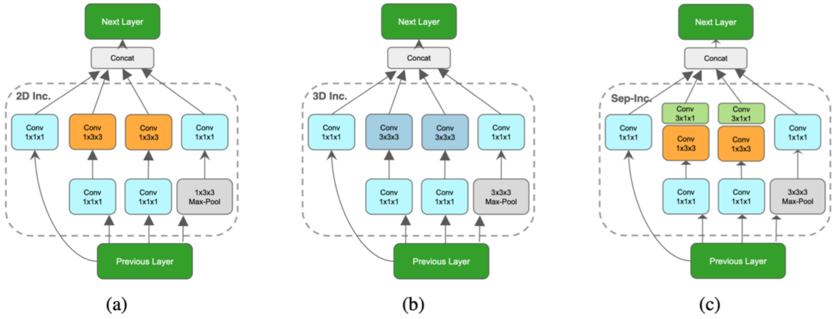
Convolutional networks are dominant



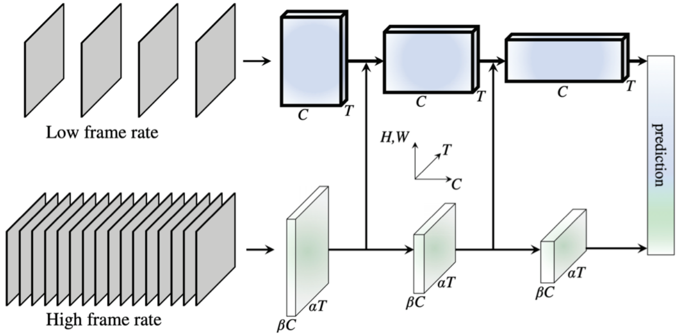
C3D [ICCV 2015]



I3D [CVPR 2017]



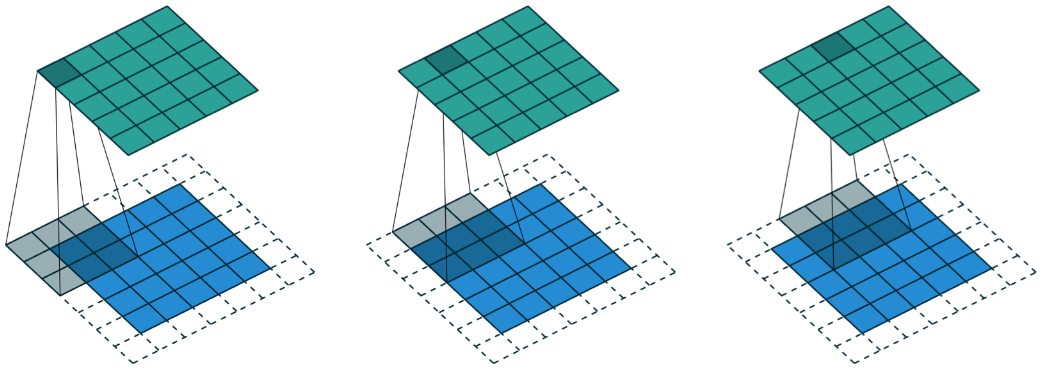
S3D [ECCV 2018]



SlowFast [ICCV 2019]

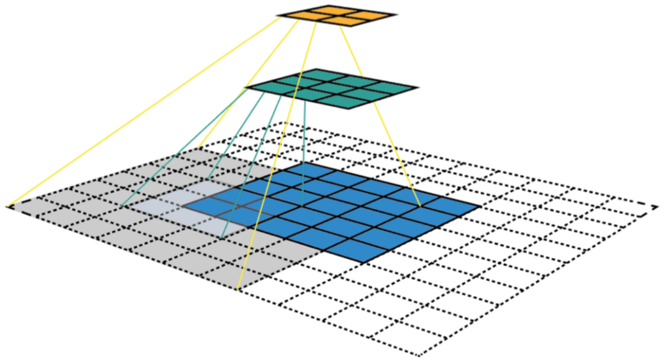
What's missing from convolution?

- Where to focus in images/videos



The same convolutional kernel
is applied at every position.

- Long-range dependencies

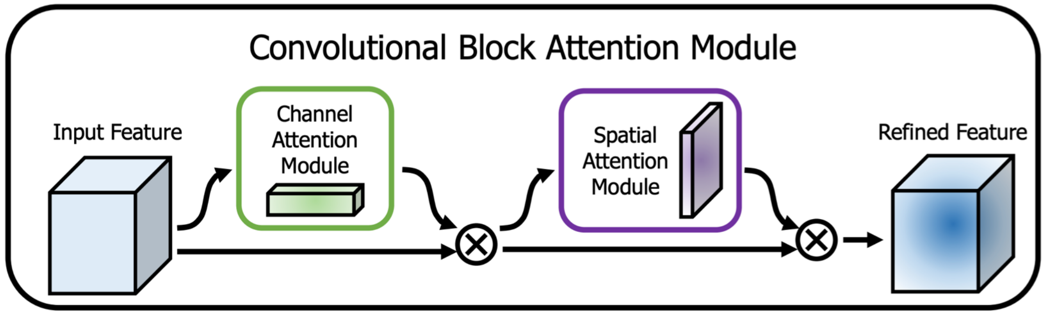


Long-range dependencies are
modeled by large receptive fields.

Photo credit: [Convolution arithmetic] [Receptive field arithmetic]

Attention is complementary to convolution

- Map-based Attention

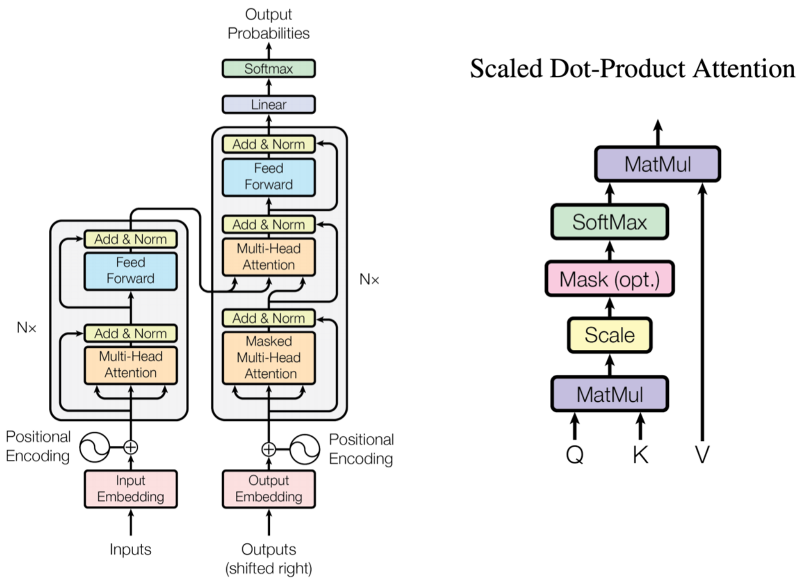


$$\mathbf{F}'' = \mathbf{M_s}(\mathbf{F}') \otimes \mathbf{F}'$$

CBAM [ECCV 2018]

Where to focus: learn a **pointwise** weighting factor for each position

- Dot-product Attention

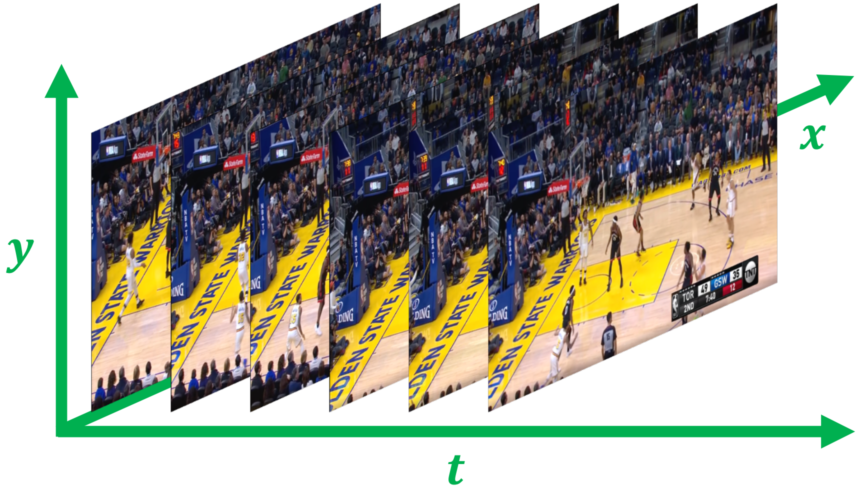


Attention is All You Need [NeurIPS 2017]

Long-range dependencies: compute **pairwise** similarity between all the positions

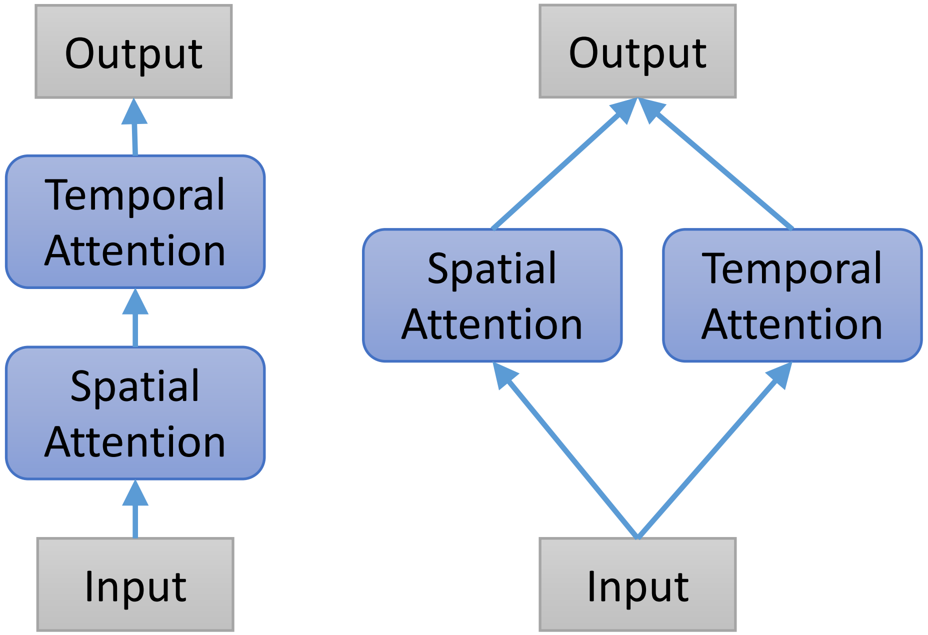
Challenge: Many design choices need to be determined to apply attention to videos

- What is the right dimension to apply attention to videos?



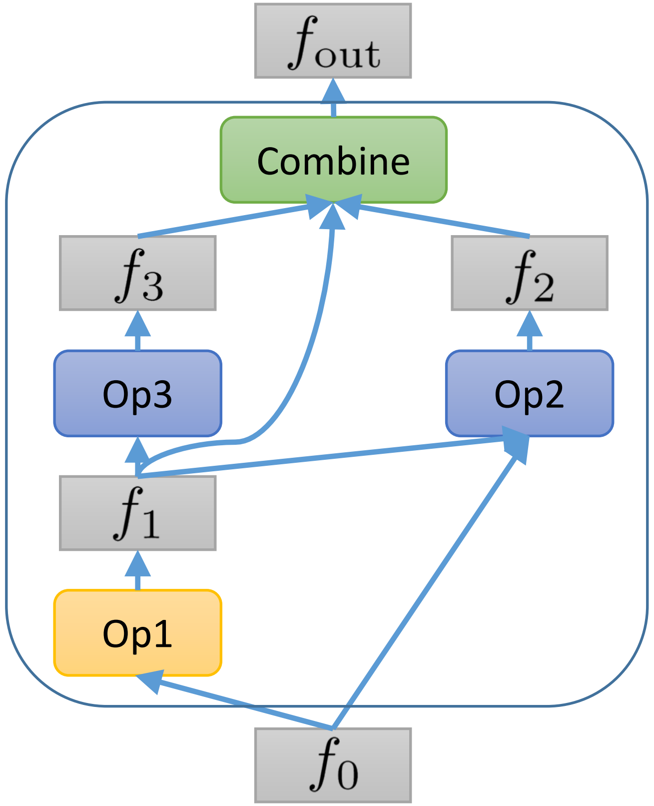
Three dimensions in video data:
spatial, temporal or spatiotemporal?

- How to compose multiple attention operations?

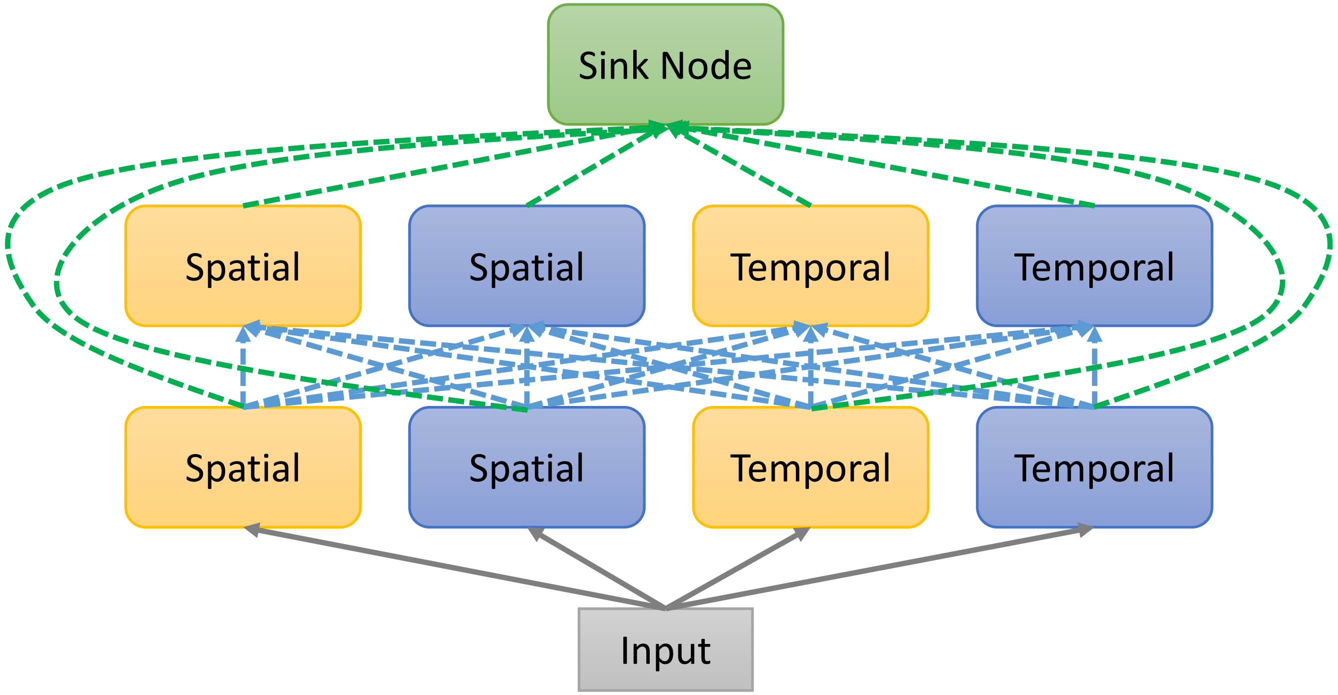


Sequential, parallel, or others?

Proposal: Automatically search for attention cells in a *data-driven* manner

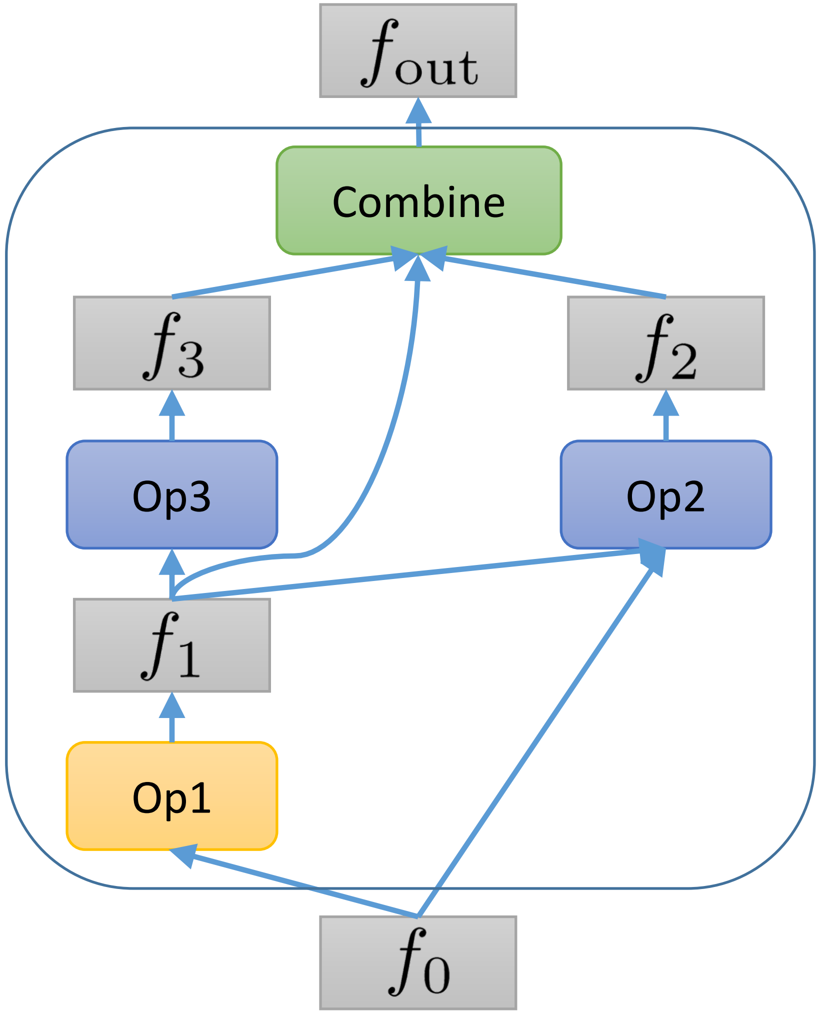


Novel Attention Cell Search Space



Efficient Differentiable Search Method

Attention Cell Search Space



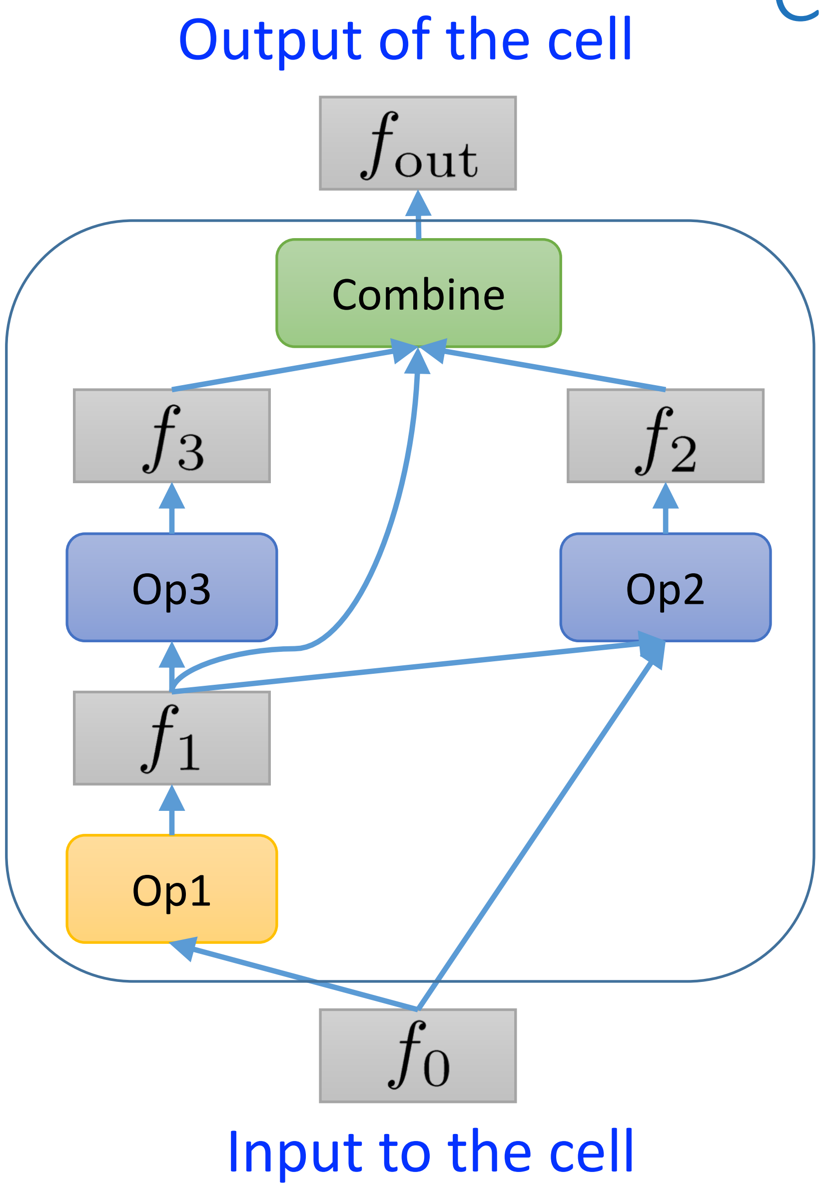
Attention Cell

- Composed of multiple attention operations
- Input shape == output shape; can be inserted anywhere in existing backbones

Search Space

- **Cell Level Search Space:** Connectivity between the operations within the cell
- **Operation Level Search Space:** Choices to instantiate an individual attention operation

Cell Level Search Space



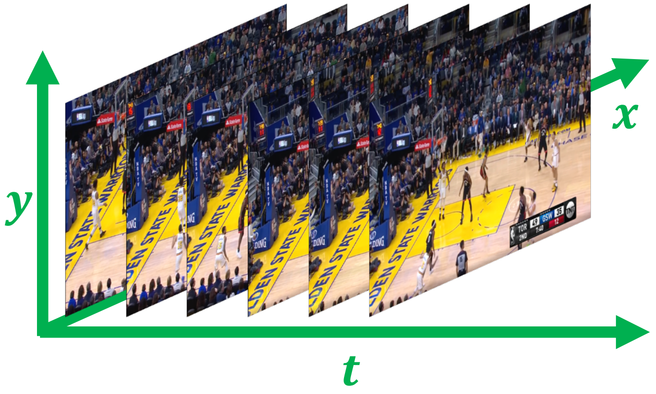
Select input to each operation

- Input to the 1^{st} operation is fixed to f_0
- Input to the k^{th} operation is a weighted sum of selected feature maps from $\{f_0, f_1, \ldots, f_{k-1}\}$

Combine $\{f_1, f_2, \ldots, f_K\}$

- Concatenate channels + CONV

Operation Level Search Space



1. Spatial 2. Temporal 3. Spatiotemporal

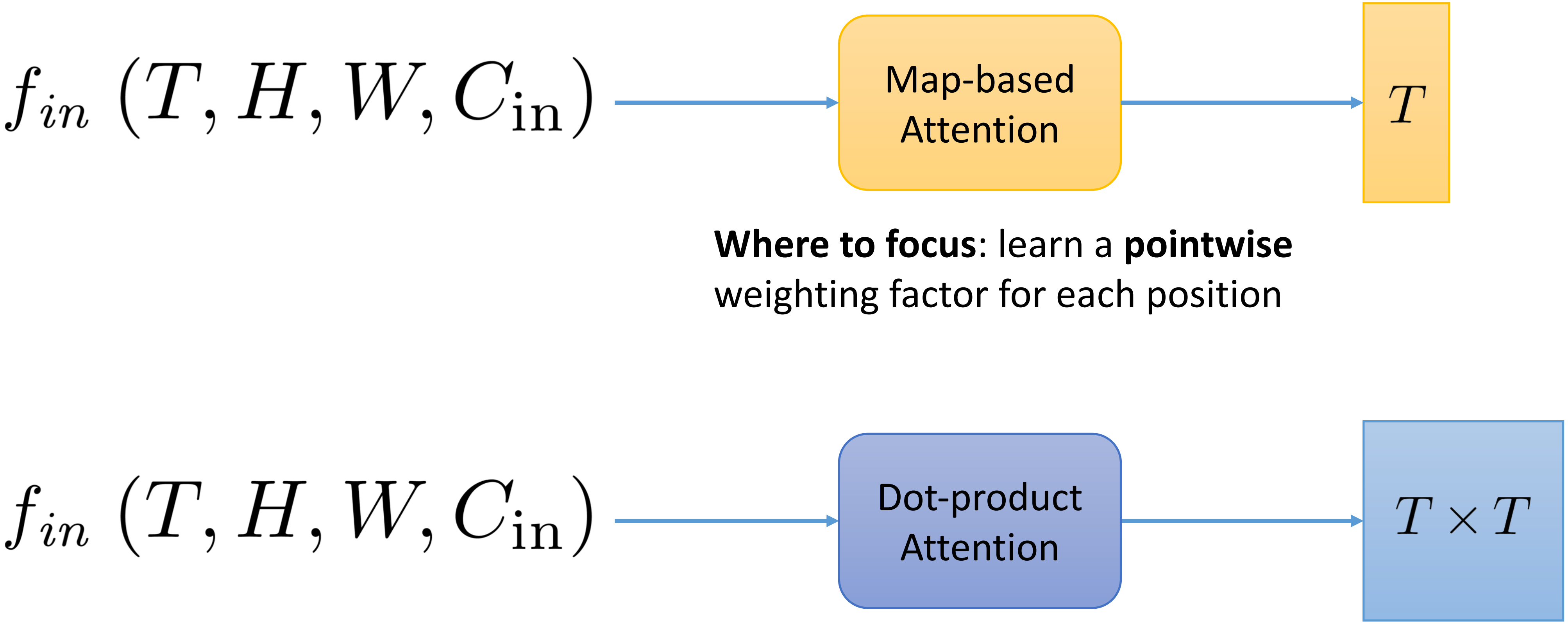
Attention Dimension

Map-based Attention

Dot-product Attention

Attention Operation Type

Map-based Attention and Dot-product Attention

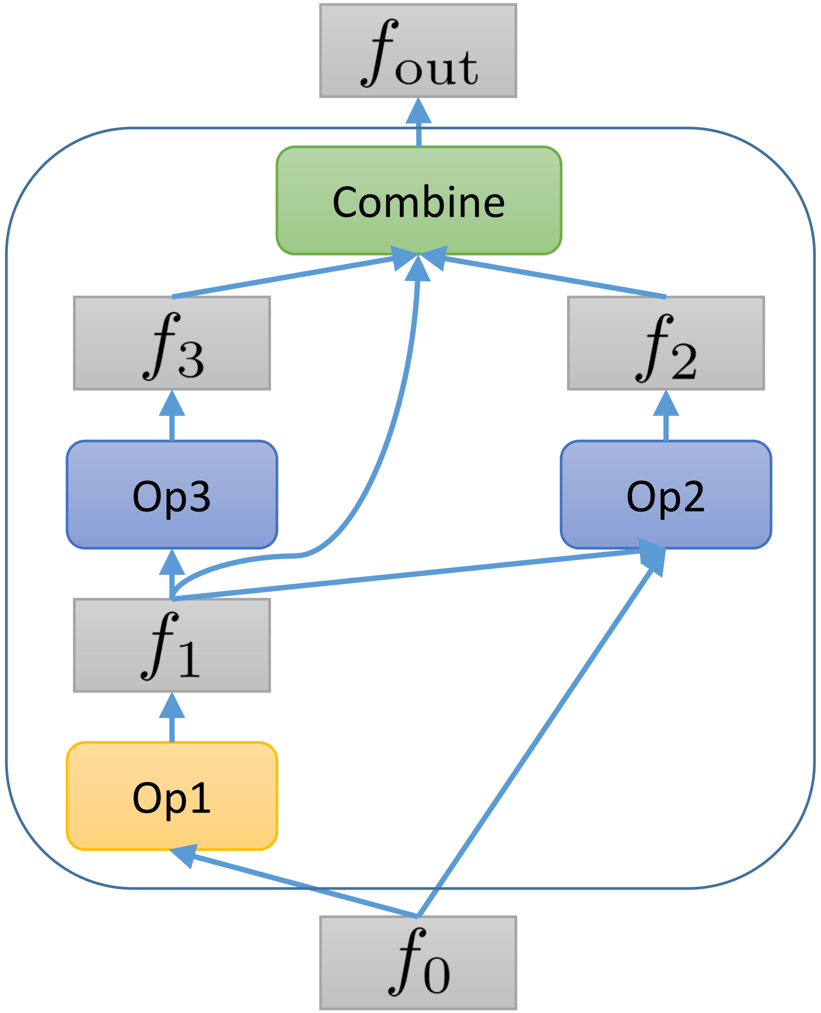


Where to focus: learn a **pointwise** weighting factor for each position

Long-range dependencies: compute **pairwise** similarity between all the positions

Assume attention dimension = temporal

Search Space Summary



- Spatial
- Temporal
- Spatiotemporal

Attention Dimension

- Map-based attention
- Dot-product attention

Attention Operation Type

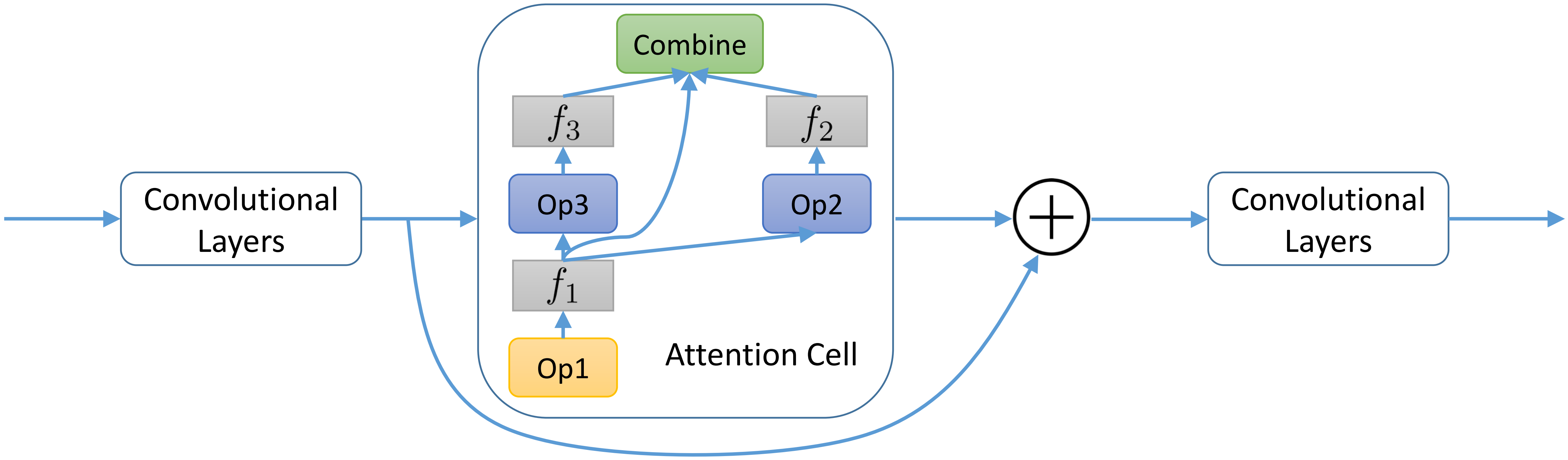
- None
- ReLU
- Softmax
- Sigmoid

Activation Function

- Input to each operation

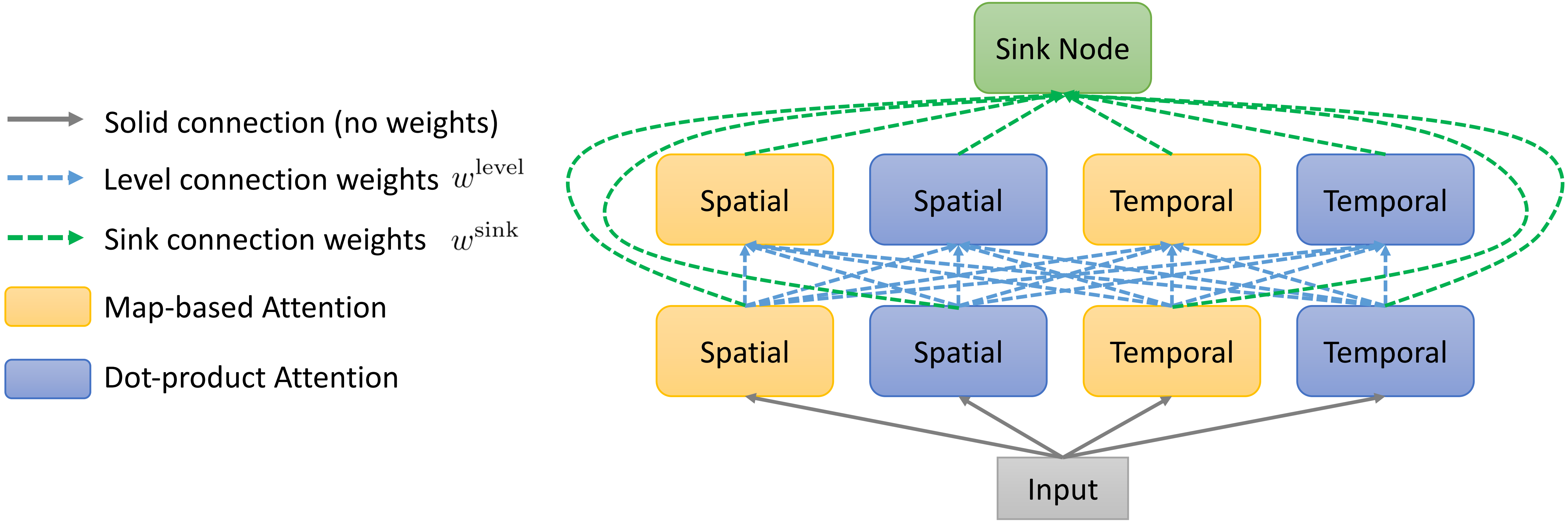
Connectivity between Operations

Insert Attention Cells into Backbone Networks



Differentiable Formulation of Search Space

- **Search algorithm**: differentiable architecture search
- **Search cost**: equals to the cost of training one network



Supergraph and Connection Weights

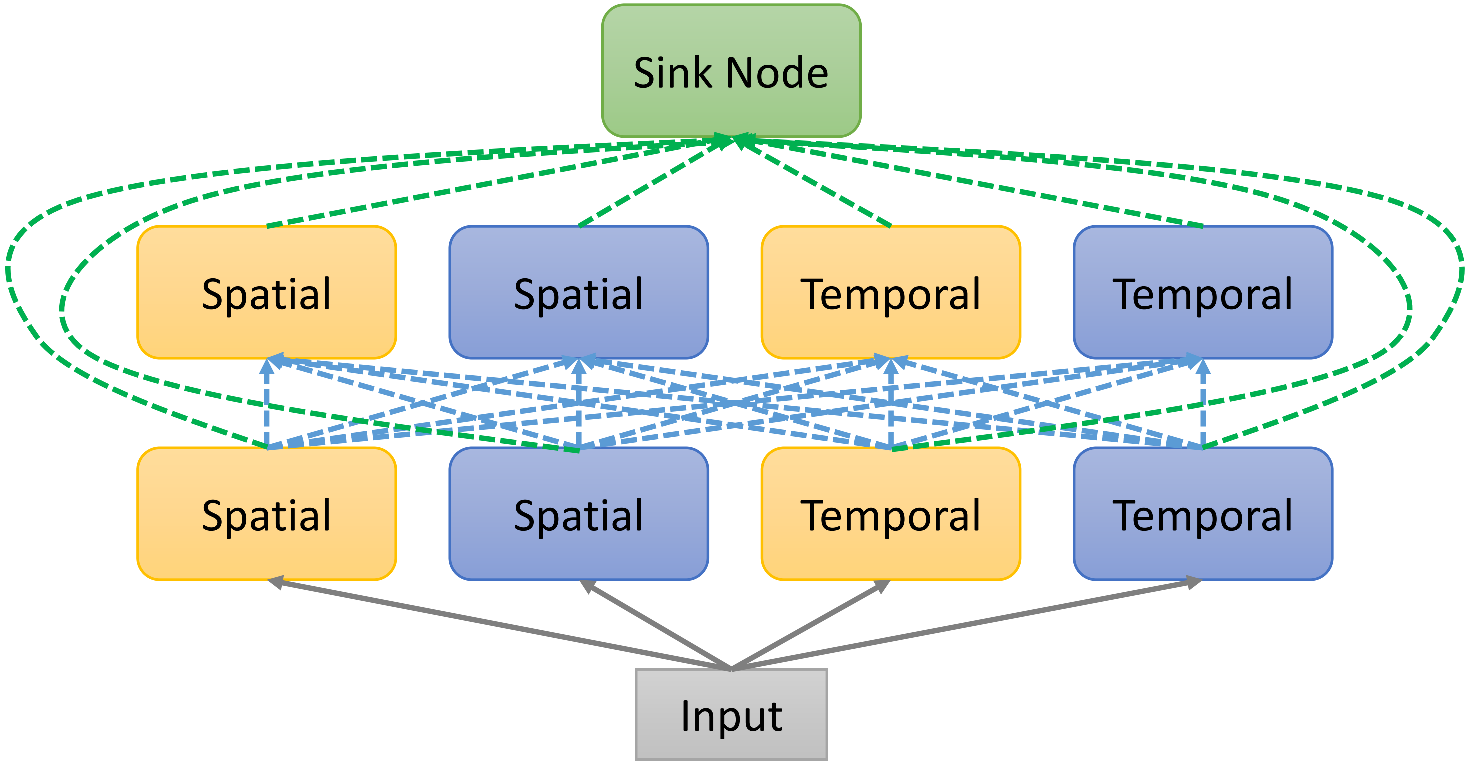
Solid connection (no weights)

Level connection weights w^{level}

Sink connection weights w^{sink}

Map-based Attention

Dot-product Attention



Supergraph: m levels; each level n nodes

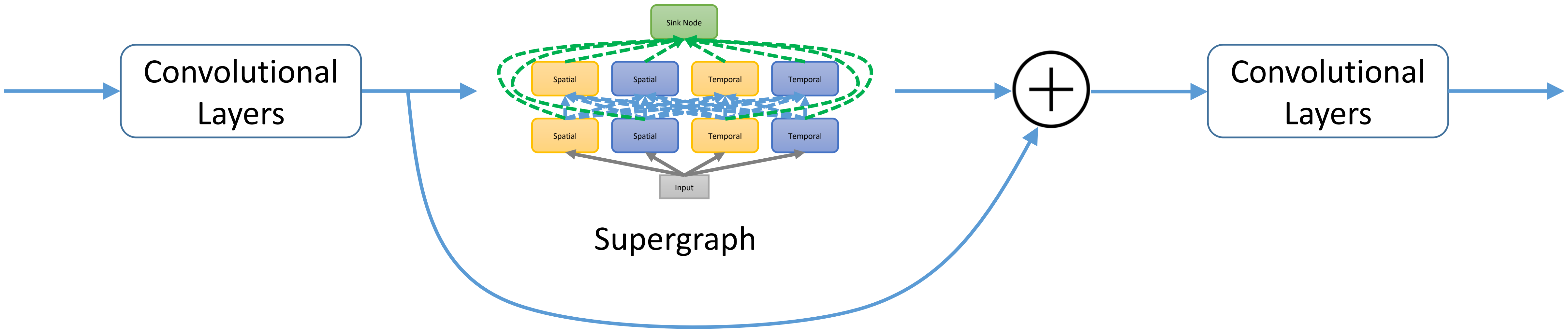
Node: an attention operation of a pre-defined attention dimension and type

$$f^{\text{in}}_{i,j} = \sum_{k=1}^{n} \boxed{w^{\text{level}}_{i,j,k}} \cdot f^{\text{out}}_{i-1}$$

$$f^{\text{out}}_{\text{sink}} = \sum_{1\le i\le m, 1\le j\le n} \boxed{w^{\text{sink}}_{i,j}} \cdot G_{i,j}(f^{\text{out}}_{i,j})$$

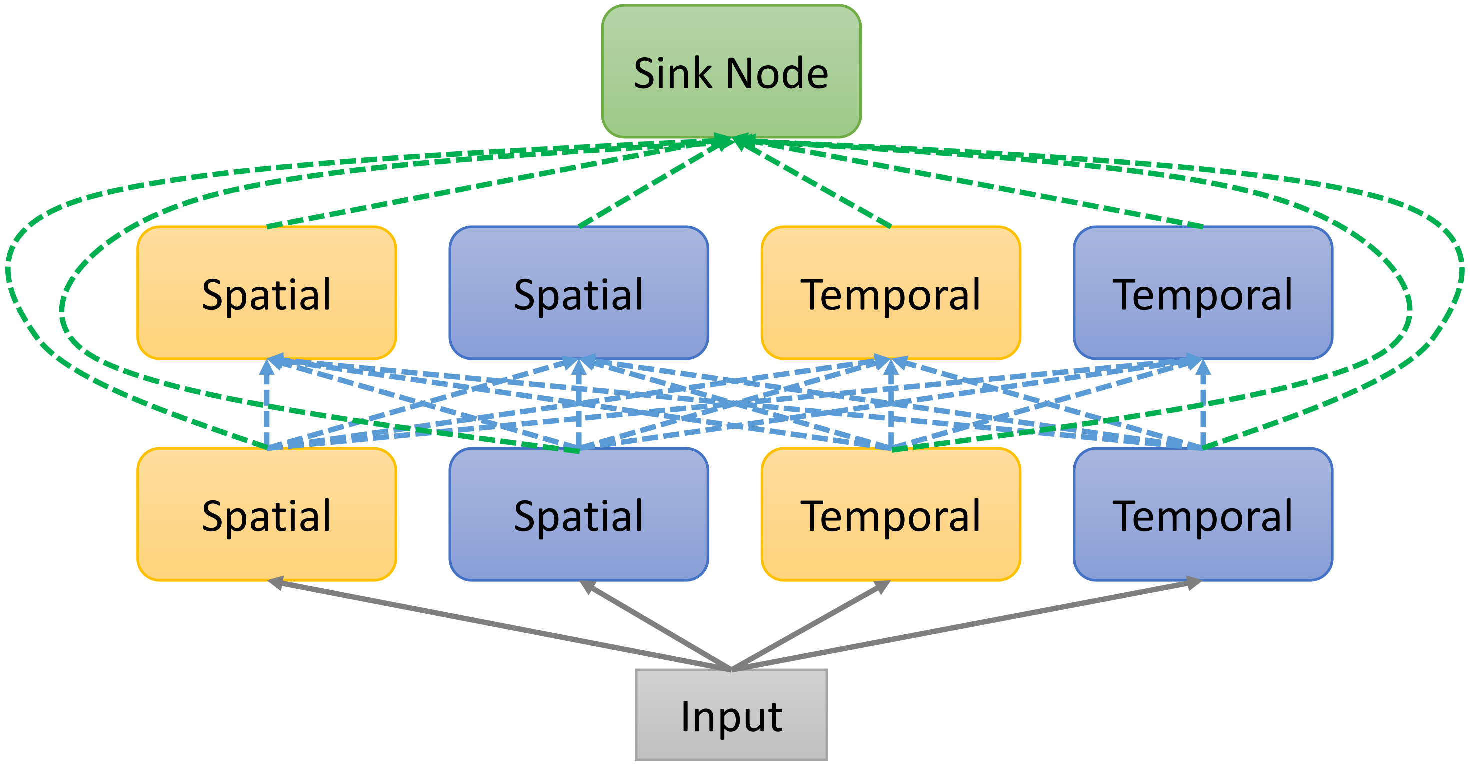
Differentiable Search

- Jointly train the network weights and connection weights with gradient descent



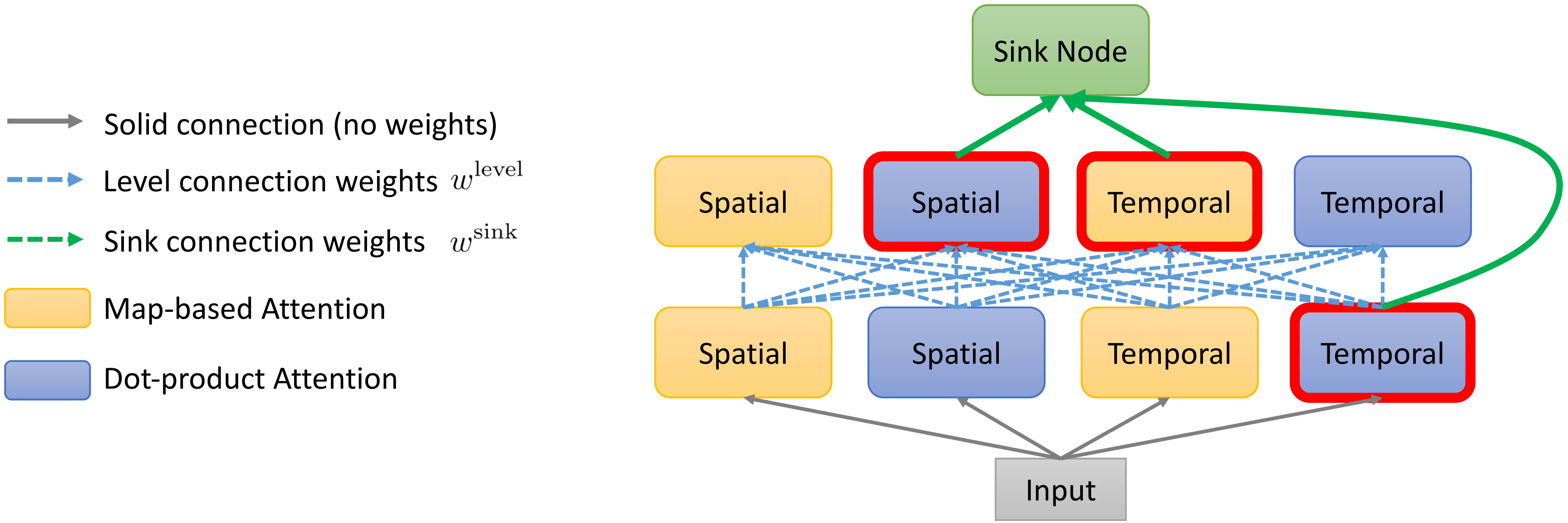
Attention Cell Design Derivation

- Solid connection (no weights)
- Level connection weights w^{level}
- Sink connection weights w^{sink}
- Map-based Attention
- Dot-product Attention



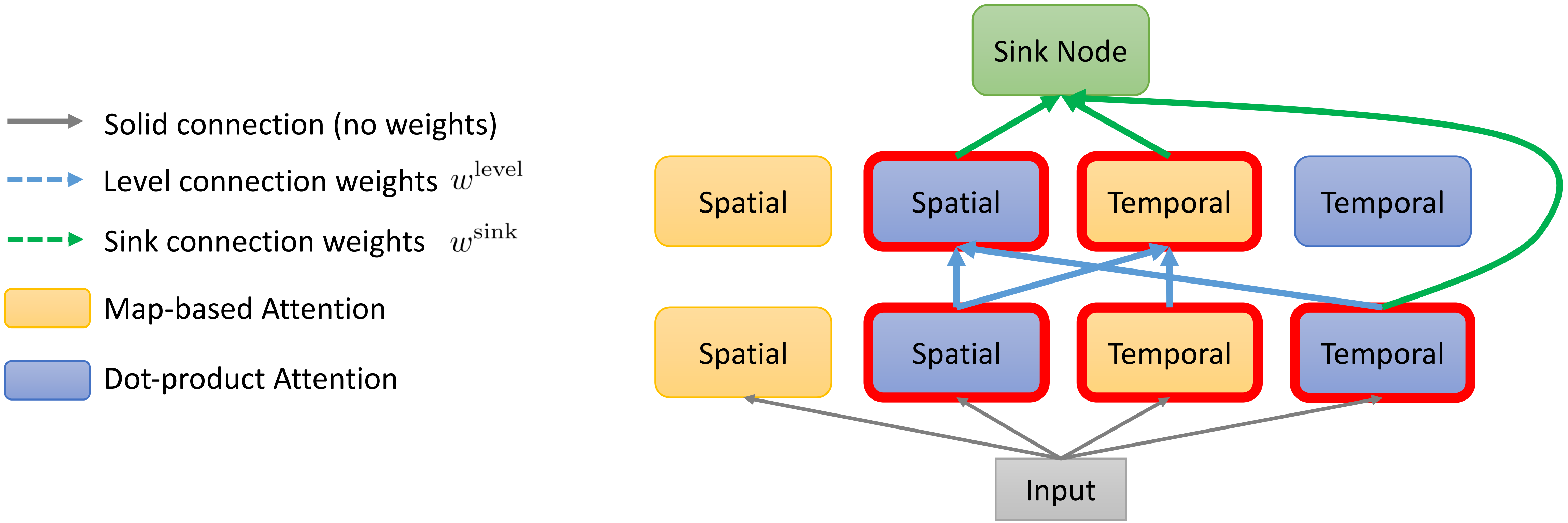
How to derive the attention cell design from the learned weights?

Attention Cell Design Derivation



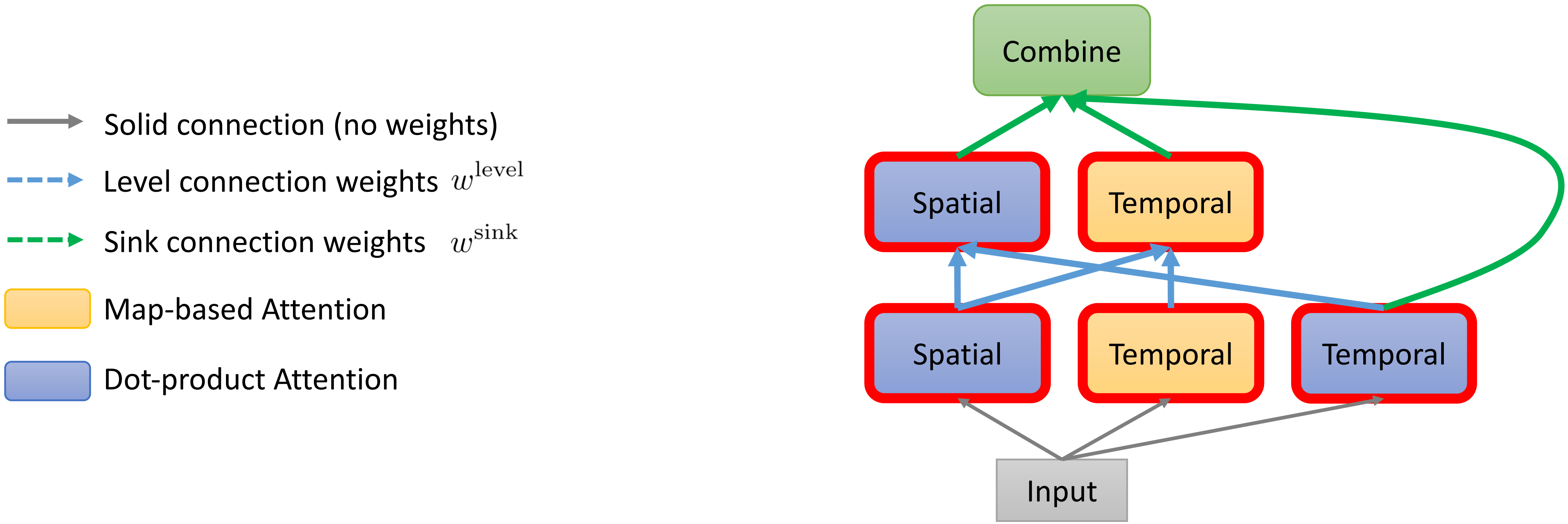
Choose top α (e.g., 3)
nodes based on w^{sink}

Attention Cell Design Derivation



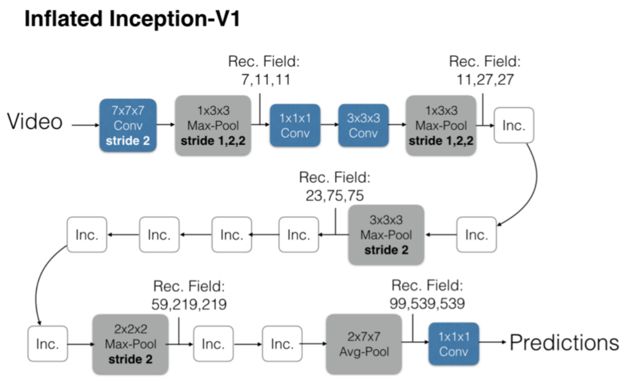
Choose top β (e.g., 2) predecessors of each selected code recursively based on w^{level} until we reach the first level

Attention Cell Design Derivation

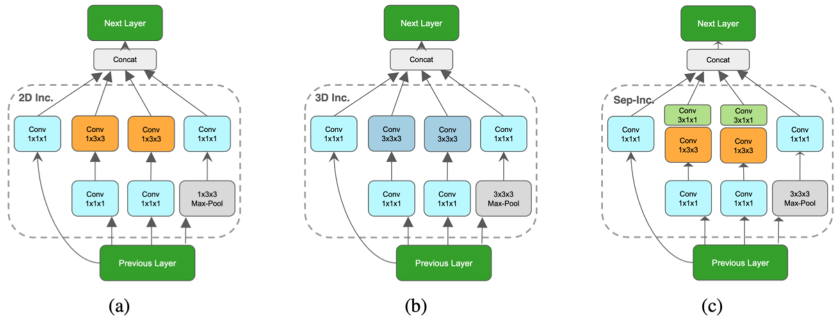


Experimental Setup

- Backbones
  - Inception-based
  - Insert 5 cells



I3D [CVPR 2017]



S3D [ECCV 2018]

- Datasets: Kinetics-600 and Moments in Time (MiT)

Comparison with Non-local Blocks

	Model	Kinetics Top-1	Kinetics Top-5	Kinetics ΔTop-1	MiT Top-1	MiT Top-5	MiT ΔTop-1
I3D	Backbone [5]	75.58	92.93	-	27.38	54.29	-
	Non-local [32]	76.87	93.44	1.29	28.54	55.35	1.16
	Ours	**77.86**	**93.75**	**2.28**	**29.58**	**56.62**	**2.20**
S3D	Backbone [36]	76.15	93.22	-	27.69	54.68	-
	Non-local [32]	77.56	93.68	1.41	29.52	56.91	1.83
	Ours	**78.51**	**93.88**	**2.36**	**29.82**	**57.02**	**2.13**

Generalization across Modalities

	Model	Kinetics Top-1	Kinetics Top-5	Kinetics ΔTop-1	MiT Top-1	MiT Top-5	MiT ΔTop-1
I3D	Backbone [5]	61.14	82.77	-	20.01	42.42	-
	Non-local [32]	64.88	85.77	3.74	21.86	**46.59**	1.85
	Ours	**66.81**	**87.85**	**5.67**	**21.94**	45.57	**1.93**
S3D	Backbone [36]	62.46	84.59	-	20.50	42.86	-
	Non-local [32]	65.79	86.85	3.33	22.13	**46.48**	1.63
	Ours	**67.02**	**87.72**	**4.56**	**22.52**	46.30	**2.02**

RGB to optical flow

Generalization across Backbones

	Model	Kinetics Top-1	Kinetics Top-5	Kinetics ΔTop-1	MiT Top-1	MiT Top-5	MiT ΔTop-1
I3D	Backbone [5]	75.58	92.93	-	27.38	54.29	-
	S3D	**77.81**	**93.74**	**2.23**	**29.26**	**56.61**	**1.88**
S3D	Backbone [36]	76.15	93.22	-	27.69	54.68	-
	I3D	**78.46**	**94.05**	**2.31**	**29.67**	**57.05**	**1.98**
I3D-R50	Backbone [32]	78.10	93.79	-	30.63	58.15	-
	I3D	**79.83**	**94.37**	**1.73**	**32.48**	**60.31**	**1.85**
	S3D	79.71	94.28	1.61	31.91	59.87	1.28

Generalization across Datasets

	Model	MiT to Kinetics			Kinetics to MiT		
		Top-1	Top-5	ΔTop-1	Top-1	Top-5	ΔTop-1
I3D	Backbone [5]	75.58	92.93	-	27.38	54.29	-
	Ours	**77.85**	**93.89**	**2.27**	**29.45**	**56.83**	**2.07**
S3D	Backbone [36]	76.15	93.22	-	27.69	54.68	-
	Ours	**78.19**	**93.98**	**2.04**	**29.33**	**56.33**	**1.64**

Comparison with State-of-the-art

(a) Kinetics-600.

Model	Top-1	Top-5	GFLOPs
I3D [5]	75.58	92.93	1136
S3D [36]	76.15	93.22	656
I3D-R50 [32]	78.10	93.79	938
D3D [27]	77.90	-	-
I3D+NL [32]	76.87	93.44	1305
S3D+NL [32]	77.56	93.68	825
TSN-IRv2 [31]	76.22	-	411
StNet-IRv2 [9]	78.99	-	440
SlowFast-R50 [7]	**79.9**	**94.5**	1971
I3D-R50+Cell	79.83	94.37	1034

(b) MiT.

Model	Top-1	Top-5	Modality
I3D [5]	27.38	54.29	RGB
S3D [36]	27.69	54.68	RGB
I3D+NL [32]	28.54	55.35	RGB
S3D+NL [32]	29.52	56.91	RGB
R50-ImageNet [18]	27.16	51.68	RGB
TSN-Spatial [31]	24.11	49.10	RGB
I3D-R50 [32]	30.63	58.15	RGB
I3D-R50+Cell	**32.48**	**60.31**	RGB
TSN-2stream [31]	25.32	50.10	R+F
TRN-Multiscale [40]	28.27	53.87	R+F
AssembleNet-50 [23]	31.41	58.33	R+F

Contributions

- Extend NAS beyond discovering convolutional cells to attention cells
- Search space for spatiotemporal attention cells
- A differentiable formulation of the search space
- State-of-the-art performance; outperforms non-local blocks
- Strong generalization across modalities, backbones, or datasets

- More analysis and visualizations of attention cells available in the paper