Spatial Mixture-of-Experts

NEURIPS 2022

This project received funding from the European Research Council under grant agreement MAELSTROM.
Data Has Spatial Structure
Data Has Spatial Structure

Weather

Geometric meshes

Registered data

Approximate alignment

And more...
Data Has Spatial Structure

Weather
Geometric meshes
Registered data
Approximate alignment
And more…
Data Has Spatial Structure

Weather

Geometric meshes

Registered data

Approximate alignment

And more...

Convolution
Data Has Spatial Structure

Weather
Geometric meshes
Registered data
Approximate alignment
And more...

Convolution
Locally-connected
Data Has Spatial Structure

- Weather
- Geometric meshes
- Registered data
- Approximate alignment
- And more…

Convolution

Locally-connected

Spatial Mixture-of-Experts

| Experts | Learned Gate |
|---------|--------------|
| ![Experts](image) | ![Learned Gate](image) |
Data Has Spatial Structure

- Weather
- Geometric meshes
- Registered data
- Approximate alignment
- And more...

Convolution

Locally-connected

Spatial Mixture-of-Experts

Experts

Learned Gate
Data Has Spatial Structure

- Weather
- Geometric meshes
- Registered data
- Approximate alignment
- And more…

Convolution

Locally-connected

Spatial Mixture-of-Experts

Experts

Learned Gate
Data Has Spatial Structure

- Weather
- Geometric meshes
- Registered data
- Approximate alignment
- And more...

Convolution

Locally-connected

Spatial Mixture-of-Experts

Experts

Learned Gate
Data Has Spatial Structure

Weather
Geometric meshes
Registered data
Approximate alignment
And more...

Convolution

Low-rank locally-connected

Locally-connected

Spatial Mixture-of-Experts

Experts

Learned Gate
Data Has Spatial Structure

Weather
Geometric meshes
Registered data
Approximate alignment
And more...

Convolution

Low-rank locally-connected
Mixture-of-Experts

Spatial Mixture-of-Experts

Experts
Learned Gate
Spatial Mixture-of-Experts
Spatial Mixture-of-Experts

\[ H \]

\[ W \]

\( C = 1 \text{ input channels} \)
Spatial Mixture-of-Experts

Experts $\mathcal{E}$

$H$

$W$

$C = 1$ input channels
Spatial Mixture-of-Experts

Experts $\mathcal{E}$

$H$

$W$

$C = 1$ input channels
Spatial Mixture-of-Experts

Experts $\mathcal{E}$

$G$, tensor routing

$D \sim |\mathcal{E}| \times H \times W$
Spatial Mixture-of-Experts

Experts $\mathcal{E}$

Gate $G$, tensor routing $\mathcal{D} \sim |\mathcal{E}| \times H \times W$

$G \sim |\mathcal{E}| \times H \times W$

$E \cdot F = 1$ output channels

$\text{top}_E$ sparse routing
Select $E = 1$ experts per pixel
Apply experts at locations

$D \sim |\mathcal{E}| \times H \times W$
Spatial Mixture-of-Experts

Experts $\mathcal{E}$

Gate $G$, tensor routing
$D \sim |\mathcal{E}| \times H \times W$

$\mathcal{E}$

top$_E$ sparse routing
Select $E = 1$ experts per pixel
Apply experts at locations

$E \cdot F = 1$ output channels

$H \times W$ channels
Spatial Mixture-of-Experts

Experts $\mathcal{E}$

Gate $G$, tensor routing
$D \sim |\mathcal{E}| \times H \times W$

$H \times W$

$C = 1$ input channels

$E \cdot F = 1$ output channels

$\text{top}_{E} \text{ sparse routing}$
Select $E = 1$ experts per pixel
Apply experts at locations

Routing Classification Loss
Spatial Mixture-of-Experts

Experts $\mathcal{E}$

Gate $G$, tensor routing $\mathcal{D} \sim |\mathcal{E}| \times H \times W$

$E \cdot F = 1$ output channels

$\text{top}_E$ sparse routing
Select $E = 1$ experts per pixel
Apply experts at locations

Routing Classification Loss

Select $E = 1$ experts per pixel
Apply experts at locations
Spatial Mixture-of-Experts

Experts $\mathcal{E}$

Gate $G$, tensor routing $\mathcal{D} \sim |\mathcal{E}| \times H \times W$

$H \cdot W = |\mathcal{E}|$ input channels

$E \cdot F = 1$ output channels

**top$_E$ sparse routing**
Select $E = 1$ experts per pixel
Apply experts at locations

Routing Classification Loss
Identify incorrect routings with error signal magnitude
Spatial Mixture-of-Experts

Experts $\mathcal{E}$

$H$ $W$

Gate $G$, tensor routing $\mathcal{D} \sim |\mathcal{E}| \times H \times W$

$E \cdot F = 1$ output channels

MSE loss $L = \frac{1}{N} \sum (X - Y)^2$

$\text{top}_E$ sparse routing
Select $E = 1$ experts per pixel
Apply experts at locations

Routing Classification Loss
Identify incorrect routings with error signal magnitude
Spatial Mixture-of-Experts

Experts $\mathcal{E}$

Gate $G$, tensor routing
$\mathcal{D} \sim |\mathcal{E}| \times H \times W$

$\text{top}_E$ sparse routing
Select $E = 1$ experts per pixel
Apply experts at locations

$E \cdot F = 1$ output channels

MSE loss
$L = \frac{1}{N} \sum (X - Y)^2$

Error signal
$\frac{dL}{dX} = \frac{2}{N} (X - Y)$

Routing Classification Loss
Identify incorrect routings with error signal magnitude
Spatial Mixture-of-Experts

Experts $\mathcal{E}$

Gate $G$, tensor routing $D \sim |\mathcal{E}| \times H \times W$

$E \cdot F = 1$ output channels

MSE loss $L = \frac{1}{N} \sum (X - Y)^2$

Error signal $\frac{dL}{dX} = \frac{2}{N} (X - Y)$

Routing Classification Loss

Identify incorrect routings with error signal magnitude

Select $E = 1$ experts per pixel
Apply experts at locations

top$_E$ sparse routing
Spatial Mixture-of-Experts

Experts $\mathcal{E}$

Gate $G$, tensor routing $D \sim |\mathcal{E}| \times H \times W$

$top_E$ sparse routing
Select $E = 1$ experts per pixel
Apply experts at locations

Routing Classification Loss
Construct labels
Identify incorrect routings with error signal magnitude

MSE loss
$L = \frac{1}{N} \sum (X - Y)^2$

Error signal
$\frac{dL}{dX} = \frac{2}{N} (X - Y)$
Spatial Mixture-of-Experts

Experts $\mathcal{E}$

Gate $G$, tensor routing $D \sim |\mathcal{E}| \times H \times W$

$E \cdot F = 1$ output channels

$\text{MSE loss } L = \frac{1}{N} \sum (X - Y)^2$

$\text{Error signal } \frac{dL}{dX} = \frac{2}{N} (X - Y)$

$\text{Select } E = 1 \text{ experts per pixel}$

$\text{Apply experts at locations}$

$\text{Routing Classification Loss}$

Construct labels

Identify incorrect routings with error signal magnitude

$1 \frac{1}{2}$ $0$ 1

$0 \frac{1}{2}$ 0 0

$1$ 0 0 0

✓ ✓ X

✓ ✓ ✓
Spatial Mixture-of-Experts

Experts $\mathcal{E}$

Gate $G$, tensor routing $\mathcal{D} \sim |\mathcal{E}| \times H \times W$

top$_E$ sparse routing
Select $E = 1$ experts per pixel
Apply experts at locations

Routing Classification Loss

Compute Loss
Construct labels
Identify incorrect routings with error signal magnitude

Binary cross-entropy

MSE loss
$L = \frac{1}{N} \sum (X - Y)^2$

Error signal
$\frac{dL}{dX} = \frac{2}{N} (X - Y)$
Spatial Mixture-of-Experts

**Experts $\mathcal{E}$**

- Gate $G$, tensor routing $\mathcal{D} \sim |\mathcal{E}| \times H \times W$

**Expert Error Damping**

- Damp error signals for incorrect routings
- $\alpha$

**top$_E$ sparse routing**

- Select $E = 1$ experts per pixel
- Apply experts at locations

**Routing Classification Loss**

- Compute Loss
- Construct labels
- Identify incorrect routings with error signal magnitude

**MSE loss**

$$L = \frac{1}{N} \sum (X - Y)^2$$

**Error signal**

$$\frac{dL}{dX} = \frac{2}{N}(X - Y)$$
Weather
Weather

Medium-range weather prediction **WeatherBench**

[Rasp et al., 2020; Rasp & Thuerey, 2021; Ashkboos et al., 2022; Elsayed et al., 2020; Dosovitskiy et al., 2021]
Weather

Medium-range weather prediction \textit{WeatherBench}

\[ T = 0 \quad \rightarrow \quad T + 3 \text{ or } 5 \text{ days} \]

[Rasp et al., 2020; Rasp & Thueray, 2021; Ashkboos et al., 2022; Elsayed et al., 2020; Dosovitskiy et al., 2021]
Medium-range weather prediction **WeatherBench**

\[ T = 0 \quad \rightarrow \quad T + 3 \text{ or } 5 \text{ days} \]

| Model                  | Z500 [m²s⁻²] | T850 [K]  |
|------------------------|--------------|-----------|
|                        | 3 days | 5 days | 3 days | 5 days |
| Rasp & Thuerey         | 316±2.4  | 563±3.1 | 1.80±0.02 | 2.84±0.03 |
| 2x wide                | 310±2.0  | 555±2.8 | 1.76±0.03 | 2.78±0.01 |
| LRLCN                  | 290±1.4  | 549±1.9 | 1.73±0.03 | 2.79±0.01 |
| ViT                    | 438±2.8  | 638±3.1 | 2.24±0.04 | 2.88±0.03 |
| SMoE                   | 270±2.0  | 525±2.0 | 1.66±0.02 | 2.60±0.01 |

[Rasp et al., 2020; Rasp & Thuerey, 2021; Ashkboos et al., 2022; Elsayed et al., 2020; Dosovitskiy et al., 2021]
Weather

Medium-range weather prediction **WeatherBench**

![Diagram: WeatherBench]

\[ T = 0 \quad \rightarrow \quad T + 3 \text{ or } 5 \text{ days} \]

| Model                  | Z500 [m²s⁻²] | T850 [K]       |
|------------------------|--------------|----------------|
|                        | 3 days       | 5 days         | 3 days       | 5 days       |
| Rasp & Thuerey         | 316±2.4      | 563±3.1        | 1.86±0.02    | 2.84±0.03    |
| ➔ 2x wide             | 310±2.0      | 555±2.8        | 1.76±0.03    | 2.78±0.01    |
| LRLCN                  | 290±1.4      | 549±1.9        | 1.73±0.03    | 2.79±0.01    |
| ViT                    | 438±2.8      | 638±3.1        | 2.24±0.04    | 2.88±0.03    |
| SMoE                   | 270±2.0      | 525±2.0        | 1.66±0.02    | 2.60±0.01    |
| R&T (pretrained)       | 267±1.8      | 500±2.4        | 1.66±0.03    | 2.43±0.02    |
| SMoE (pretrained)      | 253±2.1      | 488±1.7        | 1.57±0.02    | 2.34±0.02    |
| ➔ + extra ERA5         | 232±1.5      | 440±1.2        | 1.46±0.02    | 2.19±0.01    |
| ➔ + 1.4°               | 198±1.6      | 382±2.0        | 1.42±0.00    | 2.06±0.02    |

[Rasp et al., 2020; Rasp & Thuerey, 2021; Ashkboos et al., 2022; Elsayed et al., 2020; Dosovitskiy et al., 2021]
# Weather

## Medium-range weather prediction

**WeatherBench**

![WeatherBench Diagram](https://via.placeholder.com/150)

\[ T = 0 \quad \xrightarrow{\text{post-processing}} \quad T + 3 \text{ or } 5 \text{ days} \]

| Model                  | Z500 [m²s⁻²] | T850 [K]  |
|------------------------|--------------|-----------|
|                        | 3 days       | 5 days    | 3 days    | 5 days    |
| Rasp & Thuerey         | 316 ± 2.4    | 563 ± 3.1 | 1.86 ± 0.02 | 2.84 ± 0.03 |
| 2x wide                | 310 ± 2.0    | 555 ± 2.8 | 1.76 ± 0.03 | 2.78 ± 0.01 |
| LRLCN                  | 290 ± 1.4    | 549 ± 1.9 | 1.73 ± 0.03 | 2.79 ± 0.01 |
| ViT                    | 438 ± 2.8    | 638 ± 3.1 | 2.24 ± 0.04 | 2.88 ± 0.03 |
| SMoE                   | 270 ± 2.0    | 525 ± 2.0 | 1.66 ± 0.02 | 2.60 ± 0.01 |
| R&T (pretrained)       | 267 ± 1.8    | 500 ± 2.4 | 1.66 ± 0.03 | 2.43 ± 0.02 |
| SMoE (pretrained)      | 253 ± 2.1    | 488 ± 1.7 | 1.57 ± 0.02 | 2.34 ± 0.02 |
| + extra ERA5           | 232 ± 1.5    | 440 ± 1.2 | 1.46 ± 0.02 | 2.19 ± 0.01 |
| + 1.4°                 | 198 ± 1.8    | 382 ± 2.0 | 1.42 ± 0.00 | 2.06 ± 0.02 |

[Rasp et al., 2020; Rasp & Thuerey, 2021; Ashkboos et al., 2022; Elsayed et al., 2020; Dosovitskiy et al., 2021]
### Weather

**Medium-range weather prediction** WeatherBench

| Model                  | Z500 [m$^2$s$^{-2}$] | T850 [K] |
|------------------------|-----------------------|----------|
|                        | 3 days    | 5 days    | 3 days    | 5 days    |
| Rasp & Thuerey         | 316±2.4   | 563±3.1   | 1.86±0.02 | 2.84±0.03 |
| ➡️ 2x wide             | 310±2.0   | 555±2.8   | 1.76±0.03 | 2.78±0.01 |
| LRLCN                  | 290±1.4   | 549±1.9   | 1.73±0.03 | 2.79±0.01 |
| ViT                    | 438±2.8   | 638±3.1   | 2.24±0.04 | 2.88±0.03 |
| SMoE                   | 270±2.0   | 525±2.0   | 1.66±0.02 | 2.60±0.01 |
| R&T (pretrained)       | 267±1.8   | 500±2.4   | 1.66±0.03 | 2.43±0.02 |
| SMoE (pretrained)      | 253±2.1   | 488±1.7   | 1.57±0.02 | 2.34±0.02 |
| ➡️ + extra ERA5        | 232±1.5   | 440±1.2   | 1.46±0.02 | 2.19±0.01 |
| ➡️ + 1.4°              | 198±1.8   | 382±2.0   | 1.42±0.00 | 2.06±0.02 |

[Rasp et al., 2020; Rasp & Thuerey, 2021; Ashkboos et al., 2022; Elsayed et al., 2020; Dosovitskiy et al., 2021]
Weather

Medium-range weather prediction **WeatherBench**

\[ T = 0 \quad \rightarrow \quad T + 3 \text{ or } 5 \text{ days} \]

| Model          | Z500 [m²s⁻²] | T850 [K] |
|----------------|--------------|----------|
|                | 3 days       | 5 days   | 3 days   | 5 days   |
| Rasp & Thuerey | 316±2.4      | 563±3.1  | 1.86±0.02| 2.84±0.03|
| 2x wide        | 310±2.0      | 555±2.8  | 1.76±0.03| 2.78±0.01|
| LRLCN          | 290±1.4      | 549±1.9  | 1.73±0.03| 2.79±0.01|
| VIT            | 438±2.8      | 638±3.1  | 2.24±0.04| 2.88±0.03|
| SMoE           | 270±2.0      | 525±2.0  | 1.66±0.02| 2.60±0.01|
| R&T (pretrained)| 267±1.8      | 500±2.4  | 1.66±0.03| 2.43±0.02|
| SMoE (pretrained)| 253±2.1     | 488±1.7  | 1.57±0.02| 2.34±0.02|
| 2x + extra ERA5| 232±1.5      | 440±1.2  | 1.46±0.02| 2.19±0.01|
| 1.4°           | 198±1.8      | 382±2.0  | 1.42±0.00| 2.06±0.02|

[Emory et al., 2020; Rasp & Thuerey, 2021; Ashkboos et al., 2022; Elsayed et al., 2020; Dosovitskiy et al., 2021]

**Ensemble post-processing ENS-10**

| Model          | Z500 [m²s⁻²] | T850 [K] | T2M [K] |
|----------------|--------------|----------|---------|
|                | 5 ens        | 10 ens   | 5 ens   | 10 ens   | 5 ens   | 10 ens   |
| EMOS           | 79.12±0.12   | 78.80±0.21| 0.721±0.01| 0.706±0.04| 0.720±0.00| 0.711±0.03|
| U-Net          | 76.54±0.20   | 76.18±0.12| 0.685±0.00| 0.670±0.01| 0.657±0.01| 0.644±0.01|
| SMoE           | 68.94±0.14   | 67.43±0.12| 0.612±0.01| 0.590±0.02| 0.601±0.02| 0.594±0.02|
| CRPS           | 29.21±0.18   | 29.02±0.13| 0.247±0.00| 0.245±0.00| 0.244±0.00| 0.241±0.02|
| U-Net          | 27.78±0.11   | 27.55±0.19| 0.230±0.01| 0.229±0.01| 0.225±0.00| 0.220±0.01|
| SMoE (pretrained)| 23.79±0.20   | 23.10±0.16| 0.207±0.03| 0.197±0.03| 0.199±0.01| 0.190±0.02|