Representation learning and custom loss functions for atmospheric data

Christian Lessig (joint work with Sebastian Hoffmann)
Otto-von-Guericke-Universität Magdeburg

\[ \mathcal{L} = \sum_{j=1}^{N} \| h_j - \tilde{h}_j \| \]
Motivation

- Learning relies on informative and discriminative loss functions
Motivation

vorticity

MSE / $L_2$
Motivation
Motivation
Motivation

vorticity
Motivation

vorticity

MSE / $L_2$

© Christian Lessig, 2021
Motivation

vorticity

MSE / $L_2$
Motivation

vorticity

MSE / $L_2$

$0.2 \ 0.4 \ 0.6 \ 0.8 \ 1.0 \ 1.2 \ 1.4$

$10 \ 20 \ 30 \ 40 \ 50 \ 60 \ 70 \ h$
Motivation

vorticity

MSE / $L_2$
Motivation

vorticity

MSE / $L_2$
Motivation

vorticity

MSE / $L_2$
Motivation

vorticity

MSE / $L_2$

h

10 20 30 40 50 60 70

0.2
0.4
0.6
0.8
1.0
1.2
1.4
Motivation

vorticity

MSE / $L_2$
Motivation

vorticity

MSE / $L_2$
Motivation

vorticity
Motivation

vorticity

MSE / $L_2$

© Christian Lessig, 2021
Motivation
Motivation
Motivation

vorticity

MSE / $L_2$
Motivation

vorticity

MSE / $L_2$
Motivation
Motivation

- Learning relies on informative and discriminative loss functions
  - Mathematically grounded norms are often not particularly effective
Motivation

- Learning relies on informative and discriminative loss functions
  - Mathematically grounded norms are often not particularly effective
  - Similar issues for natural images
Motivation

- GAN-based super-resolution:\(^1\)

Motivation

- GAN-based super-resolution:¹

![SRGAN-MSE](image1)

![original HR image](image2)

Motivation

- GAN-based super-resolution:¹

SRGAN-MSE  SRGAN-VGG22  

<table>
<thead>
<tr>
<th>Dataset</th>
<th>HR</th>
<th>L-MSE</th>
<th>L-VGG</th>
<th>L-MOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set5</td>
<td>25.02</td>
<td>25.94</td>
<td>26.68</td>
<td>26.83</td>
</tr>
<tr>
<td>BSD100</td>
<td>24.64</td>
<td>25.99</td>
<td>27.18</td>
<td>27.45</td>
</tr>
<tr>
<td>Set14</td>
<td>26.26</td>
<td>28.43</td>
<td>30.07</td>
<td>30.33</td>
</tr>
</tbody>
</table>

Motivation

- GAN-based super-resolution:¹

---

Motivation

Figure 2. Visualization of features in a fully trained model. For layers 2-5 we show the top 9 activations in a random subset of feature maps across the validation data, projected down to pixel space using our deconvolutional network approach. Our reconstructions are not samples from the model: they are reconstructed patterns from the validation set that cause high activations in a given feature map. For each feature map we also show the corresponding image patches. Note: (i) the strong grouping within each feature map, (ii) greater invariance at higher layers and (iii) exaggeration of discriminative parts of the image, e.g. eyes and noses of dogs (layer 4, row 1, cols 1). Best viewed in electronic form.

Motivation

Figure 2. Visualization of features in a fully trained model. For layers 2-5 we show the top 9 activations in a random subset of feature maps across the validation data, projected down to pixel space using our deconvolutional network approach. Our reconstructions are not samples from the model: they are reconstructed patterns from the validation set that cause high activations in a given feature map. For each feature map we also show the corresponding image patches. Note: (i) the strong grouping within each feature map, (ii) greater invariance at higher layers and (iii) exaggeration of discriminative parts of the image, e.g. eyes and noses of dogs (layer 4, row 1, cols 1). Best viewed in electronic form.

Motivation

Figure 2. Visualization of features in a fully trained model. For layers 2-5 we show the top 9 activations in a random subset of feature maps across the validation data, projected down to pixel space using our deconvolutional network approach. Our reconstructions are not samples from the model: they are reconstructed patterns from the validation set that cause high activations in a given feature map. For each feature map we also show the corresponding image patches. Note: (i) the strong grouping within each feature map, (ii) greater invariance at higher layers and (iii) exaggeration of discriminative parts of the image, e.g. eyes and noses of dogs (layer 4, row 1, cols 1). Best viewed in electronic form.

Motivation
Motivation
Motivation
Motivation

Feature space is task / domain specific
Motivation

Feature space is task / domain specific

Compute norm there!
Motivation

- GAN-based super-resolution:¹

![GAN-based super-resolution results](image)

Motivation

- Neural style transfer:¹

Motivation

Feature space is task / domain specific

Compute norm there!
Motivation

Train on some task that requires the network to learn domain specific features

Feature space is task / domain specific

Compute norm there!
Motivation

- Pretext task: inpainting of randomly deleted image parts\(^1\)

---

Motivation

- Pretext task: predicting deleted color and gray scale channels

\[ \text{Input Image } X \xrightarrow{\mathcal{F}_1} \text{Predicted Image } \hat{X} \]

\[ \begin{align*}
\text{L Grayscale Channel } X_L & \quad \text{Predicted Color Channels } \hat{X}_C \\
\text{ab Color Channels } X_{ab} & \quad \text{Predicted Grayscale Channel } \hat{X}_G
\end{align*} \]

\[ \downarrow \]

\[ \begin{align*}
\mathcal{F}_1 & \quad \mathcal{F}_2
\end{align*} \]

\[ \downarrow \]

\[ \hat{X}_C \quad \hat{X}_G \]

---

Motivation

How can we adapt these ideas to atmospheric data?
Atmospheric data

- Wind field, vorticity, divergence, temperature, geopotential height, precipitation, ...
Atmospheric data

- Wind field, vorticity, divergence, temperature, geopotential height, precipitation, ...
- Image-like in grid representation
  - With usual issues but good starting point
Atmospheric data

- Wind field, vorticity, divergence, temperature, geopotential height, precipitation, ...
- Image-like in grid representation
  - With usual issues but good starting point
- ERA5 provides well curated data set for training
  - Contains effects we cannot model
Atmospheric data

- Wind field, vorticity, divergence, temperature, geopotential height, precipitation, ...
- Image-like in grid representation
  - With usual issues but good starting point
- ERA5 provides well curated data set for training
  - Contains effects we cannot model
  - But unlabelled
AtmoDist$^1$

- Custom distance metric for vorticity + divergence (wind velocity vector field)

AtmoDist\(^1\)

- Custom distance metric for vorticity + divergence (wind velocity vector field)
- GAN-based super-resolution / downscaling as validation application
  - Recent work by Stengel et al.\(^2\) as baseline

\(^1\) S. Hoffmann and C. Lessig. Towards representation learning for atmospheric data. In NEURIPS 2021 Workshop on Climate Change (poster), 2021.
AtmoDist

What pretext task can we use?

AtmoDist
AtmoDist

1979 \( t \) -- 2020
AtmoDist
AtmoDist
AtmoDist

1979 $t$ → 2020

$\Delta t$
AtmoDist
AtmoDist
AtmoDist

prediction of $\Delta t$

classification network

representation network

1979 $t$  \[\Delta t\]  2020
AtmoDist

prediction of $\Delta t$

classification network

representation network

$h_J$

1979 $t$ $\rightarrow$ 2020

$\Delta t$
AtmoDist

\[ N \]

\[ p(\Delta t) \]

\[ h_J \]

\[ 1979 \quad t \]

\[ \Delta t \]

\[ 2020 \]
AtmoDist

- Classification network
- Prediction of $\Delta t$
- Representation network

1979 $\rightarrow$ $\Delta t$ $\rightarrow$ 2020
AtmoDist

$$d(\zeta, \tilde{\zeta}) = \| h_J(\zeta) - h_J(\tilde{\zeta}) \|$$

prediction of $\Delta t$

classification network

representation network

1979 $t$  \hspace{1cm}  \Delta t \hspace{1cm} 2020
AtmoDist: data

- ERA5\(^\dagger\) reanalysis 1979-2006
    (58,440 training slices and 17,536 evaluation ones)

\(^\dagger\) Hersbach et al., The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146 (730), 2020.
AtmoDist: data

- ERA5\(^1\) reanalysis 1979-2006
    (58,440 training slices and 17,536 evaluation ones)
  - Vorticity and divergence

\(^1\) Hersbach et al., The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146 (730), 2020.
AtmoDist: data

Distribution of data:
AtmoDist: data

- ERA5\(^1\) reanalysis 1979-2006
  - Vorticity and divergence
  - 1280 × 2560 grids sampled into 160 x 160 patches

\(^1\)Hersbach et al., The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146 (730), 2020.
AtmoDist: data

- ERA5\(^1\) reanalysis 1979-2006
    (58,440 training slices and 17,536 evaluation ones)
  - Vorticity and divergence
  - 1280 × 2560 grids sampled into 160 x 160 patches
  - One vertical layer (≈ 883 hPa)

\(^1\) Hersbach et al., The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146 (730), 2020.
AtmoDist: network

vorticity, divergence
AtmoDist: network

\[ vorticity, \]
\[ divergence \]

\[ t \]

\[ t + \Delta t \]
AtmoDist: network

vorticity, divergence

\[ t \]

\[ t + \Delta t \]
AtmoDist: network

vorticity, divergence

$\mathbf{t}$

$\mathbf{t} + \Delta \mathbf{t}$

shared weights
AtmoDist: network

- **vorticity, divergence**
- **representation network**
  - 8x8 Conv - 16
  - 3x3 MaxPool
  - 3x3 ResBlock - 16
  - 3x3 ResBlock - 32
  - 3x3 ResBlock - 64
  - 3x3 ResBlock - 128

- **shared weights**
- 2.27 million parameters
AtmoDist: network

representation network

shared weights

classification network

vorticity, divergence

$t$

$t + \Delta t$
AtmoDist: training

- 23 classes with max. time difference of 69 hours
- Pre-training with subset of data to improve convergence
AtmoDist: training

- 23 classes with max. time difference of 69 hours
- Pre-training with subset of data to improve convergence
AtmoDist: evaluation
AtmoDist: evaluation
AtmoDist: evaluation

\[ d_{MSE}(\zeta_t, \zeta_{t+\Delta t}) \]
AtmoDist: evaluation

\[ d_{\text{MSE}}(\zeta_t, \zeta_{t+\Delta t}) \]

\[ d_{\text{AD}}(\zeta_t, \zeta_{t+\Delta t}) \]
AtmoDist: evaluation

- **vorticity, divergence**
- **representation network**
  - $t$
  - $t + \Delta t$

```
representation network

8x8 Conv - 16
3x3 MaxPool
3x3 ResBlock - 16
3x3 ResBlock - 32
3x3 ResBlock - 64
3x3 ResBlock - 128
6x 6x 6x 6x
6x 6x 6x 6x

Stack - 256
3x3 Conv - 128
flatten
Linear + softmax
```
AtmoDist: evaluation

divergence, vorticity

t

representation network

$t + \Delta t$
AtmoDist: evaluation
AtmoDist: evaluation

vorticity, divergence

representation network

\[ \| h_J(\zeta_t) - h_J(\zeta_{t+\Delta t}) \|^2_2 \]
AtmoDist: evaluation

\[ d_{\text{MSE}}(\zeta_t, \zeta_{t+\Delta t}) \quad \text{and} \quad d_{\text{AD}}(\zeta_t, \zeta_{t+\Delta t}) \]
AtmoDist: evaluation

![Graph of time difference between samples vs. average loss](image-url)

- Time difference between samples:
  - 0h
  - 12h
  - 24h
  - 36h
  - 48h
  - 60h

- Average loss:
  - MSE
  - Ours

© Christian Lessig, 2021
Super-resolution using AtmoDist

- Objective: down-scaling / upsampling of coarse fields
Super-resolution using AtmoDist

- Objective: down-scaling / upsampling of coarse fields
- Comparison using and with the GAN of Stengel et al.\(^1\)


Super-resolution using AtmoDist

- Objective: down-scaling / upsampling of coarse fields
- Comparison using and with the GAN of Stengel et al.¹
  - GAN is based on SRGAN² for natural images

Super-resolution using AtmoDist

- Objective: down-scaling / upsampling of coarse fields
- Comparison using and with the GAN of Stengel et al.\(^1\)
  - GAN is based on SRGAN\(^2\) for natural images
  - Our content loss replaces mean squared error


Super-resolution using AtmoDist

- Objective: down-scaling / upsampling of coarse fields
- Comparison using and with the GAN of Stengel et al.\(^1\)
  - GAN is based on SRGAN\(^2\) for natural images
  - Our content loss replaces mean squared error
  - Only 4X super-resolution in our work


Super-resolution using AtmoDist

ours  ground thruth  mse
Super-resolution using AtmoDist

ours  ground thruth  mse
Super-resolution using AtmoDist

ours  ground thruth  mse
Super-resolution using AtmoDist

ours

ground thruth

cmse
Super-resolution using AtmoDist
Super-resolution using AtmoDist

energy spectrum

energy

wavenumber

ground truth
ours
mse
Super-resolution using AtmoDist

![Semivariogram](image)

- **Groundtruth**
- **Ours**
- **MSE**

Normalized variance vs. lag distance [km]
Super-resolution using AtmoDist

![Total variation distribution](image)

- **Count**
- **Total variation difference**
- **Model:**
  - Ours
  - MSE
Super-resolution using AtmoDist

- Local statistics by averaging over super-resolution predictions for entire reanalysis data set
- 150 big cities as locations
Super-resolution using AtmoDist

- Local statistics by averaging over super-resolution predictions for entire reanalysis data set
- 150 big cities as locations
Super-resolution using AtmoDist

- Local statistics by averaging over super-resolution predictions for entire reanalysis data set
- 150 big cities as locations
Super-resolution using AtmoDist

Wasserstein distance to ground truth for local statistics
Summary

- Custom loss functions for atmospheric data are possible and useful
  - Prediction of time difference is an effective pretext task
  - Applicable to wide range of data sets
Summary

- Custom loss functions for atmospheric data are possible and useful
  - Prediction of time difference is an effective pretext task
  - Applicable to wide range of data sets
- Super-resolution using AtmoDist improves quantitative and qualitative results
  - Local statistics still need more work
Outlook: applications

◦ Super-resolution / downscaling
  › Larger amplification factor
  › Adapted to local regions / specific phenomena
Outlook: applications

- Super-resolution / downscaling
  - Larger amplification factor
  - Adapted to local regions / specific phenomena
- Detection of atmospheric patterns
  - Extreme events, blocking, ...
Outlook: applications

- Super-resolution / downscaling
  - Larger amplification factor
  - Adapted to local regions / specific phenomena
- Detection of atmospheric patterns
  - Extreme events, blocking, ...
- Hybrid dynamical core for GCMs
Outlook: applications

- Super-resolution / downscaling
  - Larger amplification factor
  - Adapted to local regions / specific phenomena
- Detection of atmospheric patterns
  - Extreme events, blocking, ...
- Hybrid dynamical core for GCMs
- Applications directly addressing climate change
Outlook: custom loss functions

- Other / more atmospheric fields
  - Separate dynamic variables and tracers?
  - More vertical layers
Outlook: custom loss functions

- Other / more atmospheric fields
  - Separate dynamic variables and tracers?
  - More vertical layers
- Transfer learning / refinement to local regions
  - Train with only local data in a final training phase
Outlook: custom loss functions

- Other / more atmospheric fields
  - Separate dynamic variables and tracers?
  - More vertical layers
- Transfer learning / refinement to local regions
  - Train with only local data in a final training phase
- Shift in distribution (e.g. global warming)
  - Parametrized model?
Outlook: representation learning

- Un-/semi-supervised learning has significant potential for atmospheric data
  - Side-steps need for labelling of data
  - Generically used for pre-training with natural images
Outlook: representation learning

- Un-/semi-supervised learning has significant potential for atmospheric data
  - Side-steps need for labelling of data
  - Generically used for pre-training with natural images
- Integrate heterogeneous data sources
  - Reanalysis and simulations
  - High frequency, high resolution satellite data
Outlook: theoretical foundations

- Theoretical integration of analytic models and machine learning
  - E.g. predict effective network architecture
- Consistency, stability, and convergence for (hybrid) simulations
- Predict effectiveness of learning / data for climate dynamics
AtmoDist: https://arxiv.org/abs/2109.09076

Slides via KITP Online Talks repository and on my hompage