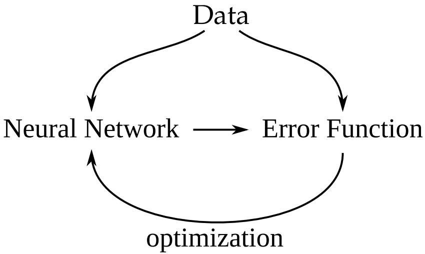
Geometric Deep Learning: From Pure Math to Applications

Jan E. Gerken

CHALMERS
UNIVERSITY OF TECHNOLOGY

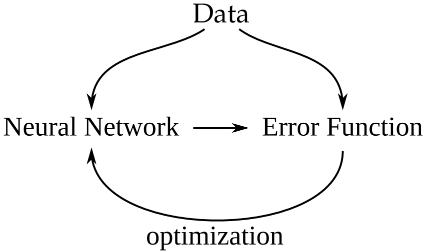
WASP
WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

12th April 2023
WASP Math/AI Meeting
KTH Stockholm


Data
Neural Network
Error Function
optimization

Deep learning is lacking a strong theoretical foundation:

- Neural networks are complicated functions
- The training process is stochastic

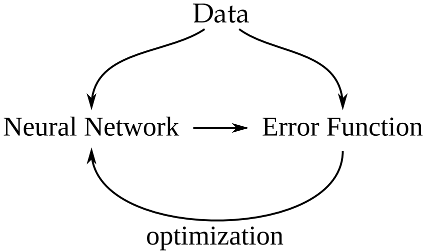


Geometric Deep Learning

Deep learning is lacking a strong theoretical foundation:

- Neural networks are complicated functions
- The training process is stochastic

Geometry can help to alleviate these problems.

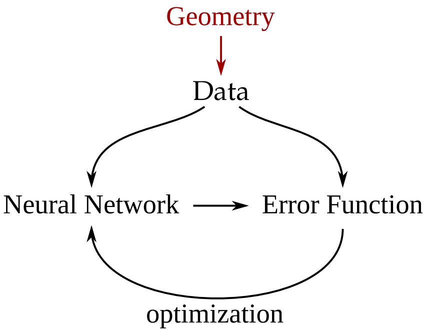


Geometric Deep Learning

Deep learning is lacking a strong theoretical foundation:

- Neural networks are complicated functions
- The training process is stochastic

Geometry can help to alleviate these problems.

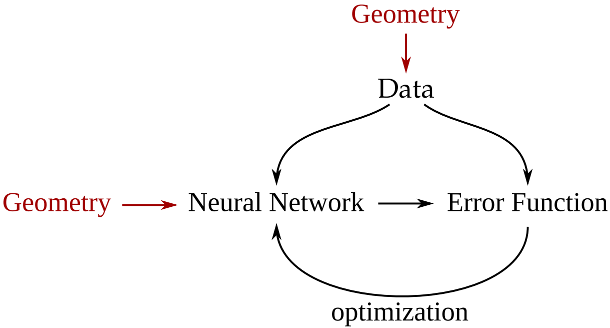


Geometric Deep Learning

Deep learning is lacking a strong theoretical foundation:

- Neural networks are complicated functions
- The training process is stochastic

Geometry can help to alleviate these problems.

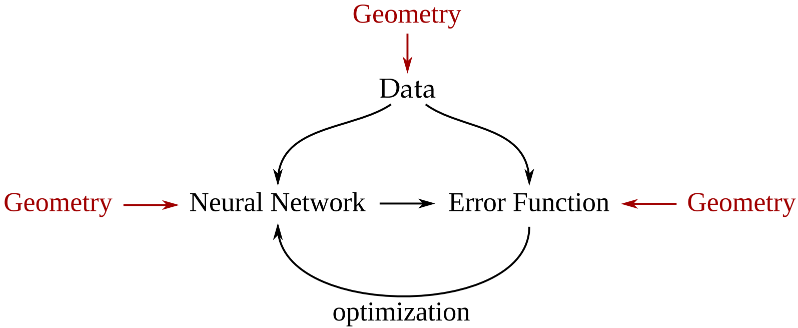


Geometric Deep Learning

Deep learning is lacking a strong theoretical foundation:

- Neural networks are complicated functions
- The training process is stochastic

Geometry can help to alleviate these problems.

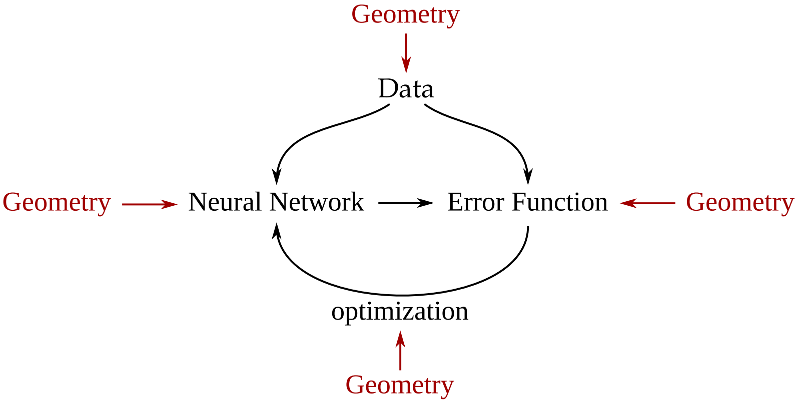


Geometric Deep Learning

Deep learning is lacking a strong theoretical foundation:

- Neural networks are complicated functions
- The training process is stochastic

Geometry can help to alleviate these problems.

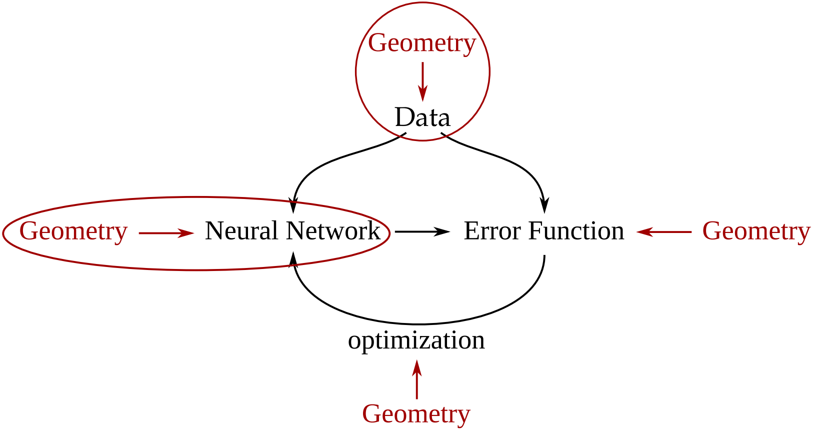


Geometric Deep Learning

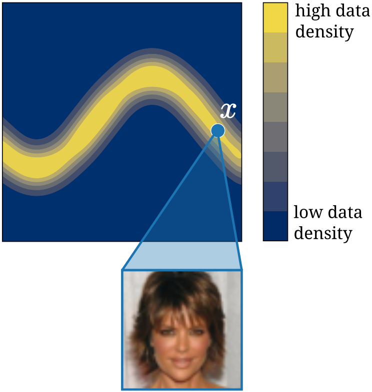
Deep learning is lacking a strong theoretical foundation:

- Neural networks are complicated functions
- The training process is stochastic

Geometry can help to alleviate these problems.



Geometry of the training data

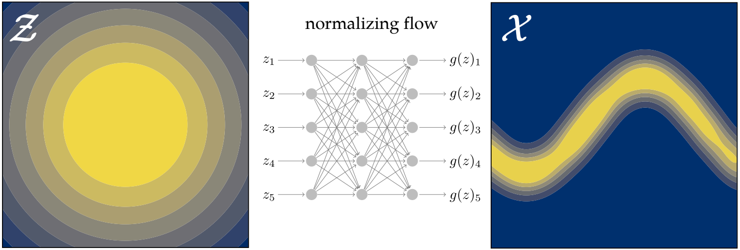


- Manifold Hypothesis: Data lies on low-dim. submanifold of high-dim. input space
- E.g. MNIST pictures lie on $\sim$ 30-dim. submanifold of $28 \times 28 = 784$ dim. input space

Learned diffeomorphisms

[Dinh, Krueger, Bengio, 2014]

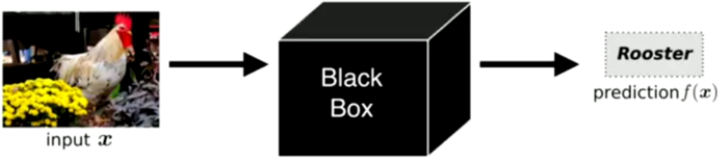
- How to characterize the data manifold?
- Can learn a diffeomorphism between a simple distribution and data distribution



- Diffeomorphism is given by another neural network, a ***normalizing flow***
- Get access to the data manifold in a functional form
- Connections to shape matching [Jansson, Modin, 2022]
- Connections to optimal transport [Chen, Karlsson, Ringh, 2021] [Bauer, Joshi, Modin, 2017] [Onken, Fung, Li, Ruthotto 2020]

Explainable AI (XAI)

[Review: Samek et al. 2021]

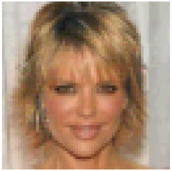


- Neural network classifiers lack inherent interpretability
- This is in contrast to more traditional methods like linear- or physical models
- For safety-critical applications this poses a serious challenge in practice
- Research progress can also be impeded
- Need explanations which provide insight into the neural network decisions

Counterfactual explanations

- ***Counterfactual of a sample***: Data point close to original but with different classification
- Difference between original and counterfactual reveals features which led to classification
- Example from CelebA dataset, classified as not-blonde:

original x



counterfactual x'

$|x - x'|$

Adversarial Examples

- Small perturbations can lead to misclassifications

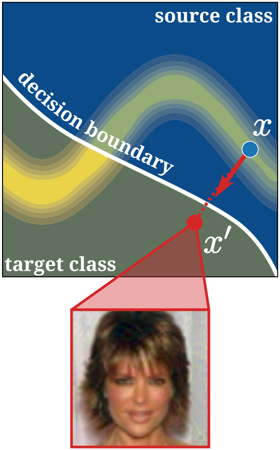
$p_{\text{blonde}}\Bigg($  $\Bigg) = 0.01$ but $p_{\text{blonde}}\Bigg($  $\Bigg) = 0.99$

Adversarial Examples

- Small perturbations can lead to misclassifications

p_{blonde} () $= 0.01$ but p_{blonde} () $= 0.99$

- Reason: Classifier only trained on the data manifold

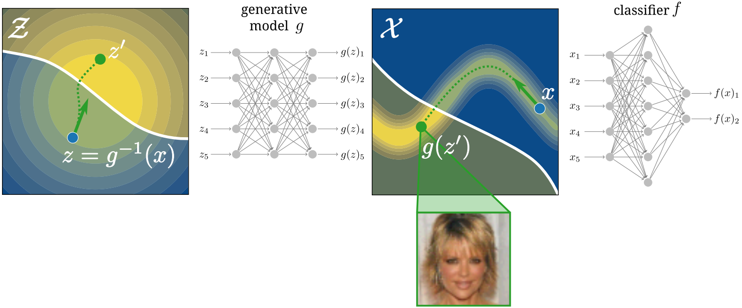


- Can use normalizing flows to optimize along the data manifold $\Rightarrow$ ***counterfactuals***

original
not blond

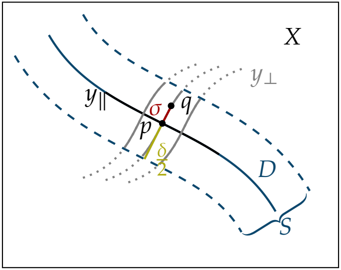
counterfactual
blond (p ≈ 0.99)

adversarial example
blond (p ≈ 0.99)



$$x^{(i+1)} = x^{(i)} + \lambda\, \gamma^{-1} \frac{\partial f_t}{\partial x}(x^{(i)}) + \mathcal{O}(\lambda^2)$$

Data coordinates



- Assume that data lies in a region $S = \mathrm{supp}(p)$ around data manifold D, in data coordinates x^α

$$S_x = \left\{ x_D + x_\delta \,\middle|\, x_D \in D_x,\ x_\delta^\alpha \in \left(-\frac{\delta}{2}, \frac{\delta}{2} \right) \right\}$$

with $\delta \ll 1$.

- Define normal coordinates y^μ in a neighborhood of D

Gradient ascent in y-coordinates

- By choosing $\{n_i\}$ orthogonal wrt γ, the inverse induced metric takes the form

$$\gamma^{\mu\nu}(y) = \begin{pmatrix} \gamma_D^{-1}(y) & & & \\ & \gamma_{\perp_1}^{-1} & & \\ & & \ddots & \\ & & & \gamma_{\perp_{N_X - N_D}}^{-1} \end{pmatrix}^{\mu\nu} .$$

- The gradient ascent update $g^{\alpha}(z^{(i+1)}) = g^{\alpha}(z^{(i)}) + \lambda\, \gamma^{\alpha\beta} \frac{\partial f_t}{\partial x^{\beta}} + \mathcal{O}(\lambda^2)$ becomes

$$\gamma^{\alpha\beta} \frac{\partial f_t}{\partial x^{\beta}} = \frac{\partial x^{\alpha}}{\partial y_{\parallel}^{\mu}} \gamma_D^{\mu\nu} \frac{\partial f_t}{\partial y_{\parallel}^{\nu}} + \frac{\partial x^{\alpha}}{\partial y_{\perp}^{i}} \gamma_{\perp_i}^{-1} \frac{\partial f_t}{\partial y_{\perp}^{i}}$$

- For $\gamma_{\perp_i}^{-1} \to 0$ and $\frac{\partial x}{\partial y_{\perp}}$ bounded we have

$$\gamma^{\alpha\beta} \frac{\partial f_t}{\partial x^{\beta}} \to \frac{\partial x^{\alpha}}{\partial y_{\parallel}^{\mu}} \gamma_D^{\mu\nu} \frac{\partial f_t}{\partial y_{\parallel}^{\nu}}$$

  and hence the update step points along the data manifold.

  $\Rightarrow$ ***In this case, obtain counterfactuals, not adversarial examples!***

The induced metric for well-trained generative models

Theorem (Diffeomorphic Counterfactuals)

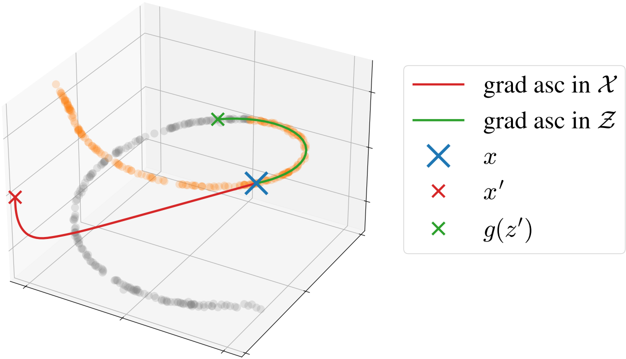
For $\epsilon \in (0, 1)$ and g a normalizing flow with Kullback–Leibler divergence $\mathrm{KL}(p, q) < \epsilon$,

$$\gamma_{\perp_i}^{-1} \longrightarrow 0 \qquad \text{as} \qquad \delta \longrightarrow 0$$

for all $i \in \{1, \ldots, N_X - N_D\}$.

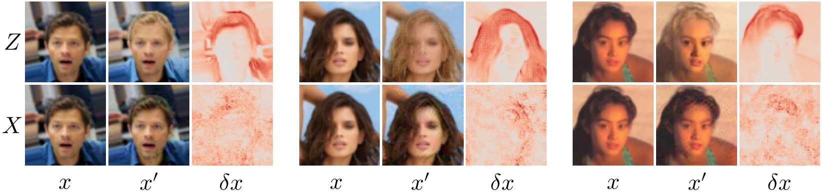
$\Rightarrow$ ***For well-trained generative models, the gradient ascent update in Z stays on the data manifold***

Toy example



Diffeomorphic Counterfactuals with CelebA

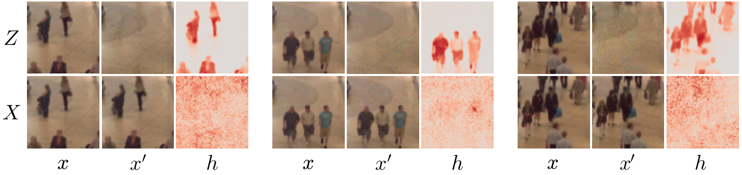
- Classifier: Binary CNN trained on ***blonde/not blonde*** attribute (test accuracy: 94%)
- Flow: Glow [Kingma et al., NeurIPS 2018]
- Task: Change classification from ***not blonde*** to ***blonde***



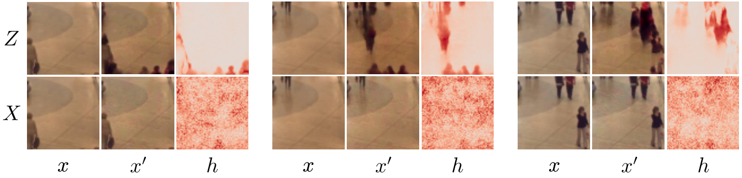
- Top row: Counterfactual computed in base space
- Bottom row: Adersarial example computed in data space

Diffeomorphic Counterfactuals for regression

- Consider crowd-counting dataset of mall images
- Count number of people in the image [Ribera et al., CVPR 2019]
- Flow: Glow [Kingma et al., NeurIPS 2018]
- Optimize for low number of people



- Optimize for high number of people



Conclusions

- Deep learning is a transformative technology lacking a strong theoretical basis

Conclusions

- Deep learning is a transformative technology lacking a strong theoretical basis
- Geometry can help in several key parts of the learning process, bringing abstract mathematics to practical applications
  - Counterfactuals explain black-box classifiers by providing a realistic sample close to the original but with a different classification
  - Naive gradient ascent leads off the data manifold, yielding adversarial examples
  - Gradient ascent in the base space of a normalizing flow leads to optimization on the data manifold
  - Concept be used for a wide range of different problems

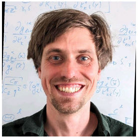
Conclusions

- Deep learning is a transformative technology lacking a strong theoretical basis
- Geometry can help in several key parts of the learning process, bringing abstract mathematics to practical applications
  - Counterfactuals explain black-box classifiers by providing a realistic sample close to the original but with a different classification
  - Naive gradient ascent leads off the data manifold, yielding adversarial examples
  - Gradient ascent in the base space of a normalizing flow leads to optimization on the data manifold
  - Concept be used for a wide range of different problems
- New areas of mathematics enter the study of neural networks

Collaborators


Daniel Persson


Klaus-Robert Müller


Pan Kessel


Christoffer Petersson


Fredrik Ohlsson


Hampus Linander


Ann-Kathrin Dombrowsik


Oscar Carlsson


Jimmy Aronsson

Thank you!