META-LEARNING AND ONE-SHOT LEARNING

NIPS 2016 TAKE-AWAYS

Olof Mogren

Chalmers University of Technology

2017-02-02

COMING SEMINARS

- Today: Olof Mogren
 Meta-learning and one-shot learning
- February 9: Devdatt Dubhashi
- February 16: **Up for grabs! Get in touch:** mogren@chalmers.se



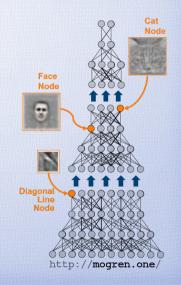
http://www.cse.chalmers.se/research/lab/seminars/

http://mogren.one/



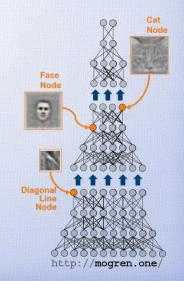
• But they were engineered by hand

• ...

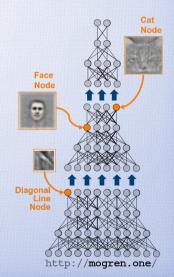


- But they were engineered by hand
- Along came feature learning

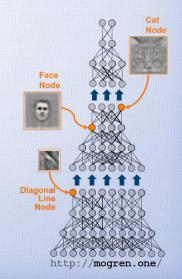
.



- But they were engineered by hand
- Along came feature learning
- Deep NNs learn a larger part of the problem



- But they were engineered by hand
- Along came feature learning
- Deep NNs learn a larger part of the problem
- Can we learn more?



• "Learning to learn"



- "Learning to learn"
- Schmidthuber 1987, 1992, 1993 nets modifying their own weights



- "Learning to learn"
- Schmidthuber 1987, 1992, 1993 nets modifying their own weights
- Schmidthuber 1997 The Success Story Algorithm

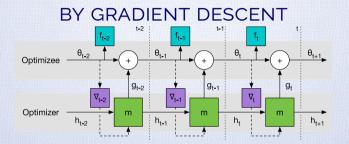


- "Learning to learn"
- Schmidthuber 1987, 1992, 1993 nets modifying their own weights
- Schmidthuber 1997 The Success Story Algorithm
- Daniel, et.al. 2016 Reinforcement learning

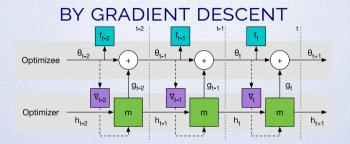


- "Learning to learn"
- Schmidthuber 1987, 1992, 1993 nets modifying their own weights
- Schmidthuber 1997 The Success Story Algorithm
- Daniel, et.al. 2016 Reinforcement learning
- Santoro, et.al. 2016, **Vinyals, et.al.** 2016 One-shot learning, mem-augmented NNs (multi-task learning is generalization)

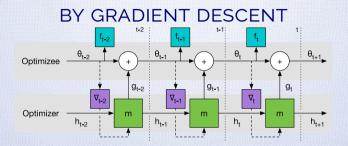




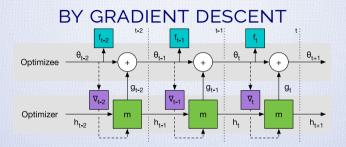
A learned coordinate-wise optimizer



- A learned coordinate-wise optimizer
- Two-layered LSTM net



- A learned coordinate-wise optimizer
- Two-layered LSTM net
- Inputs: optimizee gradient for one coordinate, optimizer state



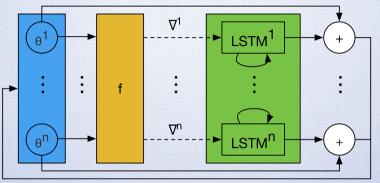
- A learned coordinate-wise optimizer
- Two-layered LSTM net
- Inputs: optimizee gradient for one coordinate, optimizer state
- Outputs: update for the specific coordinate

BY GRADIENT DESCENT

$$\mathcal{L}(\phi) = \mathbb{E}_f\left[\sum_{t=1}^T w_t f(heta_t)
ight]$$

where $\theta_{t+1} = \theta_t + g_t$, (1) $\begin{bmatrix} g_t \\ h_{t+1} \end{bmatrix} = m(\nabla_t, h_t, \phi)$.

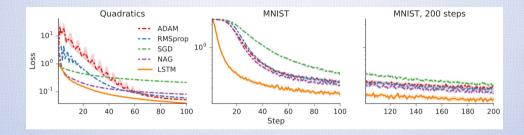
COORDINATE-WISE LSTM OPTIMIZER



LSTM¹...LSTMⁿ have shared weights, but separate hidden states.

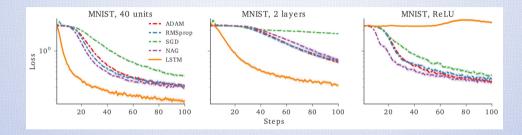
Andrychowicz, et.al., Learning to learn by gradient descent by gradient descent (NIPS 2016) http://mogren.one/

LEARNING CURVES



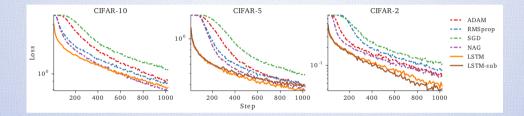
Andrychowicz, et.al., Learning to learn by gradient descent by gradient descent (NIPS 2016)

GENERALIZATION, MODEL LAYOUT



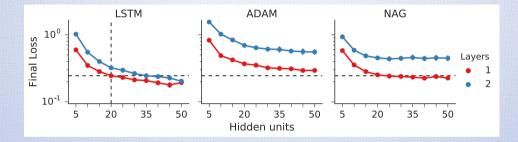
Andrychowicz, et.al., Learning to learn by gradient descent by gradient descent (NIPS 2016)

GENERALIZATION, CIFAR-10/5/2



Andrychowicz, et.al., Learning to learn by gradient descent by gradient descent (NIPS 2016)

GENERALIZATION, MODEL SIZE



Andrychowicz, et.al., Learning to learn by gradient descent by gradient descent (NIPS 2016)



• Meta-learning allows us to learn how to learn



- Meta-learning allows us to learn how to learn
- Generalize to new problems

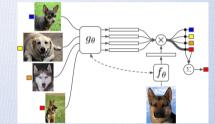
• How do we learn when we have very limited data?

- How do we learn when we have very limited data?
- Exampe: k-nearest neighbours (no learning/zero-shot)

- How do we learn when we have very limited data?
- Exampe: k-nearest neighbours (no learning/zero-shot)
- One-shot learning: learn from few examples

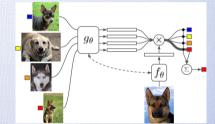
- How do we learn when we have very limited data?
- Exampe: k-nearest neighbours (no learning/zero-shot)
- One-shot learning: learn from few examples
- Also meta-learning; learn from one large training set how to make use of smaller data.

Related to metric learning



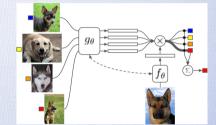
Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, Daan Wierstra (NIPS 2016)

- Related to metric learning
- Deep neural features



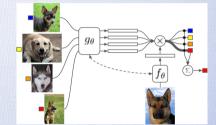
Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, Daan Wierstra (NIPS 2016)

- Related to metric learning
- Deep neural features
- Small labelled support set S,



Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, Daan Wierstra (NIPS 2016)

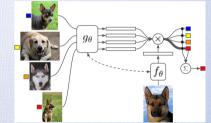
- Related to metric learning
- Deep neural features
- Small labelled support set S,
- Larns to map S to a cassifier c(x).



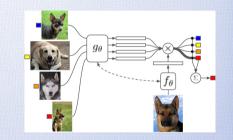
Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, Daan Wierstra (NIPS 2016)

- Related to metric learning
- Deep neural features
- Small labelled support set S,
- Larns to map S to a cassifier c(x).
- *S* may contain unseen classes!

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, Daan Wierstra (NIPS 2016)

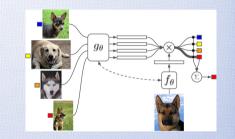


• $S = (x_i, y_i)_{i=1}^k$



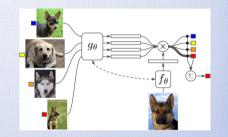
Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, Daan Wierstra (NIPS 2016)

• $S = (x_i, y_i)_{i=1}^k$ • $\hat{y} = \sum_{i=1}^k a(\hat{x}, x_i) y_i$



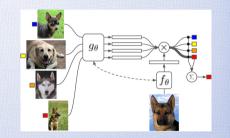
Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, Daan Wierstra (NIPS 2016)

- $S = (x_i, y_i)_{i=1}^k$
- $\hat{y} = \sum_{i=1}^{k} \alpha(\hat{x}, x_i) y_i$
- (If a is a kernel, then this is a kernel density estimator)



Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, Daan Wierstra (NIPS 2016)

- $S = (x_i, y_i)_{i=1}^k$
- $\hat{y} = \sum_{i=1}^{k} a(\hat{x}, x_i) y_i$
- (If a is a kernel, then this is a kernel density estimator)
- Subsumes both KDE and kNN

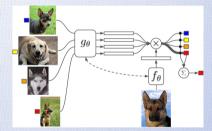


Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, Daan Wierstra (NIPS 2016)

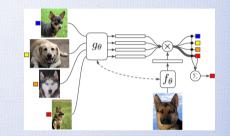
http://mogren.one/

- $S = (x_i, y_i)_{i=1}^k$
- $\hat{y} = \sum_{i=1}^{k} \alpha(\hat{x}, x_i) y_i$
- (If *a* is a kernel, then this is a kernel density estimator)
- Subsumes both KDE and kNN
- $a(\hat{x}, x_i) = \frac{e^{c(f(\hat{x}), g(x_i))}}{7}$

(Softmax of cosine sim) Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, Daan Wierstra (NIPS 2016)

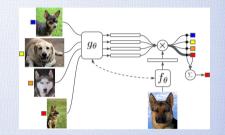


• $a(\hat{x}, x_i) = \frac{e^{cos(f(\hat{x}), g(x_i))}}{Z}$ (Softmax of cosine sim)



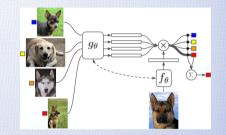
Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, Daan Wierstra (NIPS 2016)

- $a(\hat{x}, x_i) = \frac{e^{cos(f(\hat{x}), g(x_i))}}{\overline{Z}}$ (Softmax of cosine sim)
 - f (embedding of new instances)



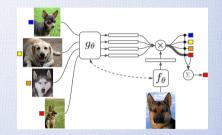
Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, Daan Wierstra (NIPS 2016)

- $a(\hat{x}, x_i) = \frac{e^{\cos(f(\hat{x}), g(x_i))}}{Z}$ (Softmax of cosine sim)
 - f (embedding of new instances)
 - g (embedding of support set)



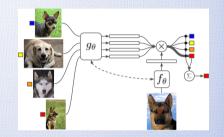
Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, Daan Wierstra (NIPS 2016)

- $a(\hat{x}, x_i) = \frac{e^{\cos(f(\hat{x}), g(x_i))}}{Z}$ (Softmax of cosine sim)
 - f (embedding of new instances)
 - g (embedding of support set)
 - VGG, Inception, word embeddings



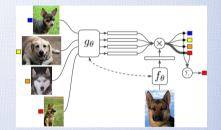
Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, Daan Wierstra (NIPS 2016)

- $a(\hat{x}, x_i) = \frac{e^{\cos(f(\hat{x}), g(x_i))}}{Z}$ (Softmax of cosine sim)
 - f (embedding of new instances)
 - g (embedding of support set)
 - VGG, Inception, word embeddings
 - $f(\hat{x}, S) = LSTM_{attention}(f'(\hat{x}), g(S), K)$



Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, Daan Wierstra (NIPS 2016)

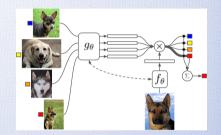
- $a(\hat{x}, x_i) = \frac{e^{cos(f(\hat{x}), g(x_i))}}{Z}$ (Softmax of cosine sim)
 - f (embedding of new instances)
 - g (embedding of support set)
 - VGG, Inception, word embeddings
 - $f(\hat{x}, S) = LSTM_{attention}(f'(\hat{x}), g(S), K)$



• $\theta = \operatorname{argmax}_{\theta} E_{L \sim T} \left[E_{S \sim L, B \sim L} \left[\sum_{(x, y) \in B} \log P(y | x, S) \right] \right]$

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, Daan Wierstra (NIPS 2016)

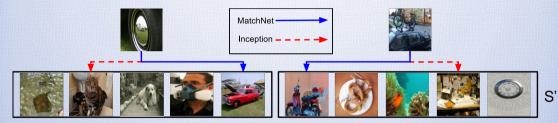
- $a(\hat{x}, x_i) = \frac{e^{\cos(f(\hat{x}), g(x_i))}}{Z}$ (Softmax of cosine sim)
 - f (embedding of new instances)
 - g (embedding of support set)
 - VGG, Inception, word embeddings
 - $f(\hat{x}, S) = LSTM_{attention}(f'(\hat{x}), g(S), K)$



- $\theta = \operatorname{argmax}_{\theta} E_{L \sim T} \left[E_{S \sim L, B \sim L} \left[\sum_{(x, y) \in B} \log P(y | x, S) \right] \right]$
- (learns to learn from a support set).

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, Daan Wierstra (NIPS 2016)

MATCHING NETWORKS: EXAMPLES



Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, Daan Wierstra (NIPS 2016)

 Search in molecular space is challenging; large, discrete, and unstructured

D, D,)-()-() blordordordordordor-()-14 ya ya sa ar Brod propo III droga De to brand ba a to brand by and by individing proposed you ~ & by XY~ tate an thand had a sea 20 pt pt (too ~ ~ ha hand bet baxy gart xy gxy gxy g to x, a ha x, a b, x b, x b, x b, x ha b, x x, a b, x 1 Stor Stor bixx d Dix + to yandare Noncon Str Xxom botho X x Norchen & Julienanan

- Search in molecular space is challenging; large, discrete, and unstructured
- Variational autoencoder

D- D-)-()-() by ord ord ord ord ord ()-()-()-1479 72 20 Dr B wa propo III drogo De to an mar ad ta Dr at the por a po individing proposed you ~ & by XY~ tata axhahabdax ga zapa pa loo ~ ~+d+a+dbdbdbdxrdartxrdxraxra to x, a ha x, a b, x b, x b, x b, x ha b, x x, a b, x 1 from francor francor -ing ~~~~ d lt ~~ & tod by xx, & By x By x p/x p/+ 1/2 Noncon Str Xxom botho X x Norchen & Julienanan

- Search in molecular space is challenging; large, discrete, and unstructured
- Variational autoencoder
- Convert discrete representations to and from continuous

1479 72 20 Dr B wa propo III drogo De to be a sold of a sold of a bor a sold of individina proportion for many by XY~ tata axhahabdax ga zapa pa loo ~ ~ hd + a + d bet bd xyd art xyd xya xya L to x, a ha x, a b, x b, x b, x b, x ha b, x x, a b, x 1 from francor francor -ing ~~~~ d lt ~~ & tod by xx, & By x By x p/x p/+ 1/2 Noncon Str Xxom botho X x Norchen & Julian

- Search in molecular space is challenging; large, discrete, and unstructured
- Variational autoencoder
- Convert discrete representations to and from continuous
- Optimize molecule properties

1479 72 20 Dr B wa propo III drogo De to an mar ad ta Dr at the por mar to individina proportion for many by XY~ tata axhahabdax ga zapa pa loo ~ ~+d+a+dbdbdbdxrdartxrdxraxra to x, a ha x, a b, x b, x b, x b, x ha b, x x, a b, x 1 from francor francor -ing ~~~~ d lt ~~ & tod by xx, & By x By x p/x p/+ 1/2 Noncon Str Xxom botho X x Norchen & Julienanan

- Search in molecular space is challenging; large, discrete, and unstructured
- Variational autoencoder
- Convert discrete representations to and from continuous
- Optimize molecule properties
- Best paper award at Constructive machine learning workshop at NIPS

1479 72 20 Dr B wa propo III drogo De to be a sold of a sold of a bor a sold of ord ond ond pro pag tod 30 ~~ & bax XY~ tate an xhand bd ax ga zapa pa l ba ~ ~ ha hand bet baxy gart xy gxy gxy g to x, a ha x, a b, x b, x b, x b, x ha b, x x, a b, x 1 from front and and the for the -ing ~~~~ d lt ~~ & tod by xx, & By x By x p/x p/+ 1/2 Noncon Str Xxom botho X x Norchen & Julienanan

Variational autoencoder, seq2seq

Gómez-Bombarelli, et.al., 2016

~O~O pro pro pro pro pro pro Prd ()~()~(,~Q ,~Q $\sim_{-} \bigcirc = \cdots$ ord ord ord ord ord $p_{\alpha'} = \cdots$ 1479 79 20 or Brod propro III drogo De to an an walka va to may to An - Pa ord ond ond pro pag tod 30 ~~ & bdx XY~ tata and had had a sea so se it bo ~ ~ ha + a + a b t b a x of a + x of x of x of x to x, a ha x, a b, x b, x b, x b, x ha b, x x, a b, x 1 from francista and the galance -ing ~~~~ d lt ~~ for the by xx, to by x By x by + the Noncon fr Xxom botho X-x Norchen & Julian

- Variational autoencoder, seq2seq
- String representation of molecules: SMILES

Gómez-Bombarelli, et.al., 2016

~O~O~Opopropropropro RdO~O~~~ $\sim_{-} \bigcirc = \cdots$ ord ord ord ord ord $p_{\alpha'} = \cdots$ 1479 79 20 or Brod propro III drogo De to be a work of the or the bor a bor and the ord ond ond pro pag tod 30 ~~ & bdx XY~ tata and had had a sea so se it bo ~ ~+d+a+dbd+bdxydar+xydxyaxya to x, a ha x, a b, x b, x b, x b, x ha b, x x, a b, x 1 from francor francor -ing ~~~~ d lt ~~ for the by xx, to by x By x by + h to Noncon fr Xxom botho X-x Norchen & Julian

- Variational autoencoder, seq2seq
- String representation of molecules: SMILES
- Decode random vectors

Gómez-Bombarelli, et.al., 2016

~O~O pro pro pro pro pro pro Prd ()~()~(,~Q ,~Q ~... m m and and and and par = " ~ ~... ~. 1479 79 20 or Brod propro III drogo De to an an walka va to may to An - Pa ord ond ond pro pag tod 30 ~~ & bdx XY~ tata and had had a sea so se it bo ~ ~+d+a+dbd+bdxydar+xydxyaxya to x, a ha x, a b, x b, x b, x b, x ha b, x x, a b, x 1 from francor francor -ing ~~~~ d lt ~~ for the by xx, to by x By x by + h to Noncon fr Xxom botho X-x Norchen & Julian

- Variational autoencoder, seq2seq
- String representation of molecules: SMILES
- Decode random vectors
- Perturb known chemical structures

Gómez-Bombarelli, et.al., 2016

~O~O pro pro pro pro pro pro Prd ()~()~(,~Q ,~Q ~... m m and and and and par = " ~ ~... ~. 1479 79 20 or Brod propro III drogo De to an an walka va to may to An - Pa ord ond ond pro pag tod 30 ~~ & bdx XY~ tataaxhahabaaxga zapa at to ~ ~+d+a+dbd+bdxydar+xydxyaxya to x, a ha x, a b, x b, x b, x b, x ha b, x x, a b, x 1 from francor francor -ing ~~~~ d lt ~~ for the by xx, to by x By x by + h to Noncon fr Xxom botho X-x Norchen & Julian

- Variational autoencoder, seq2seq
- String representation of molecules: SMILES
- Decode random vectors
- Perturb known chemical structures
- Interpolate between molecules

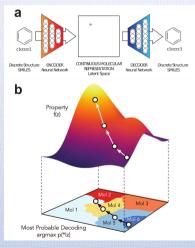
Gómez-Bombarelli, et.al., 2016

~O~O pro pro pro pro pro pro Prd ()~()~(,~Q ,~Q ~... m m and and and and par = " ~ ~... ~. 1479 79 20 or Brod propro III drogo De to an an vale of the or the bar wand of ord ond ond pro pag tod 30 ~~ & bdx XY~ tata and had had a sea so se it bo ~ ~ ha + a + a b t b a x of a + x of x of x of x to x, a ha x, a b, x b, x b, x b, x ha b, x x, a b, x 1 Stody & to dix x d Dix + to yalance -ing ~~~~ d lt ~~ for the by xx, to by x By x by + h to Noncon fr Xxom botho X-x Norchen & Julian

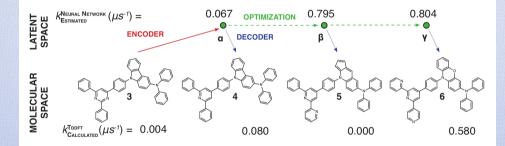
- Variational autoencoder, seq2seq
- String representation of molecules: SMILES
- Decode random vectors
- Perturb known chemical structures
- Interpolate between molecules
- Train a model to predict medical properties based on representation

Gómez-Bombarelli, et.al., 2016

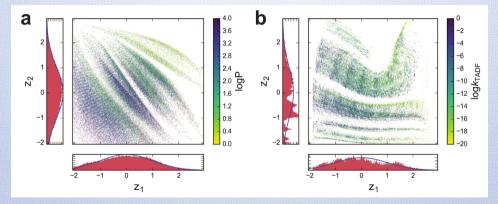
~... m m and and and and par = " ~ ~... ~. 1479 79 20 or Brod propro III drogo De to an an vale of the or the bar wand of ord ond ond pro pag tod 30 ~~ & bdx XY~ tata and had had a sea so se it bo ~ ~ ha + a + a b t b a x of a + x of x of x of x to x, a ha x, a b, x b, x b, x b, x ha b, x x, a b, x 1 Stody & to dix x d Dix + to yalance -ing ~~~~ d lt ~~ for the by xx, to by x By x by + the Noncon fr Xxom botho X-x Norchen & Julian



Gómez-Bombarelli, et.al., 2016



Gómez-Bombarelli, et.al., 2016



Gómez-Bombarelli, et.al., 2016

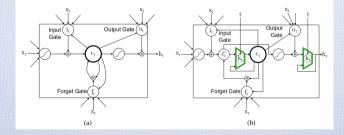
mogren@chalmers.se

http://mogren.one/



APPENDIX

PHASED LSTM



PHASED LSTM: STATE VISUALIZATION

