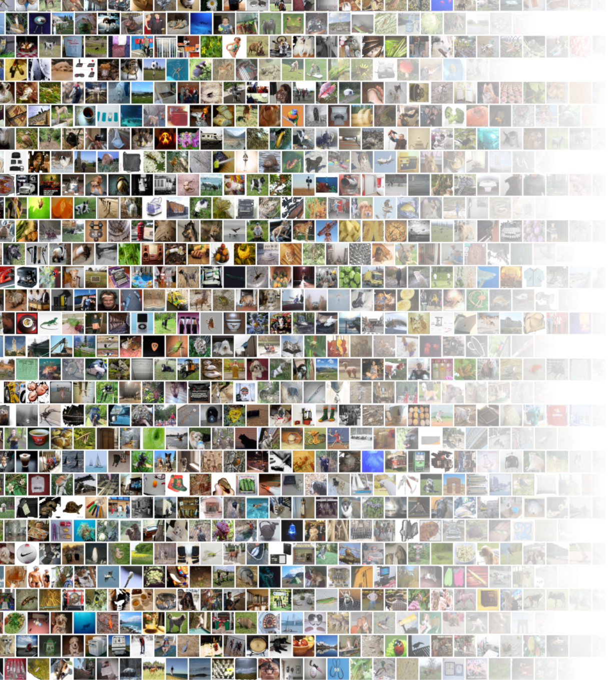


Transfer and privacy

MLDS GBG Meetup, Nov. 2019

Olof Mogren, Research institutes of Sweden

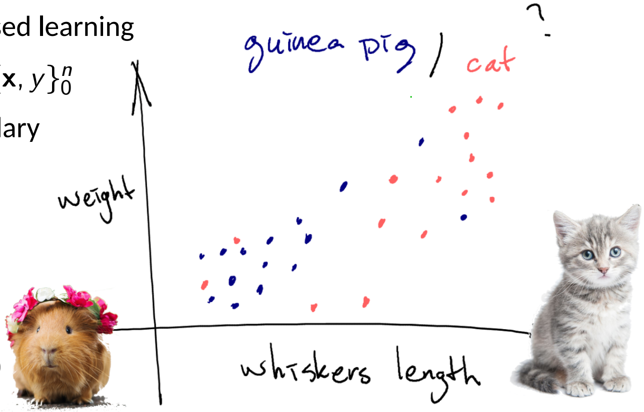
AI at RISE

- STHLM, GBG, LKPG, V-ås, Luleå, Lund
- Research projects
  - Industry
  - Public authorities
  - Academia
- Gothenburg deep learning group
  - Machine learning seminars
    Every **Thursday at 15** *
    Lindholmspiren 3A
    Open to the public

* *14/11: John Martinsson; Adversarial privacy*

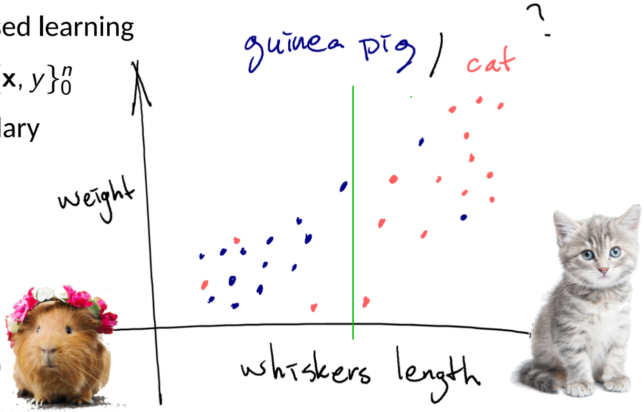
Foundations of machine learning

- Today: supervised learning
- Training data: $\{\mathbf{x}, y\}_0^n$
- Decision boundary
- Underfitting
- Overfitting
- Generalization
- No free lunch (Wolpert, 1996)



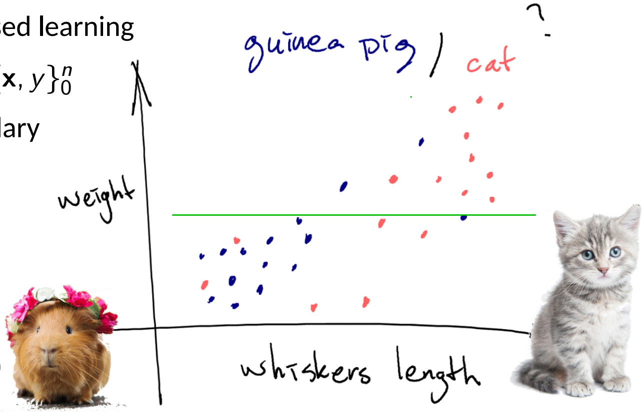
Foundations of machine learning

- Today: supervised learning
- Training data: $\{\mathbf{x}, y\}_0^n$
- Decision boundary
- Underfitting
- Overfitting
- Generalization
- No free lunch (Wolpert, 1996)



Foundations of machine learning

- Today: supervised learning
- Training data: $\{\mathbf{x}, y\}_0^n$
- Decision boundary
- Underfitting
- Overfitting
- Generalization
- No free lunch (Wolpert, 1996)



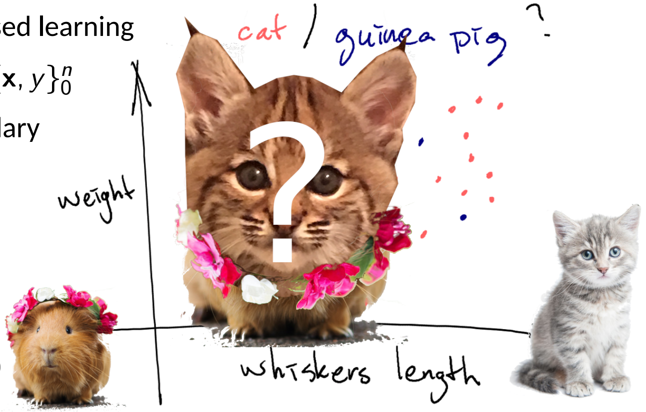
Foundations of machine learning

- Today: supervised learning
- Training data: $\{\mathbf{x}, y\}_0^n$
- Decision boundary
- Underfitting
- Overfitting
- Generalization
- No free lunch (Wolpert, 1996)

RISE

Foundations of machine learning

- Today: supervised learning
- Training data: $\{\mathbf{x}, y\}_0^n$
- Decision boundary
- Underfitting
- Overfitting
- Generalization
- No free lunch (Wolpert, 1996)



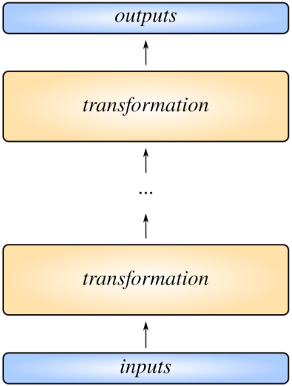
http://mogren.one/

Foundations of machine learning

- Today: supervised learning
- Training data: $\{\mathbf{x}, y\}_0^n$
- Decision boundary
- Underfitting
- Overfitting
- Generalization
- No free lunch (Wolpert, 1996)

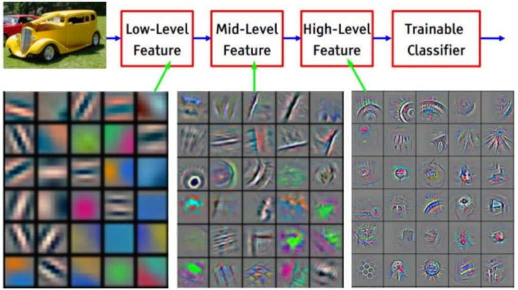
RISE

Deep learning

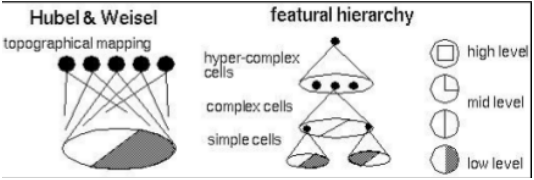


- Sequence of transformations
- Learning to compute representations
- Depth adds representation power
- (Zero hidden layers $\rightarrow$ linear model)

Levels of abstractions



Feature visualization of convolutional net trained on ImageNet from [Zeiler & Fergus 2013]



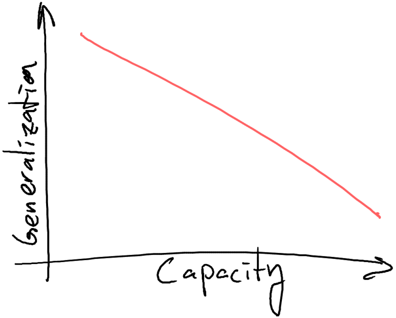
enough units

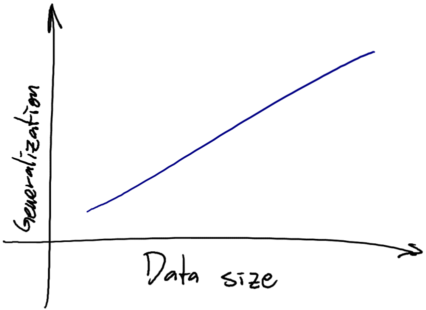
$\downarrow$

universal approximation

Universal approximation theorem: A feed forward net with enough hidden units can approximate any continuous function with arbitrary precision. (Balázs, et.al., 2001)

Remember





Require large amounts of training data

- High representational capacity $\rightarrow$ large data requirements

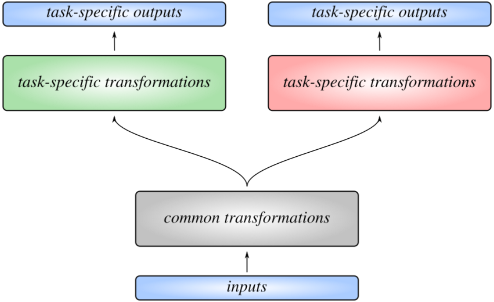
Choices when data is limited

- Go get some more!
- Data augmentation
- Generate synthetic data
- Use some other data

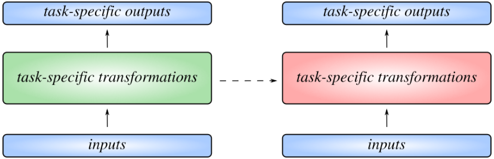
Multi-task learning

Transfer learning

One model, two tasks

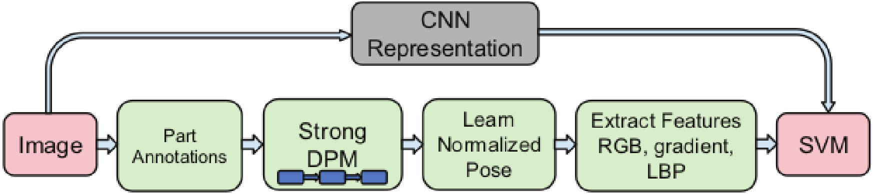


1. Pretrain 2. Fine-tune



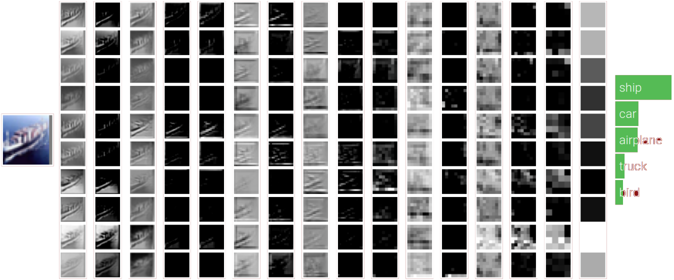
E.g., Un-, Semi-, Supervised

Imagenet pretraining (1/3)



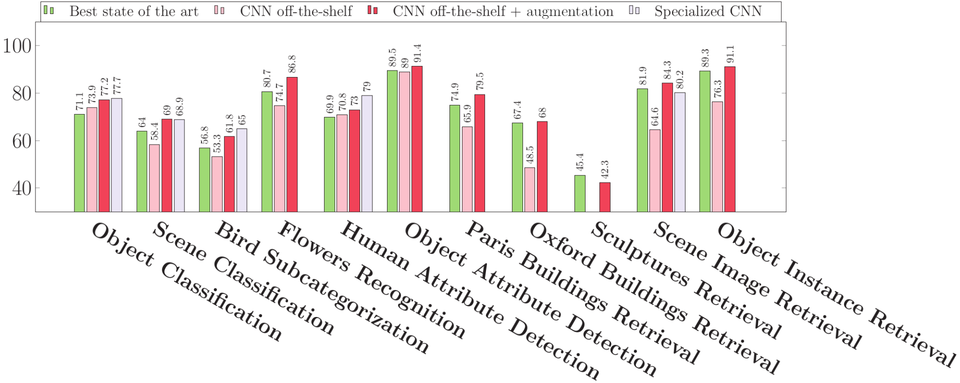
Razavian, et.al., 2014

Imagenet pretraining (2/3)



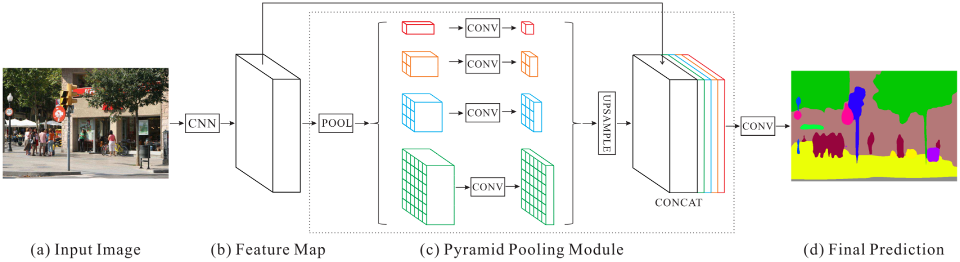
Razavian, et.al., 2014

Imagenet pretraining (3/3)



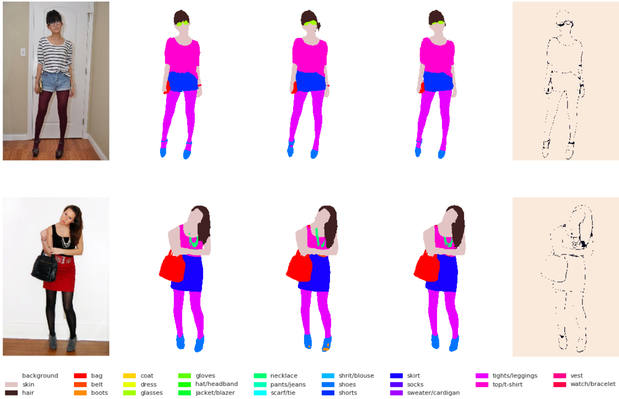
Razavian, et.al., 2014

Semantic segmentation



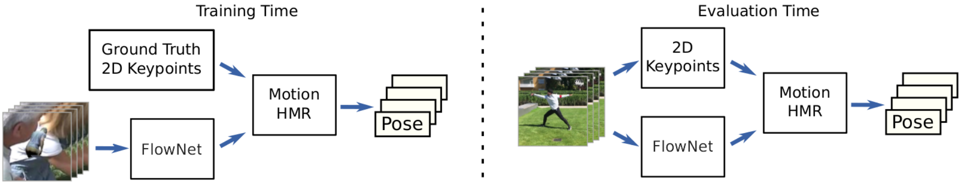
Pyramid Scene Parsing Network, Zhao, et.al., 2016

Semantic segmentation of fashion images



*Martinsson, **Mogren** (Extra)*

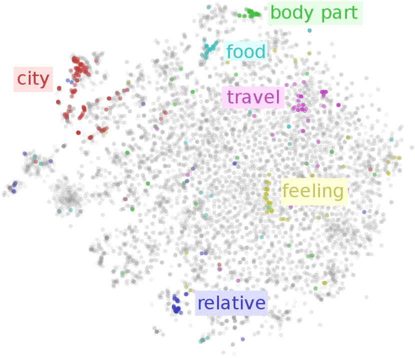
3D pose estimation



Doersch and Doersch, NeurIPS 2019 (Extra)

Transfer learning in language

- Natural language processing, NLP
  - Discrete data
  - Large sources of text available ((weekly or un-) annotated)
  - Embeddings
    - bag-of-words (Schütze, 1993)
    - word2vec (Mikolov, et.al., 2013)
    - Glove (Pennington, et.al., 2015)
  - End-to-end; not yet always



NLP, Transformers

- Deep transfer learning for language
- Transfer learning/unsupervised pretraining

Attention Is All You Need, Vaswani, et.al., 2017

Applying Transformers

- Representation learning, e.g. QA (Nadhan, Mondal, 2019)
- Finetuning, e.g. summarization of podcasts (Risne, Siitova, 2019)

Multilingual Transformers

- Multilingual BERT
- XLM-R
  - Pretrain on 100 languages
  - Fine-tune on one language
  - Improved performance on low-resouce languages

Conneau, et.al., 2019

RISE

Differential privacy; a definition

If the output from an algorithm **does not change much** with **small** changes in the input dataset, the algorithm is **differentially private**.

Age	Gender	BMI	Fever	Nause	Headacl	Diarrhea	Fatigue	Jaund	Epi	WBC	RBC	HGB	Plat	AST 1
56	1	35	2	1	1	1	2	2	2	7425	4248807	14	112132	99
46	1	29	1	2	2	1	2	2	1	12101	4429425	10	129367	91
57	1	33	2	2	2	2	1	1	1	4178	4621191	12	151522	113
49	2	33	1	2	1	2	1	2	1	6490	4794631	10	146457	43
59	1	32	1	1	2	1	2	2	2	3661	4606375	11	187684	99
58	2	22	2	2	2	1	2	2	1	11785	3882456	15	131228	66
42	2	26	1	1	2	2	2	2	2	11620	4747333	12	177261	78
48	2	30	1	1	2	2	1	1	2	7335	4405941	11	216176	119
44	1	23	1	1	2	2	2	1	2	10480	4608464	12	148889	93
45	1	30	2	1	2	2	1	1	2	6681	4455329	12	98200	55
37	2	24	2	1	2	1	2	2	1	4437	4265042	12	166027	103
36	1	22	2	2	1	1	1	1	1	6052	4130219	13	144266	75
45	2	25	2	1	1	1	2	1	2	9279	4116937	13	203003	97

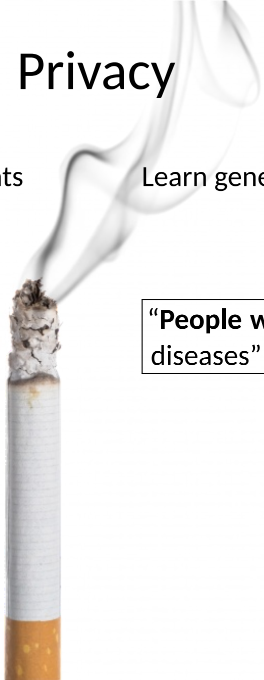
But. . .

ML models work by looking at data to learn patterns from it.

Privacy

Learn details about individual data points

Learn general patterns about data



“**Jane Smith** has a heart disease”

“**People who smoke** risk getting heart diseases”

Privacy

~~Learn details about individual data points~~

~~“**Jane Smith** has a heart disease”~~

Learn general patterns about data

“**People who smoke** risk getting heart diseases”

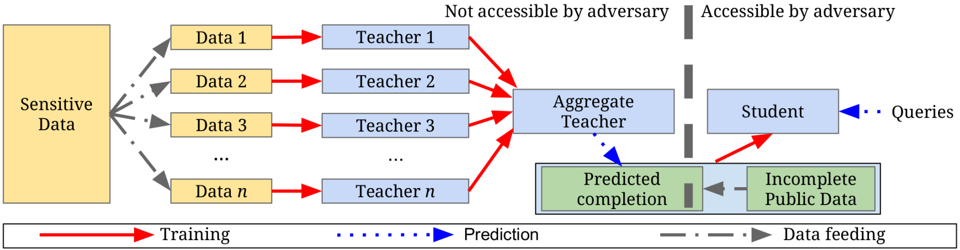
Does deep learning memorize data?

Good generalization

$\downarrow$

general patterns, not specific details

Private aggregation of teacher ensembles



1. Divide training set into disjoint parts
2. Train ensemble on parts; noisy voting
3. Train student with ensemble as oracle
4. Adversary can query student model

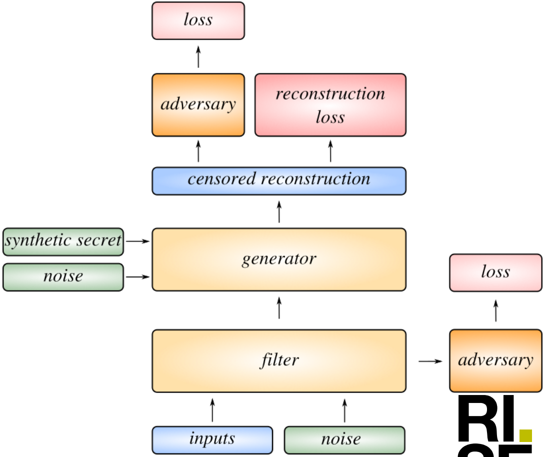
PATE, Papernot, et.al., 2016

Privacy

- Overfitting; memorizing specifics about the training data.
- Limiting overfitting can lead to improving privacy but this neat side-effect may not be enough in practice.

Adversarially learned privacy (1/2)

- Learn to fool adversary for sensitive attribute
- Produce sensitive attribute from population-level distribution



Martinsson, Listo Zec, Gillblad, Mogren, 2019

Adversarially learned privacy (2/2)

Top row: input. Middle row non-smiling output. Bottom: smiling output.

Martinsson, Listo Zec, Gillblad, Mogren, 2019

More adversarial representation learning on Thursday!