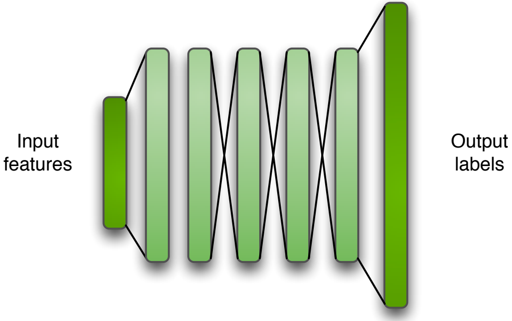
Multitask learning & related supervised methods

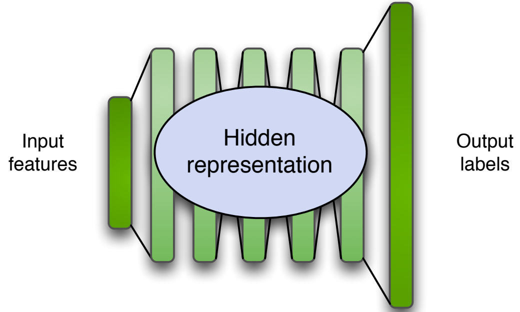
Peter Bell

Machine learning practical— MLP Lecture 11
25 January 2017

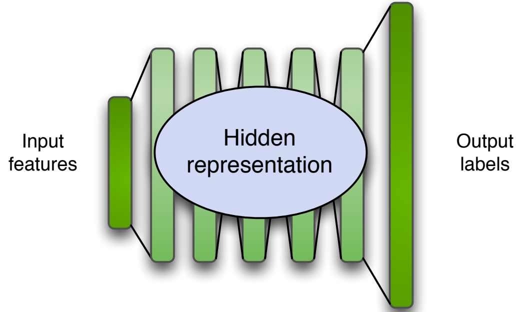
Learning hidden representations



Learning hidden representations

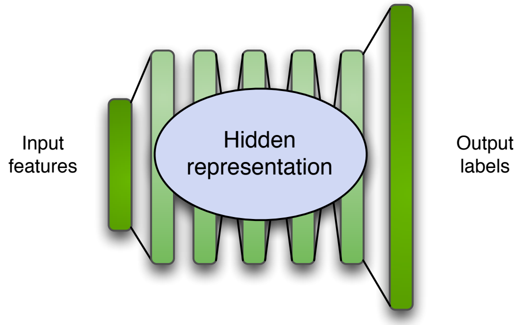


Learning hidden representations



- Higher layers of deep neural networks are assumed to learn increasingly more abstract representations of the data

Learning hidden representations



- Higher layers of deep neural networks are assumed to learn increasingly more abstract representations of the data
- Learning a good hidden representation enables the network to generalise well to unseen examples

Training deep neural networks

- Hard to find a good minimum when the training criterion is highly non-convex

Training deep neural networks

- Hard to find a good minimum when the training criterion is highly non-convex
- **Unsupervised** pre-training: start the optimisation in a "good" region of parameter space that describes observed (unlabelled) samples

Training deep neural networks

- Hard to find a good minimum when the training criterion is highly non-convex
- **Unsupervised** pre-training: start the optimisation in a "good" region of parameter space that describes observed (unlabelled) samples
- Alternatively – consider better methods of **supervised training**

The label problem

- Supervised training assumes we have a suitable label for each training sample

The label problem

- Supervised training assumes we have a suitable label for each training sample
- Even if the labels are hand-generated, and “correct”, there are problems:

The label problem

- Supervised training assumes we have a suitable label for each training sample
- Even if the labels are hand-generated, and "correct", there are problems:
    - Does the labelling describe all the important properties of the data?

The label problem

- Supervised training assumes we have a suitable label for each training sample
- Even if the labels are hand-generated, and "correct", there are problems:
    - Does the labelling describe all the important properties of the data?
    - Is it too simple?

The label problem

- Supervised training assumes we have a suitable label for each training sample
- Even if the labels are hand-generated, and "correct", there are problems:
  - Does the labelling describe all the important properties of the data?
  - Is it too simple?
  - Or too difficult to learn from a flat start?

The label problem

- Supervised training assumes we have a suitable label for each training sample
- Even if the labels are hand-generated, and "correct", there are problems:
  - Does the labelling describe all the important properties of the data?
  - Is it too simple?
  - Or too difficult to learn from a flat start?
  - Is it well-defined?

The label problem

- Supervised training assumes we have a suitable label for each training sample
- Even if the labels are hand-generated, and "correct", there are problems:
  - Does the labelling describe all the important properties of the data?
  - Is it too simple?
  - Or too difficult to learn from a flat start?
  - Is it well-defined?
- This lecture will explore these issues.

The rest of this lecture

- **Curriculum learning**
- Multitask learning
- Student-teacher models

Curriculum learning

- If a task is difficult, it may be hard to learn from scratch from a limited quantity of data

Curriculum learning

- If a task is difficult, it may be hard to learn from scratch from a limited quantity of data
- Learn how humans do – from simpler examples first, moving on to more complex examples

Curriculum learning

- If a task is difficult, it may be hard to learn from scratch from a limited quantity of data
- Learn how humans do – from simpler examples first, moving on to more complex examples
- Difficulty can be defined with respect to the entropy of the training data distribution

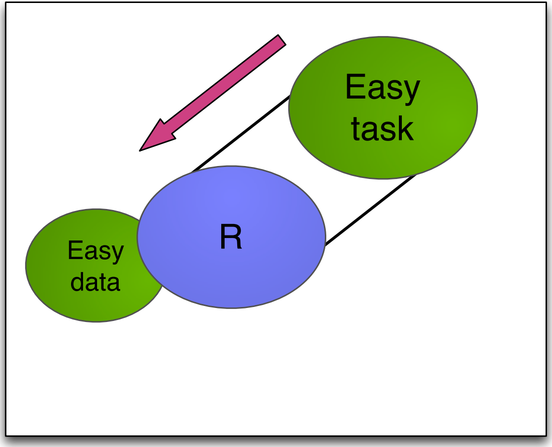
Curriculum learning

- If a task is difficult, it may be hard to learn from scratch from a limited quantity of data
- Learn how humans do – from simpler examples first, moving on to more complex examples
- Difficulty can be defined with respect to the entropy of the training data distribution
- Easier samples $\rightarrow$ less noise in the error signals

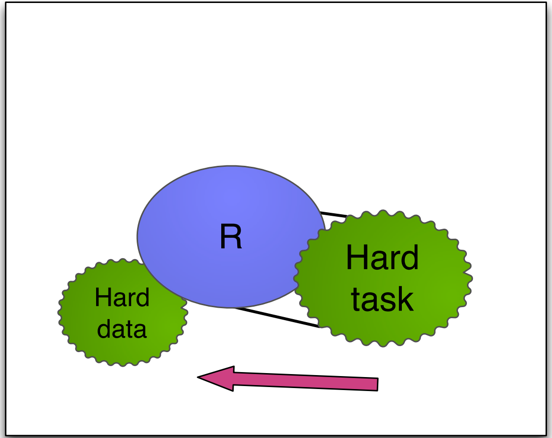
Curriculum learning

- If a task is difficult, it may be hard to learn from scratch from a limited quantity of data
- Learn how humans do – from simpler examples first, moving on to more complex examples
- Difficulty can be defined with respect to the entropy of the training data distribution
- Easier samples $\rightarrow$ less noise in the error signals
- Greater size of label space $\rightarrow$ samples harder to classify

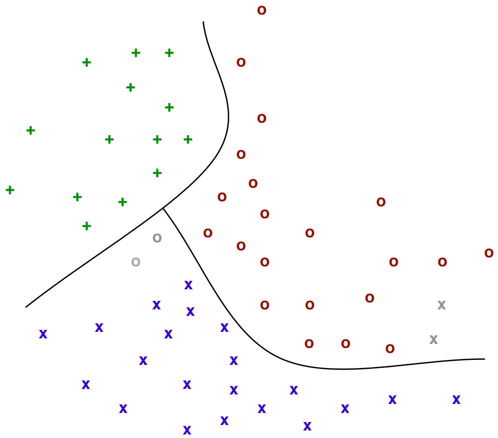
Train on easy data



Then train on harder data

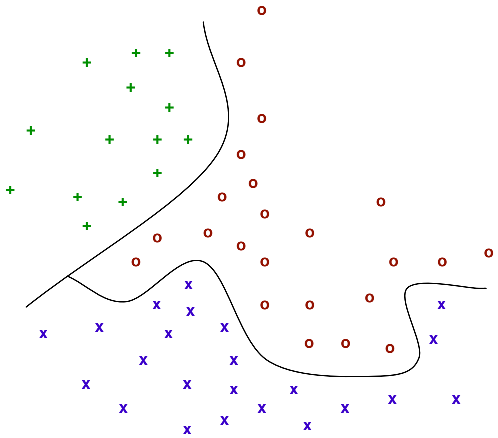


Learn approximate decision boundaries...



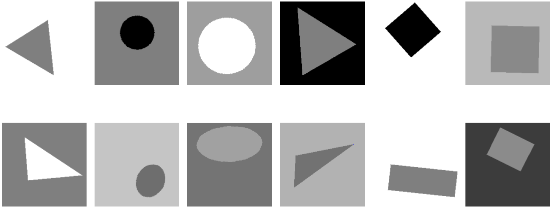
Harder data

Then learn fine-grained decision boundaries...



Examples

- Recognising shapes in images (toy example)



- Increasing the vocabulary of a language model
- In automatic speech recognition, modelling phonetic units with and without context

- Curriculum learning
- **Multitask learning**
- Student-teacher models

Motivation

- In machine learning, we normally break a complex problem down into tractable sub-problems, and learn to solve one problem at a time.

Motivation

- In machine learning, we normally break a complex problem down into tractable sub-problems, and learn to solve one problem at a time.
- This potentially ignores rich sources of information found in the training signals of other tasks

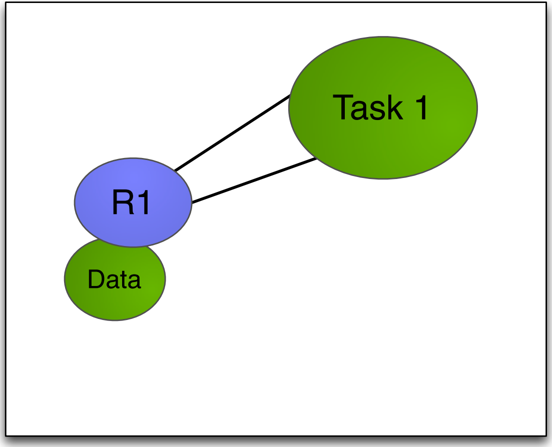
Motivation

- In machine learning, we normally break a complex problem down into tractable sub-problems, and learn to solve one problem at a time.
- This potentially ignores rich sources of information found in the training signals of other tasks
- Caruana [1997] proposed multitask learning as a means of **inductive transfer** between tasks

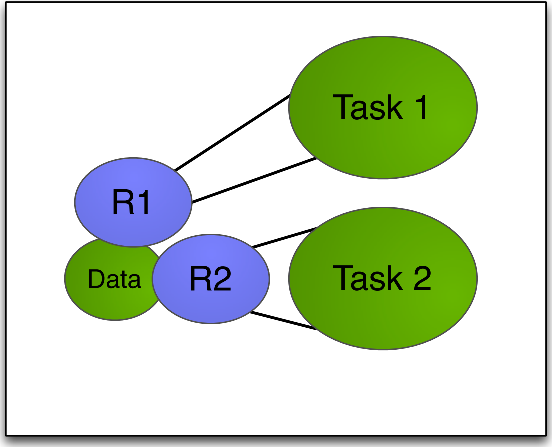
Motivation

- In machine learning, we normally break a complex problem down into tractable sub-problems, and learn to solve one problem at a time.
- This potentially ignores rich sources of information found in the training signals of other tasks
- Caruana [1997] proposed multitask learning as a means of **inductive transfer** between tasks
- This acts as a form of **bias**, causing the classifier to prefer hypotheses that explain more than one task, improving generalisation

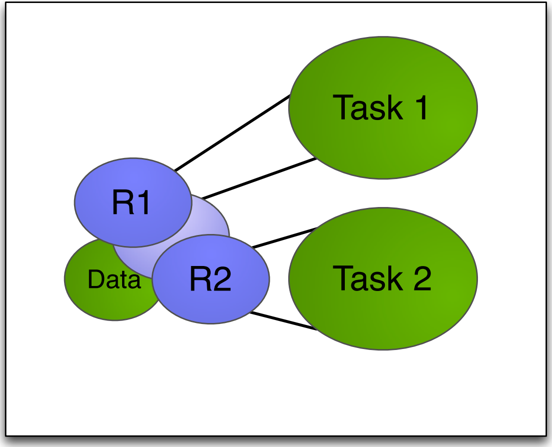
Multitask learning illustrated



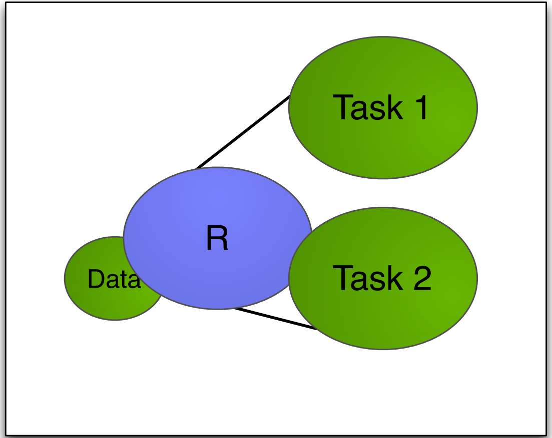
Add a related task...



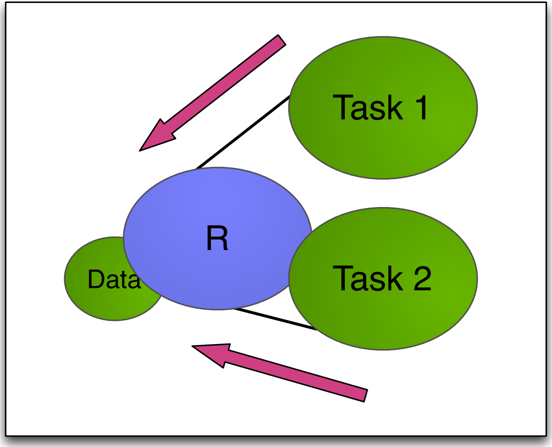
Representation may overlap...



Share the representation...



… and update with both error signals



Examples

Data	Primary task	Secondary task
Images	Face detections	Facial landmarks
Audio	Speech recognition	Speaker recognition
Biological	Gene expression	?

Often autoencoding is used as a secondary task.

Why does it work?

- **Data amplification** to minimise the effect of noise in the training signals

Why does it work?

- **Data amplification** to minimise the effect of noise in the training signals
- Better **selection** of shared hidden representations, reducing the effect of irrelevent inputs when there is limited data

Why does it work?

- **Data amplification** to minimise the effect of noise in the training signals
- Better **selection** of shared hidden representations, reducing the effect of irrelevent inputs when there is limited data
- **Eavesdropping** on a good underlying representation that may be easily learned for one task but not for another.

Why does it work?

- **Data amplification** to minimise the effect of noise in the training signals
- Better **selection** of shared hidden representations, reducing the effect of irrelevent inputs when there is limited data
- **Eavesdropping** on a good underlying representation that may be easily learned for one task but not for another.
- **Representation bias** – tasks prefer representations that other tasks also prefer

Why does it work?

- **Data amplification** to minimise the effect of noise in the training signals
- Better **selection** of shared hidden representations, reducing the effect of irrelevent inputs when there is limited data
- **Eavesdropping** on a good underlying representation that may be easily learned for one task but not for another.
- **Representation bias** – tasks prefer representations that other tasks also prefer
- Using extra features as **output** may be better than using them as **input**

Taking multitask learning further

- We've seen that adding incorporating additional label information in training can result in better hidden representations

Taking multitask learning further

- We've seen that adding incorporating additional label information in training can result in better hidden representations
- What if we derive different variants of a single labelling and train in a multitask way?

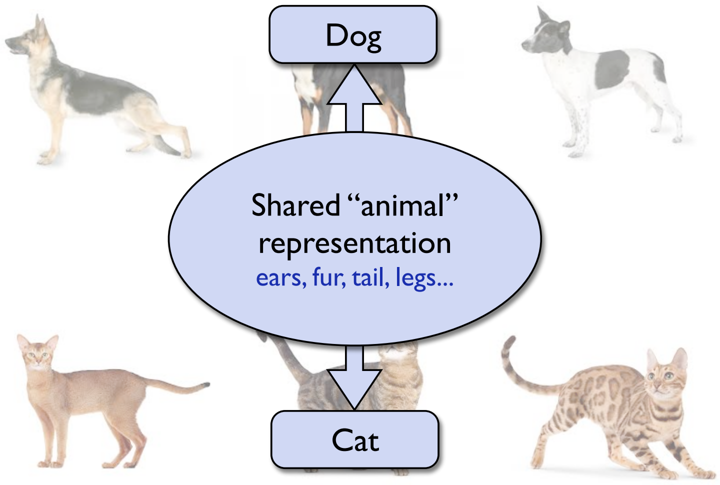
Taking multitask learning further

- We've seen that adding incorporating additional label information in training can result in better hidden representations
- What if we derive different variants of a single labelling and train in a multitask way?
- Motivated by curriculum learning

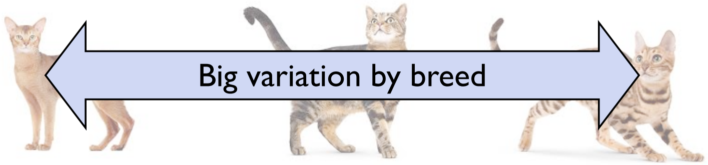
Illustration: classifying cats and dogs

How could a machine learn to tell them apart?

Learn useful discriminative features

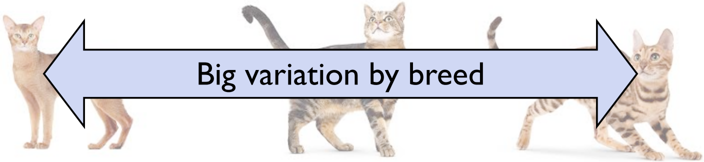


Could we learn the right features

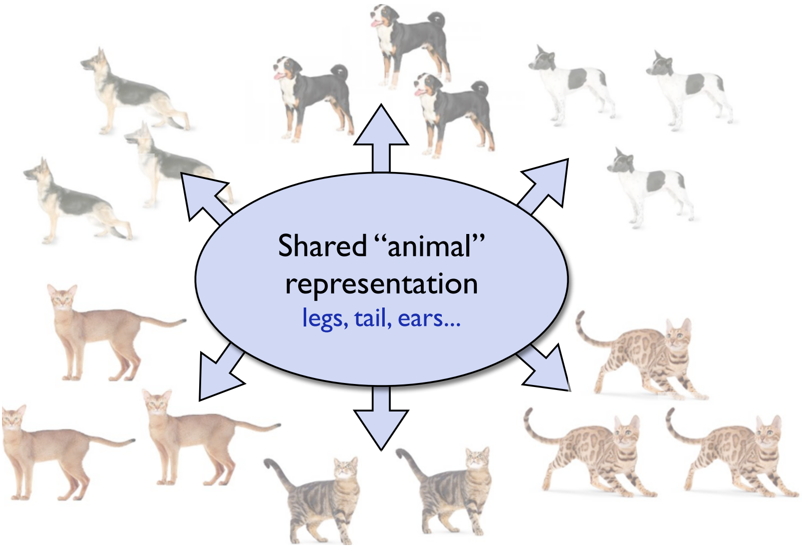


Could we learn the right features

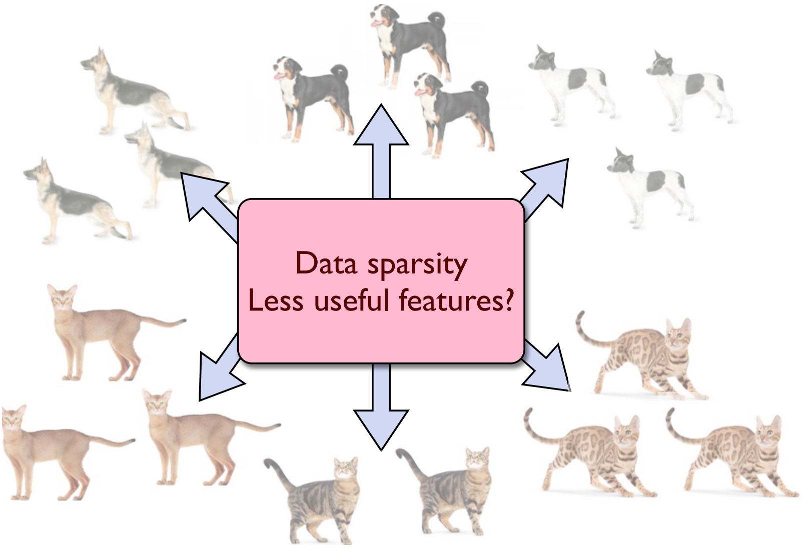
Hard to learn the most relevant features



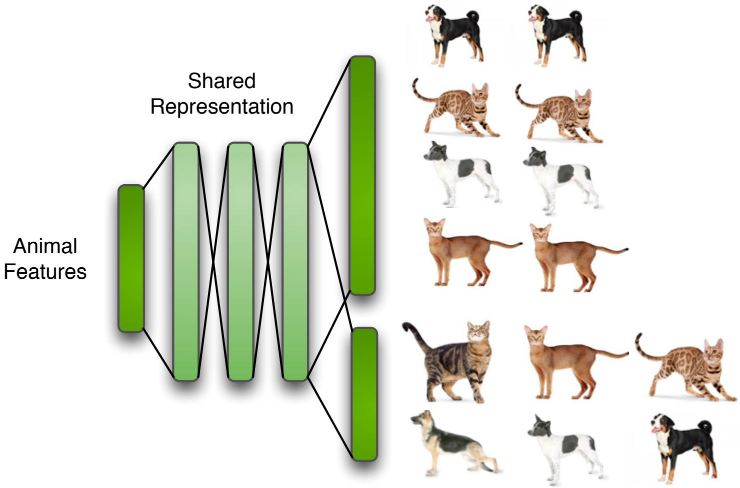
Better to discriminate between breeds?



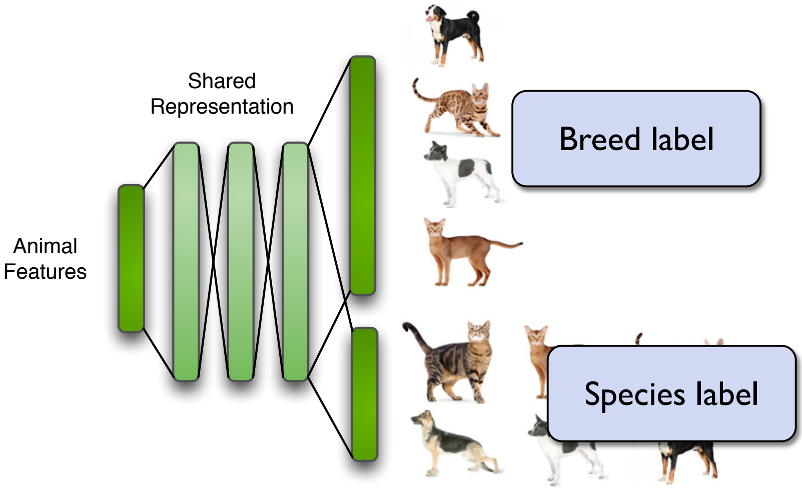
Better to discriminate between breeds?



Solution: learn both sets of labels



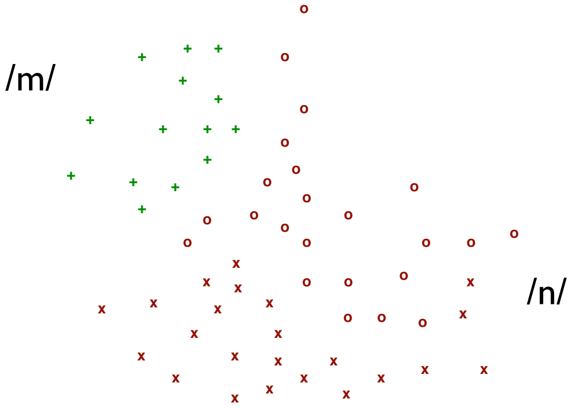
Solution: learn both sets of labels



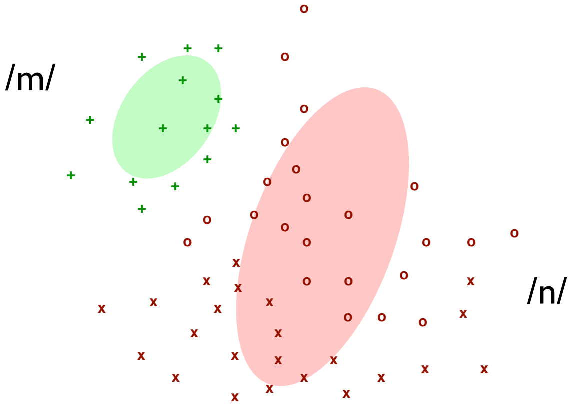
Example: phone modelling for speech recognition

- We want to model *phones*, the distinct units of speech (48 in English)
- But the placement of a phone in the input acoustic feature space is highly dependent on the surrounding phones
- Usually, DNNs model a phone together with both adjacent phones
- Clustering used to reduce 110,000 labels to around 5,000-10,000.

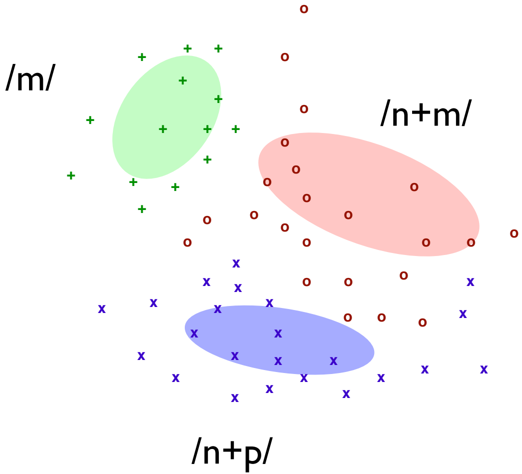
Modelling phones with context



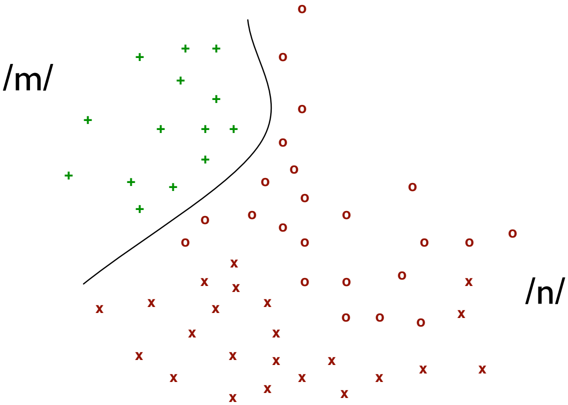
Modelling phones with context



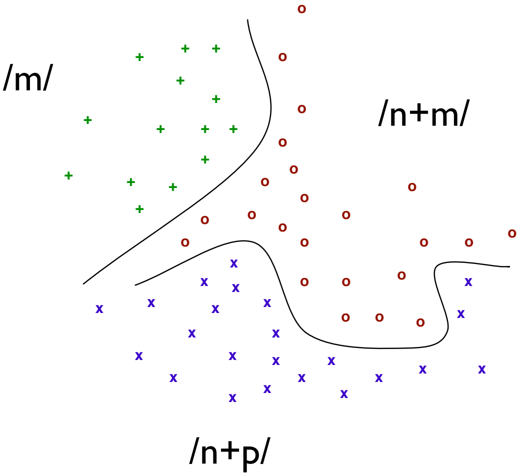
Modelling phones with context



Modelling phones with context

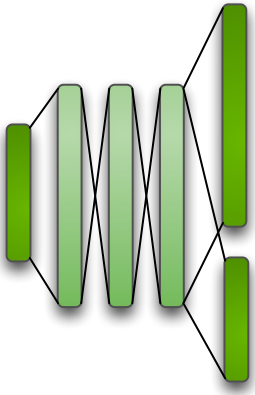


Modelling phones with context

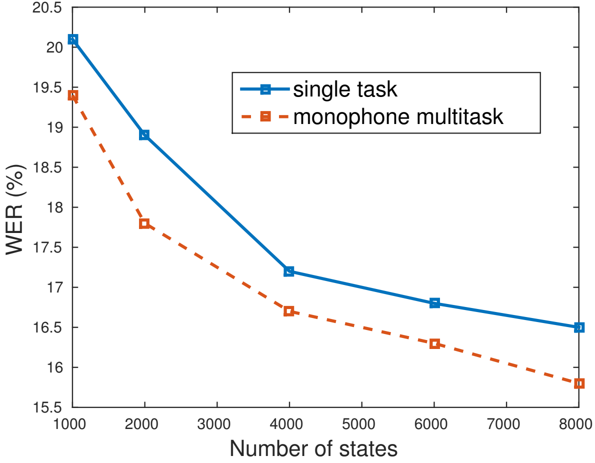


Modelling phones with context

- Use multitask learning to avoid over-fitting to a single set of targets



Speech recognition results



- Curriculum learning
- Multitask learning
- **Student-teacher models**

Student teacher models

- We have seen that it's possible to learn a better model from alternative labellings of the data, rather than fixing on a single hard set of labels

Student teacher models

- We have seen that it's possible to learn a better model from alternative labellings of the data, rather than fixing on a single hard set of labels
- What if we replaced the labelling with the predictions from another model?

Student teacher models

- We have seen that it's possible to learn a better model from alternative labellings of the data, rather than fixing on a single hard set of labels
- What if we replaced the labelling with the predictions from another model?
- Effectively "soft" labels $\rightarrow$ richer and more informative

Student teacher models

- We have seen that it's possible to learn a better model from alternative labellings of the data, rather than fixing on a single hard set of labels
- What if we replaced the labelling with the predictions from another model?
- Effectively "soft" labels $\rightarrow$ richer and more informative
- This is the idea behind *student-teacher* models

Student teacher models

- Train a smaller, weaker model to mimic the outputs of a larger model

Student teacher models

- Train a smaller, weaker model to mimic the outputs of a larger model
- Minimise the KL divergence between the two:

$$KL(P_T, P_S) = \sum_n \sum_i P_T(s_i|x_n) \log \frac{P_T(s_i|x_n)}{P_S(s_i|x_n)}$$

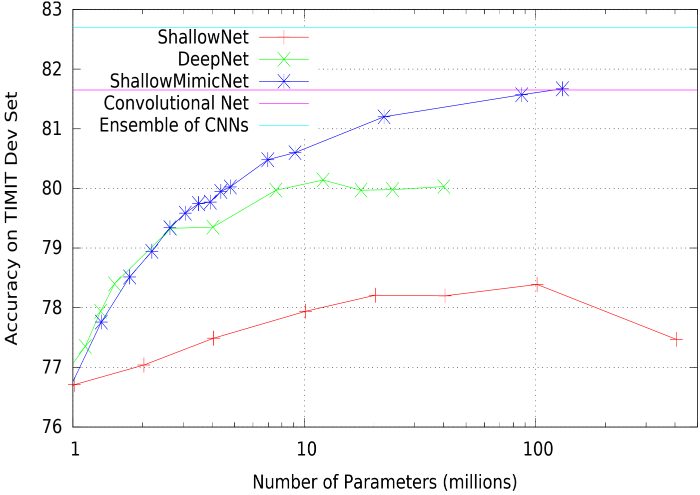
Student teacher models

- Train a smaller, weaker model to mimic the outputs of a larger model
- Minimise the KL divergence between the two:

$$KL(P_T, P_S) = \sum_n \sum_i P_T(s_i|x_n) \log \frac{P_T(s_i|x_n)}{P_S(s_i|x_n)}$$

- Training shallow nets to mimic deep nets has given performance on speech recognition data sets previously achievable only by deeper models

Example (Ba and Caruana, 2015)



Conclusions

- When training a discriminative model, we should be careful about the labelling that is used...

Conclusions

- When training a discriminative model, we should be careful about the labelling that is used...
- ... especially if the labelling is in some way arbitrary

Conclusions

- When training a discriminative model, we should be careful about the labelling that is used...
- ... especially if the labelling is in some way arbitrary
- View multiple labelling schemes, or soft labelling, as an additional source of information about the samples

Conclusions

- When training a discriminative model, we should be careful about the labelling that is used...
- ... especially if the labelling is in some way arbitrary
- View multiple labelling schemes, or soft labelling, as an additional source of information about the samples
- Learn more general representations by fitting to multiple tasks

Reading

- Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proc. ICML*, 2009.
- R. Caruana, "Multitask learning," *Machine learning*, vol. 28, pp. 41–75, 1997.
- J. Ba and R. Caruana, "Do deep nets really need to be deep?," in *Proc. NIPS*, 2014.
- P. Bell, P. Swietojanski, and S. Renals, "Multitask learning of context-dependent targets in deep neural network acoustic models", in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, issue 2, 2017