CS/DS 541: Class 9

Jacob Whitehill

Group Exercise

https://www.wellandgood.com/missing-richard-simmons-podcast-fitness-history/

Group Exercise

Suppose you have a very deep NN (e.g., where the number of layers $l = 10000$). Let the element-wise activation function of every hidden layer be σ, and suppose that, given the current set of network weights $\mathcal{W} = \{\mathbf{W}^{(k)}\}_{k=1}^{l}$ and some input $\mathbf{x}$, the pre-activation vector $\mathbf{z}^{(k)}$ at each layer k contains all zeros. What happens to $\nabla_{\mathbf{W}^{(k)}} f(\mathbf{x}; \mathcal{W})$, where f is the loss function (e.g., MSE), as $k \to 0$, for the following 3 cases:

- $\sigma = \tanh$.
- $\sigma = \text{ReLU}$.
- $\sigma(z) = (1 + \exp(-z))^{-1}$, i.e., logistic sigmoid.

To get started, first (1) draw each of these activation functions on paper and/or the screen. Next, (2) determine each of their derivatives evaluated at 0. Finally, (3) think about how the chain-rule yields each $\nabla_{\mathbf{W}^{(k)}} f(\mathbf{x}; \mathcal{W})$, and how the corresponding product-of-Jacobians (some of which are directly affected by σ') changes for each of the three cases above.

Data normalization

Data normalization

- In many ML settings (not just DL), it is important or useful to normalize the input data to improve accuracy or ease of training.

- Common normalization methods include:
  - Mapping *each* feature into a fixed range (e.g., [0,1], [-1,1]).
  - Z-scoring *each* feature (so that the mean and stddev are 0 and 1, respectively).
  - Whitening *all* features jointly (to have 0 mean and *I*-covariance).

Data normalization

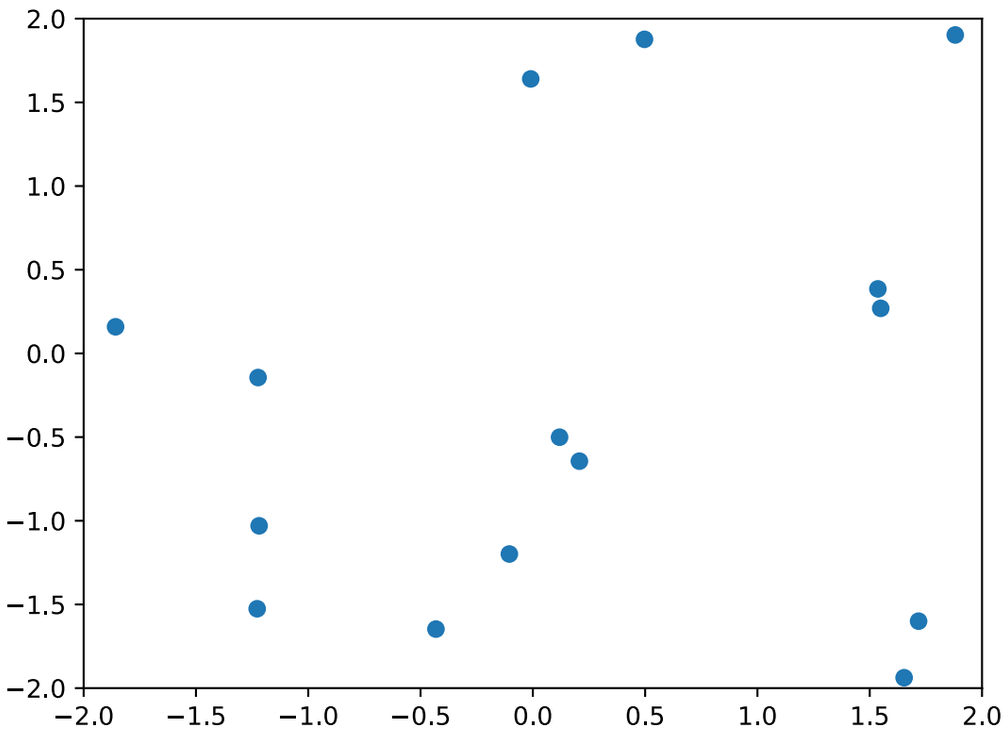
- There are (at least) two ways to achieve this.
- Strategy 1: Learn transformation on training data:
  - Compute the normalization parameters on the training set; save them.
  - Apply the normalization to the training data *and* each of the testing data.

Data normalization

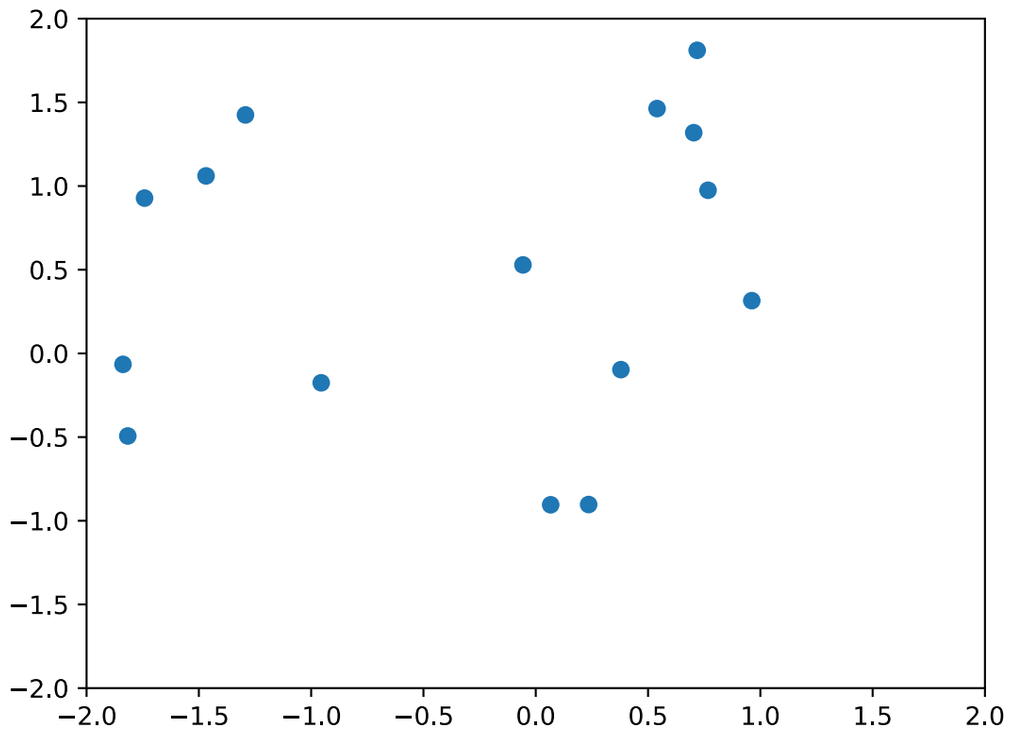
- There are (at least) two ways to achieve this.
- Strategy 2: Learn transformations for training and testing data *separately*:
  - Compute the normalization parameters on the *training* set, and apply it to each of the *training* data.
  - Compute the normalization parameters on the *testing* set, and apply it to each of the *testing* data.

Strategy 1

Unnormalized training data

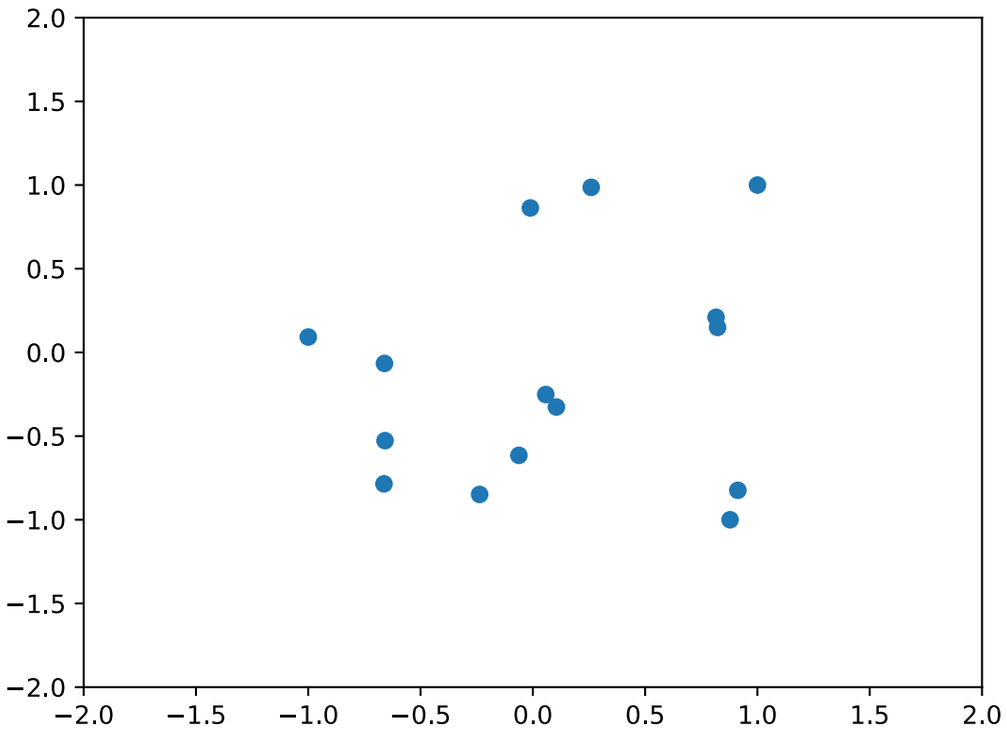


Unnormalized testing data

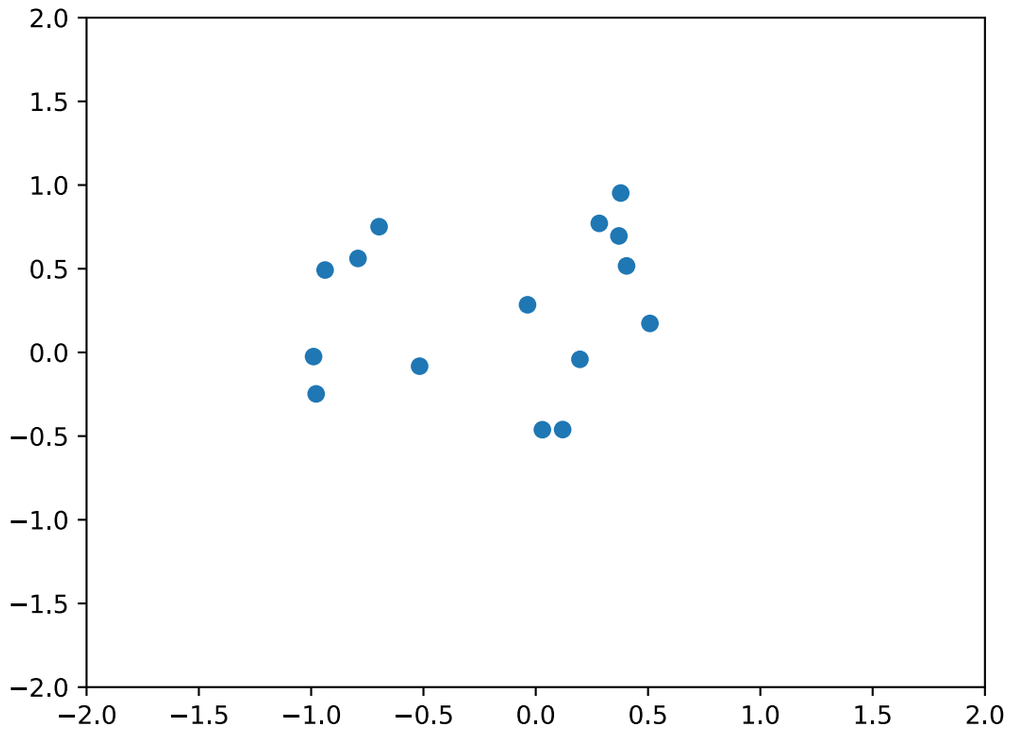


Strategy 1

Normalized training data

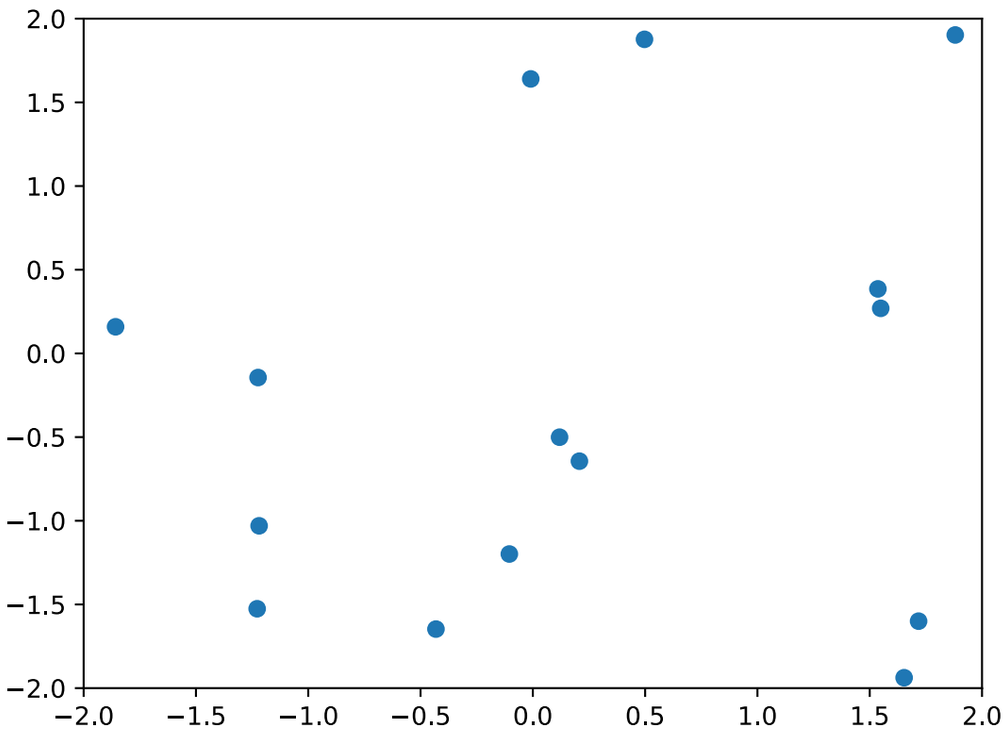


Normalized testing data

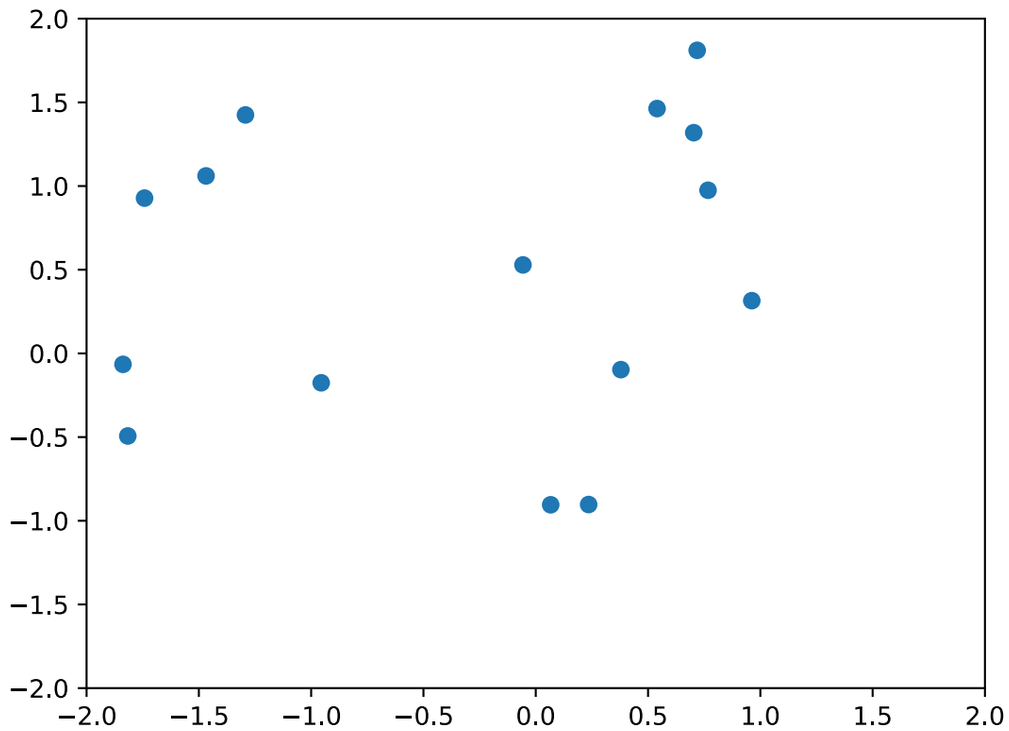


Strategy 2

Unnormalized training data

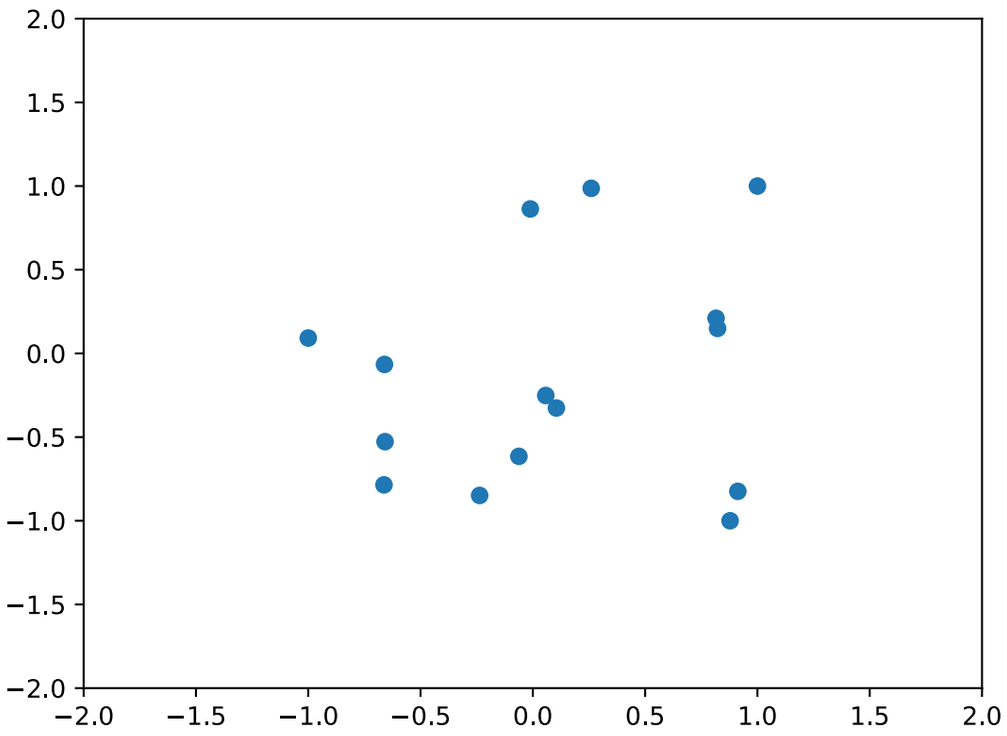


Unnormalized testing data

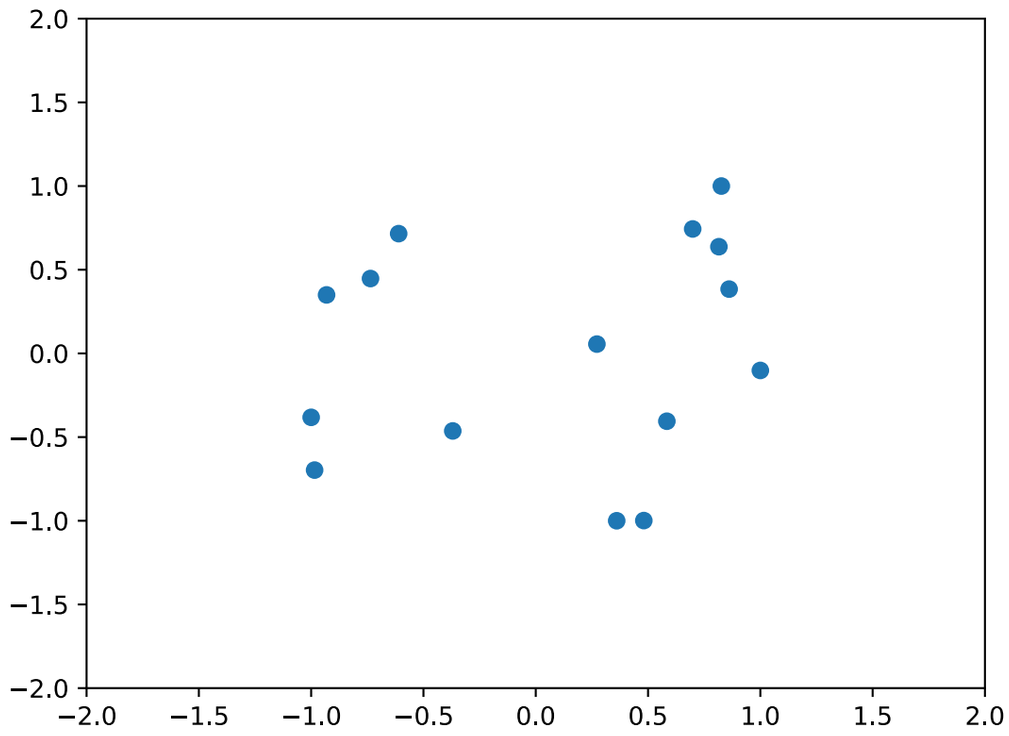


Strategy 2

Normalized training data



Normalized testing data



Strategy 1 vs. Strategy 2

- Strategy 2 results in normalized testing data that exactly fit the range [-1, +1].

Strategy 1 vs. Strategy 2

- Strategy 2 results in normalized testing data that exactly fit the range [-1, +1].

- However, normalizing in this way requires knowledge of the test distribution.

- This means it is necessary to collect a sufficiently large sample of testing data *before* running the classifier.

Data augmentation

Data augmentation

- The more training data you have, the less is the risk of overfitting.
- Unfortunately, training data are often hard to find.
- Can we synthesize new training examples automatically?

Data augmentation

- **Data augmentation** is the creation of new examples based on existing ones.

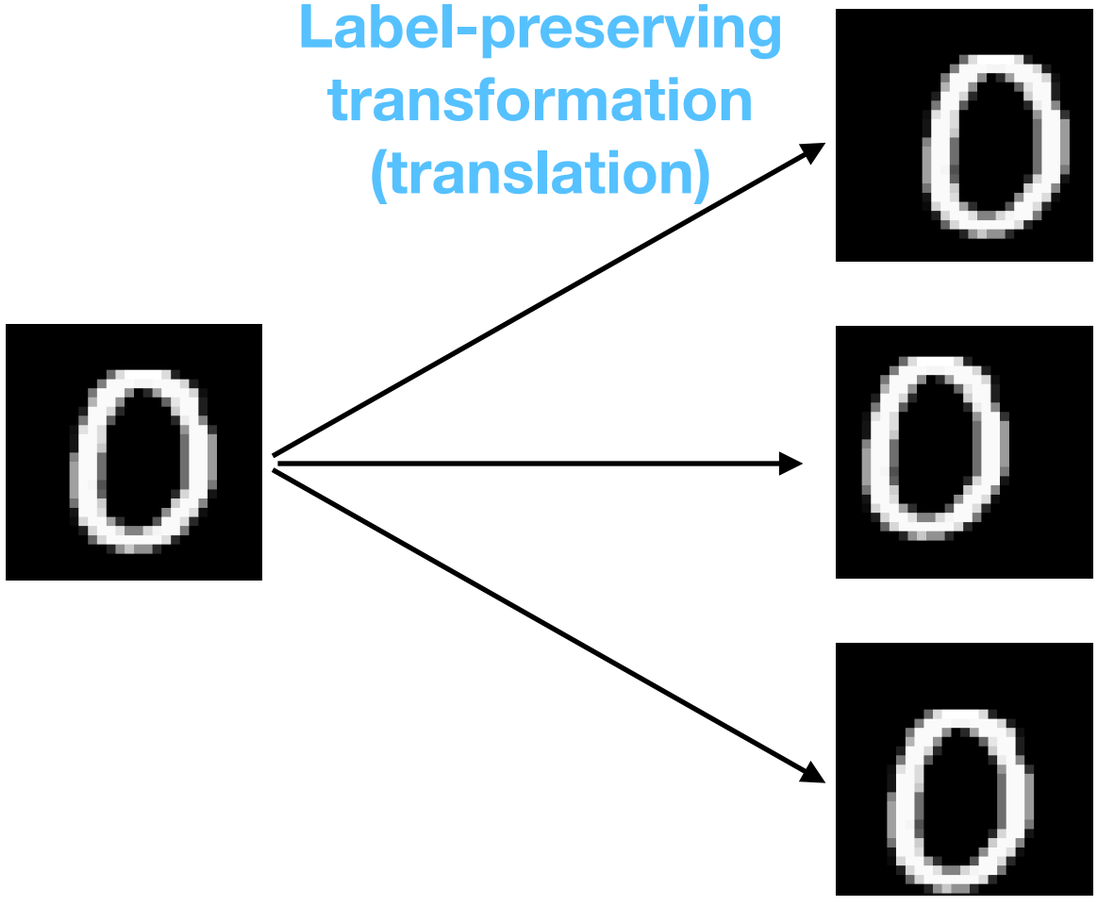
- If we can alter an existing training example without affecting its associated label, then we can generate many new training examples and train on them.

Data augmentation

- Several commonly used methods of data augmentation:
  - Adding noise to existing examples (e.g., Gaussian, Laplacian).
  - Geometric transformations (e.g., flip left/right, rotate, translate).

Example: translation

- From an existing MNIST image, translate all the pixels by some random amount (dx, dy).



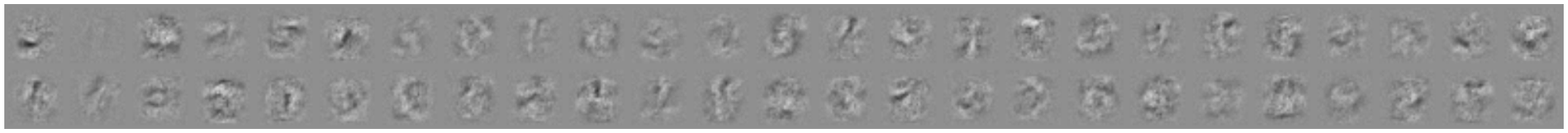
Example: translation

- Data augmentation via translation encourages the NN to learn **translation-invariant** features — they are useful for classification no matter where in the image they occur.

Example: translation

- Here are the weights $\mathbf{W}^{(1)}$ (transformed to 100x28x28) of a MNIST classification network **without** data augmentation:

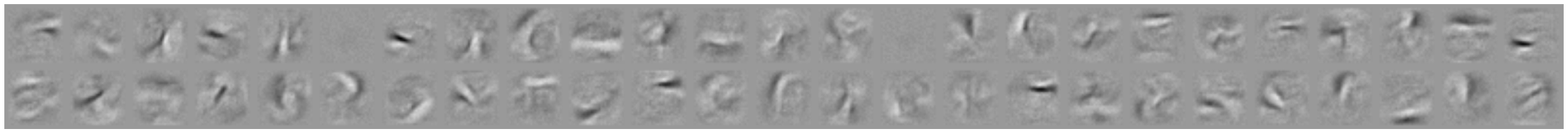
Acc=98.07



Example: translation

- Here are the weights $\mathbf{W}^{(1)}$ (transformed to 100x28x28) of a MNIST classification network **with** data augmentation:

Acc=98.44



- Compared to the previously shown weights, these show visually more well-defined contours.

(Supervised)
pre-training

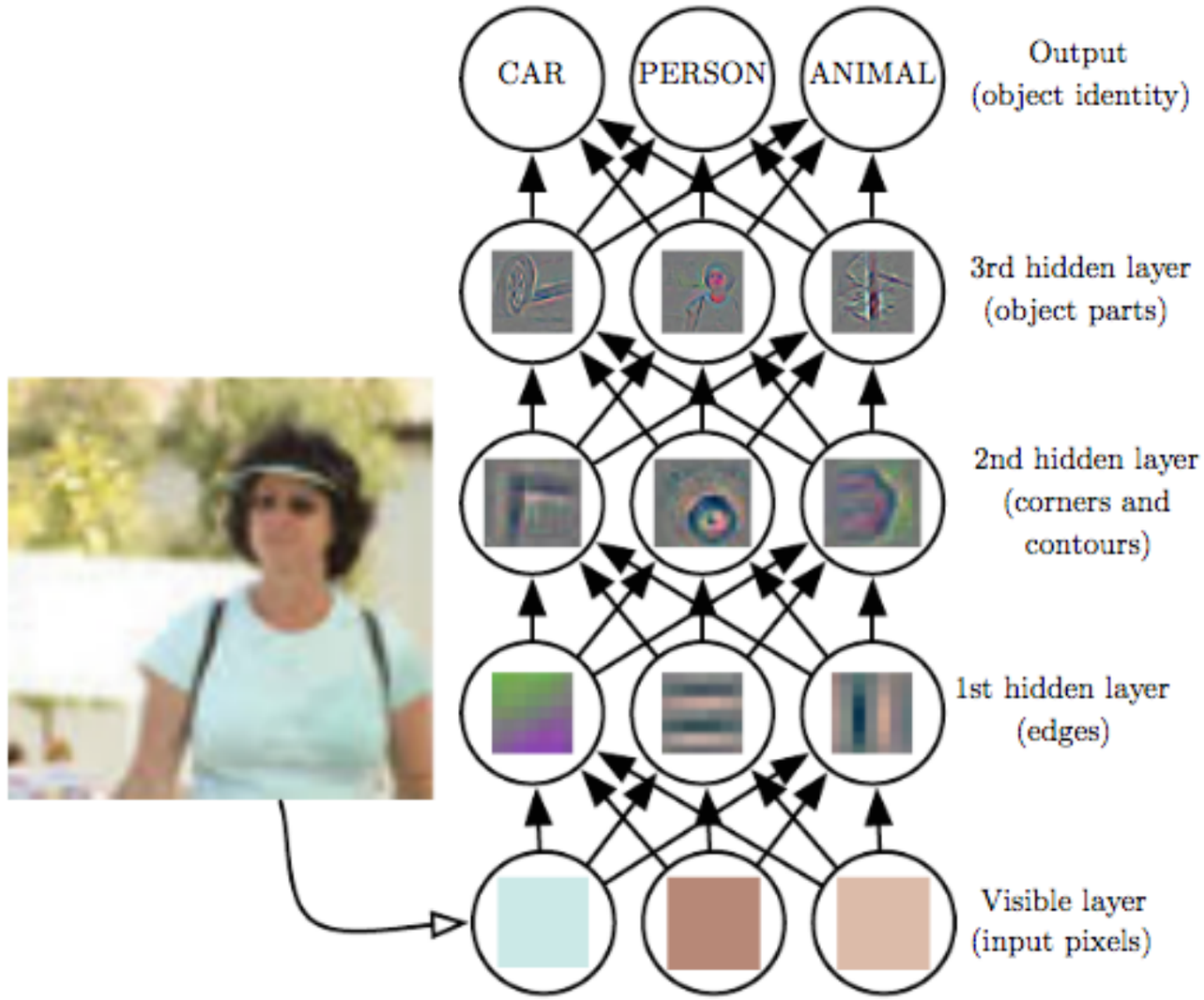
Supervised pre-training

- An alternative strategy to finding good feature representations is to borrow a NN from a related task.

- For instance, there now exist high-accuracy networks for recognizing 1000+ object categories from images (next slide).

- We can “borrow” the feature representation from one ML model and apply it to another application domain…

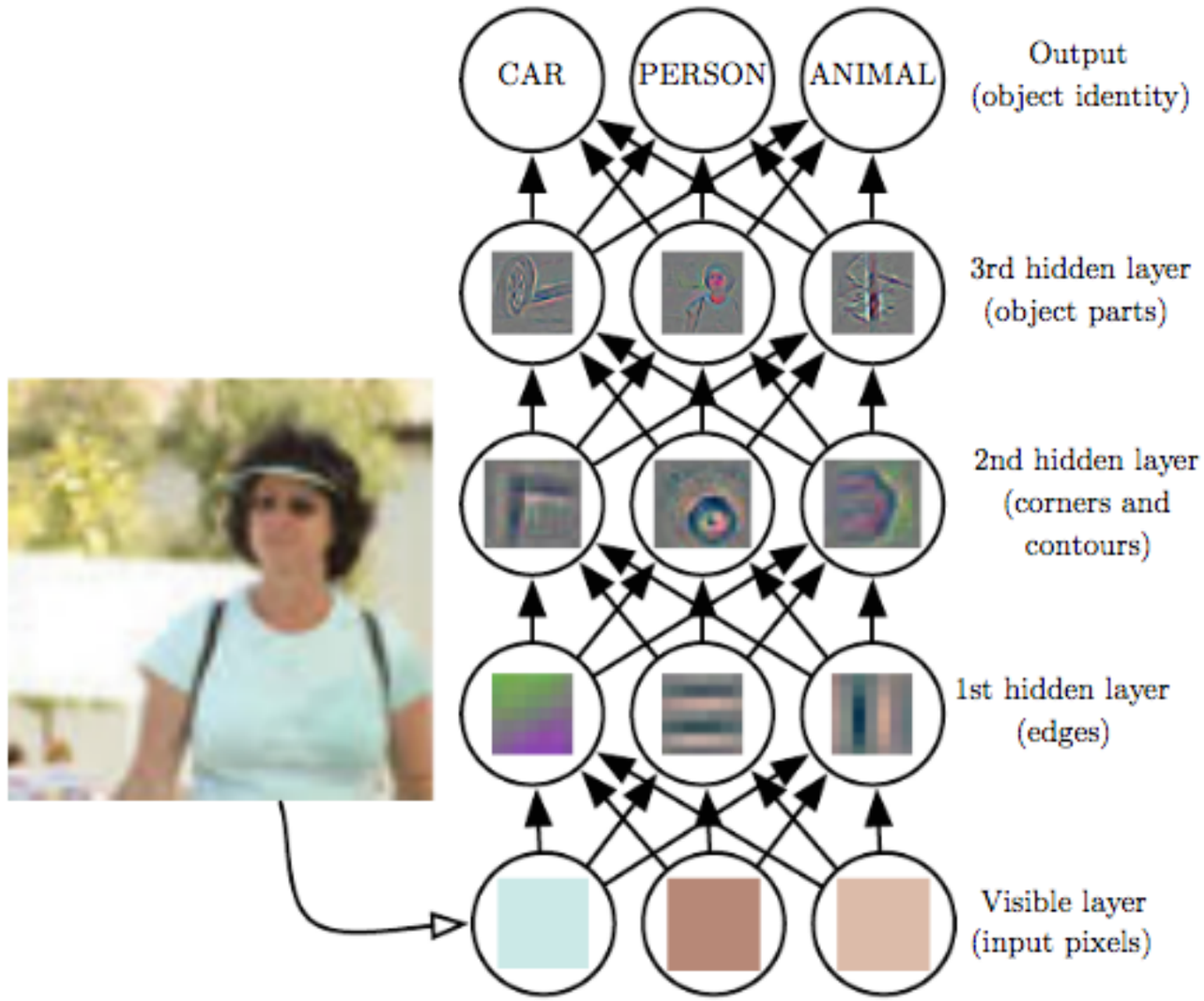
Learning representations



- The first feature representation looks vaguely like the representation learned by my MNIST network.

http://www.deeplearningbook.org/contents/intro.html

Learning representations



- Each layer of the network finds successively more abstract feature representations.
  - This was not “hard-coded” — it just turned out that these representations were useful for predicting the target labels.

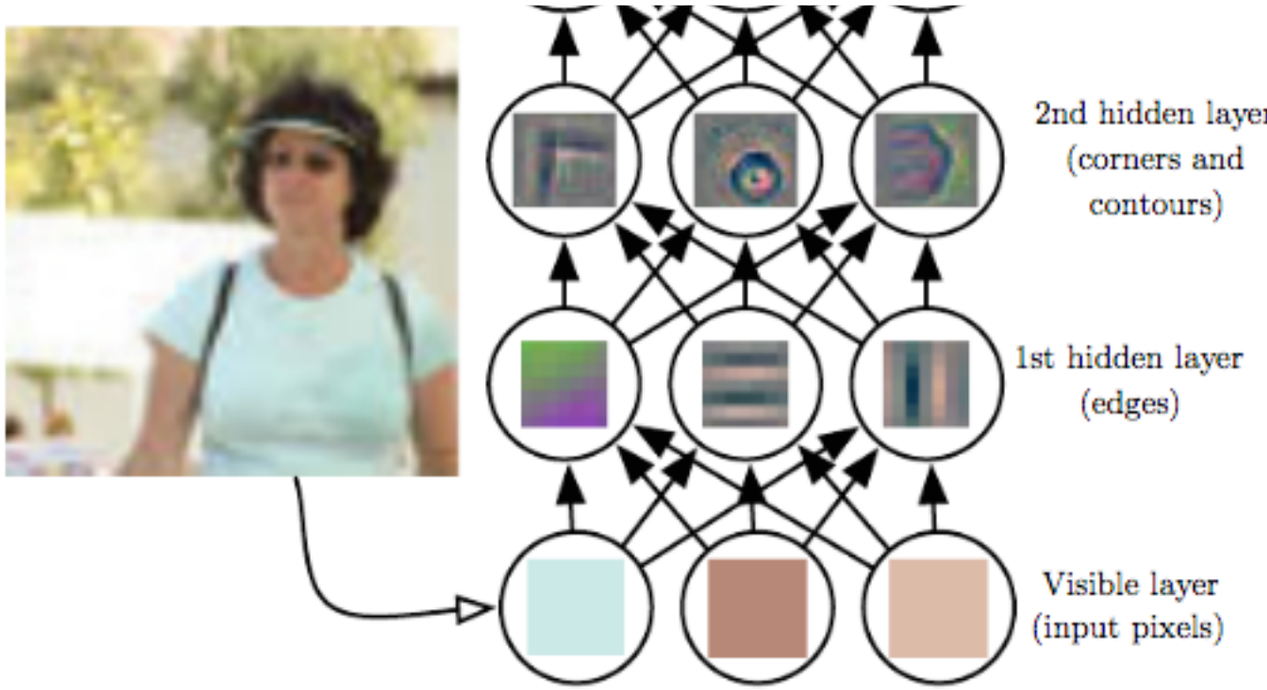
http://www.deeplearningbook.org/contents/intro.html

Supervised pre-training

- Might one (or more) of the feature representations from this NN do well on a different but related problem, e.g., smile detection or age estimation?
- Strategy:
  1. Pre-train a NN on a large dataset (e.g., ImageNet) for a general-purpose image recognition task.
  2. “Chop off” the final layer(s).
  3. Add a secondary network in place of the deleted layers, and train it for the new prediction task.
  4. Optional: fine-tune the rest of the NN.

Supervised pre-training

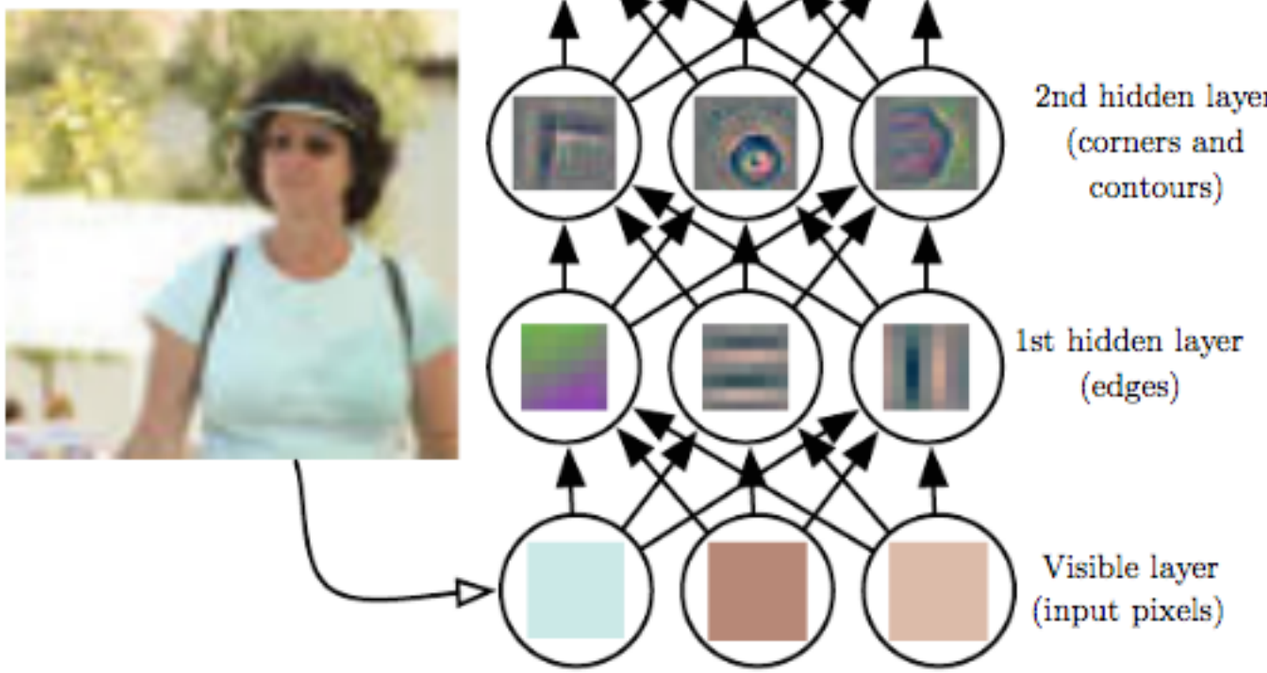
Chop off.



http://www.deeplearningbook.org/contents/intro.html

Supervised pre-training

Replace with secondary network for new application domain.



http://www.deeplearningbook.org/contents/intro.html

Supervised pre-training

- This strategy is known as **supervised pre-training** and can be highly effective for application domains for which only a small number of labeled data are available.

Multi-task learning (MTL)

Multi-task learning (MTL)

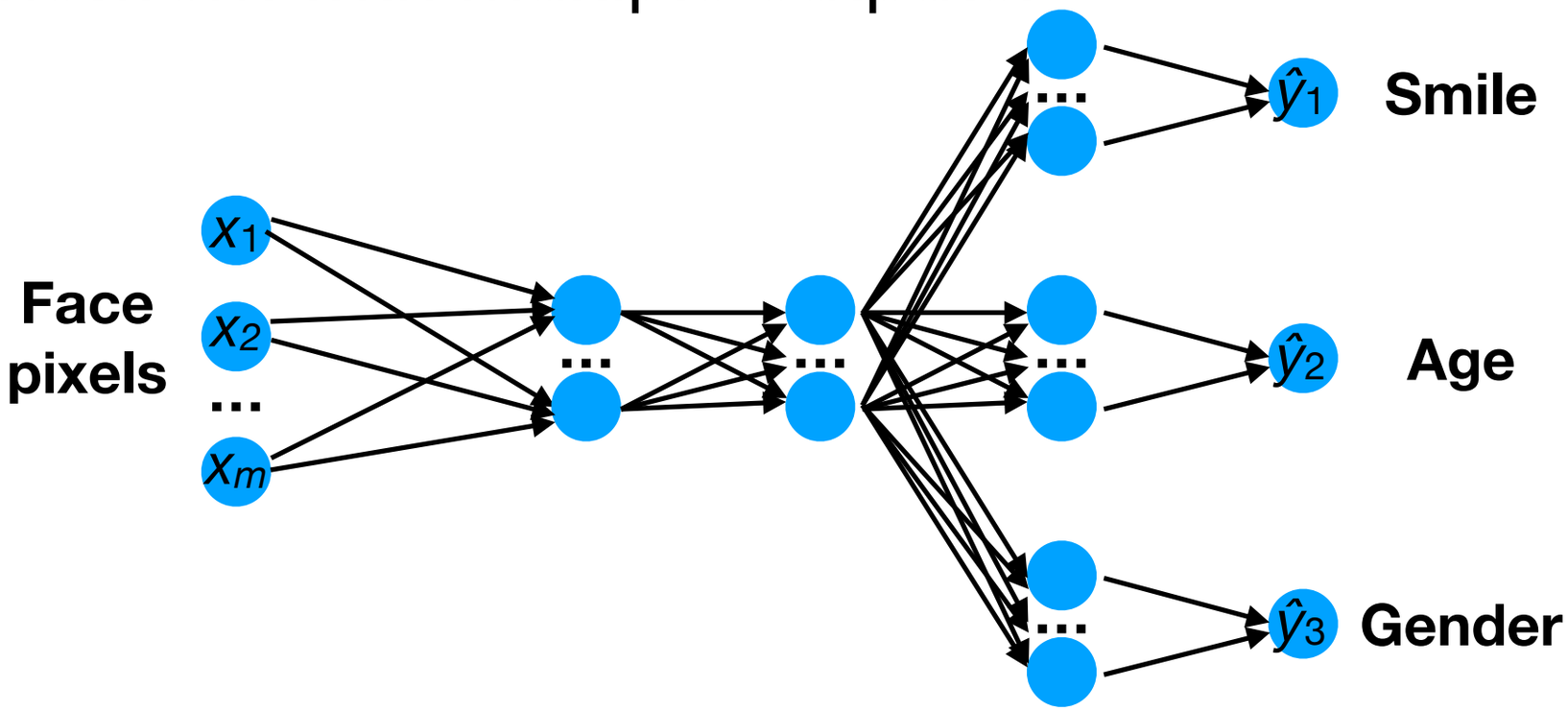
- A NN can generalize to unseen data when it computes a hidden representation that explains the data well.

- We can sometimes encourage the NN to learn a general hidden representation by training it to solve multiple related tasks.

- E.g., for automated face analysis:
  - Smile detection
  - Age estimation
  - Gender detection

Multi-task learning (MTL)

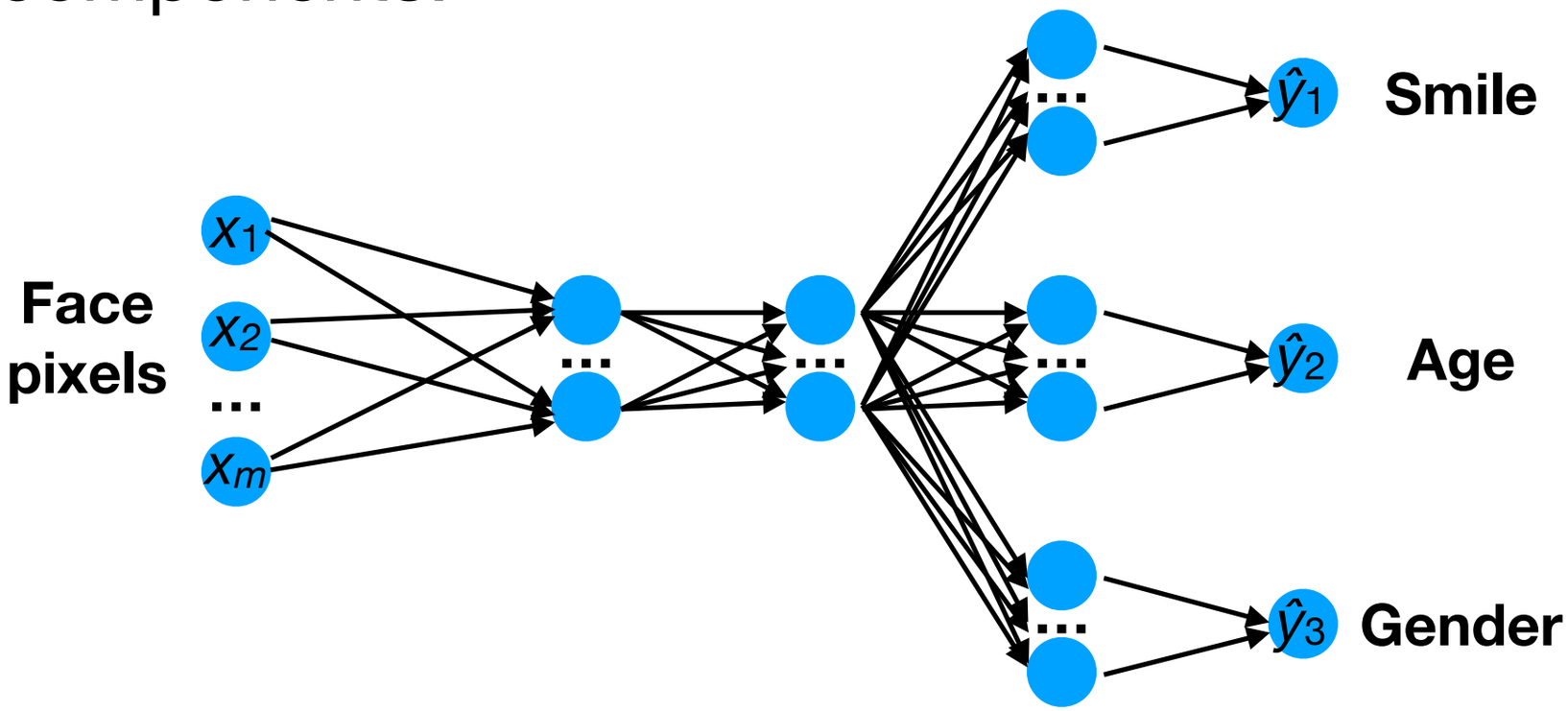
- If we have labeled data for all these tasks, we can train a network with multiple outputs:



- Note that the labels for the auxiliary tasks can be useful even if we only care about one task.

Multi-task learning (MTL)

- With MTL, our loss function consists of multiple components:



$$f_{\text{MTL}}(\mathbf{W}^{(1)}, \mathbf{b}^{(1)}, \ldots) = f_{\text{smile}}(\mathbf{W}^{(1)}, \mathbf{b}^{(1)}, \ldots) + f_{\text{age}}(\mathbf{W}^{(1)}, \mathbf{b}^{(1)}, \ldots) + f_{\text{gender}}(\mathbf{W}^{(1)}, \mathbf{b}^{(1)}, \ldots)$$

Fairness in ML

Fairness in ML

- Suppose we are training a classifier to perceive whether a person is smiling based on their face image.


vs.


- Suppose the joint probability distribution of our training labels is:

	Male	Female
Smile	0.47	0.02
Non-smile	0.45	0.06

Almost all data are male faces

Fairness in ML

- Suppose we are training a classifier to perceive whether a person is smiling based on their face image.


vs.


- Suppose the joint probability distribution of our training labels is:

	Male	Female
Smile	0.47	0.02
Non-smile	0.45	0.06

Almost all female faces are non-smiling

Fairness in ML

- To minimize prediction error, the training algorithm can harness the fact that *most women do not smile* ***in this dataset***.


vs.


	Male	Female
Smile	0.47	0.02
Non-smile	0.45	0.06

Fairness in ML

- At test time, that machine might implicitly infer that the face is female, and then use that to “downgrade” the estimate of the smile probability.


vs.


	Male	Female
Smile	0.47	0.02
Non-smile	0.45	0.06

Fairness in ML

- In contrast, a male face has a roughly equal chance of being classified as smile/ non-smile.

vs.

	Male	Female
Smile	0.47	0.02
Non-smile	0.45	0.06

Fairness in ML

- Consider the following definition of ML fairness:
  - For all subgroups (i.e., smiling males, non-smiling males, smiling females, non-smiling females), the prediction accuracy should be equal.

Fairness in ML

- Consider the following definition of ML fairness:
  - For all subgroups (i.e., smiling males, non-smiling males, smiling females, non-smiling females), the prediction accuracy should be equal.
- Then the classifier described above is unfair because female faces would likely be perceived less accurately compared to male faces.

Fairness in ML

- Mitigating strategies:
    - Collect more training data for specific populations.
    - Train the classifier with structural constraints that prevent the exploitation of correlated but non-causal features:
        - In this dataset, gender is correlated with smile, but does not cause smile!