Deep Learning

6.1a Regularization for Deep Learning

Dr. Konda Reddy Mopuri
kmopuri@iittp.ac.in
Dept. of CSE, IIT Tirupati

Deep Regularization

1. Most of the regularization techniques for deep learning are based on regularizing estimators

Deep Regularization

1. Most of the regularization techniques for deep learning are based on regularizing estimators
2. Trade increased bias for decreased variance

Deep Regularization

1. An overly complex model family need not include the target function

Deep Regularization

1. An overly complex model family need not include the target function
2. In practice we almost never have access to the true data generating process, and which is almost certainly outside the model family

Deep Regularization

1. Most often the best-fitting model is a large model that has been appropriately regularized

Deep Regularization

- Parameter Norm penalties (l_2, l_1, etc.)
- Dataset Augmentation
- Noise Robustness
- Semi-Supervised Learning
- Multi-Task Learning (Parameter sharing)
- Sparse Representation
- Dropout
- etc.

Parameter Norm Penalties

1. For neural networks, typically only the weights of the affine transformations are regularized leaving the biases unregularized

Parameter Norm Penalties

1. For neural networks, typically only the weights of the affine transformations are regularized leaving the biases unregularized
2. Bias controls only a single variable as opposed to weight which connects two

Parameter Norm Penalties

1. For neural networks, typically only the weights of the affine transformations are regularized leaving the biases unregularized
2. Bias controls only a single variable as opposed to weight which connects two
3. Regularizing biases induces underfitting

Parameter Norm Penalties

1. L_2 parameter regularization: $\tilde{\mathcal{J}} = \frac{\alpha}{2} w^T w + \mathcal{J}(w; X, y)$

Parameter Norm Penalties

1. L_2 parameter regularization: $\tilde{\mathcal{J}} = \frac{\alpha}{2} w^T w + \mathcal{J}(w; X, y)$
2. L_1 regularization: $\tilde{\mathcal{J}} = \alpha |w|_1 + \mathcal{J}(w; X, y)$

Parameter Norm Penalties

1. L_2 parameter regularization: $\tilde{\mathcal{J}} = \frac{\alpha}{2} w^T w + \mathcal{J}(w; X, y)$
2. L_1 regularization: $\tilde{\mathcal{J}} = \alpha |w|_1 + \mathcal{J}(w; X, y)$
3. Norm penalties induce different desired behaviors based on the exact penalty imposed

Dataset Augmentation

1. Bestway to make ML model generalize better is to train with more data

Dataset Augmentation

1. Bestway to make ML model generalize better is to train with more data
2. In practice training data is limited

Dataset Augmentation

1. Bestway to make ML model generalize better is to train with more data
2. In practice training data is limited
3. Create fake data and add it to the training data, called Dataset augmentation

Dataset Augmentation

1. Easier for classification

Dataset Augmentation

1. Easier for classification
2. Difficult for density estimation task (unless we have solved the estimation problem)

Dataset Augmentation

1. Has been particularly effective for specific classification problems such as object recognition

Dataset Augmentation

1. Has been particularly effective for specific classification problems such as object recognition
2. Operations such as translation by few pixels, rotating slightly, adding mild noise, etc. greatly improve generalization

Dataset Augmentation

1. Has been particularly effective for specific classification problems such as object recognition
2. Operations such as translation by few pixels, rotating slightly, adding mild noise, etc. greatly improve generalization
3. Hand-designed augmentations in some domains can result in dramatic improvements
4. Should restrict to label preserving transformations

Multi-Task Learning

1. Improves generalization by collecting samples arising out of multiple taks

Multi-Task Learning

1. Improves generalization by collecting samples arising out of multiple taks
2. Similar to additional data samples, multi-task samples also put more pressure on the parameters of the shared layers to be better

Multi-Task Learning

1. Improves generalization by collecting samples arising out of multiple taks
2. Similar to additional data samples, multi-task samples also put more pressure on the parameters of the shared layers to be better