1] Vanishing gradients

- Activation functions
    - Linear
    - Sigmoid
    - Tanh
    - ReLU
        - Relu variants
            - Linear
                - Leaky ReLU
                - Parametric ReLU
            - Non linear
                - ELU [Exponential linear unit]
                - SeLU [Scaled exponential linear unit]
- Weight initialization
    - Xavier / Glorot
        - Normal
        - Uniform
    - He initialization
        - Normal
        - Uniform

2] Overfitting

- Reduce complexity / Increase data
- Dropout layers
- Regularization (L1 & L2)
- Early stopping

3] Normalization

- Normalizing inputs
- Batch normalization
- Normalizing activations

4] Gradient checking and clipping

5] Optimizers

- Momentum
- NAG [Nesterov accelerated gradient]
- Adagrad [Adaptive gradient]
- RMSprop [Root mean squared propagation]
- Adam [Adaptive moment estimation]

6] Learning rate scheduling

7] Hyperparameter tuning

- Number of hidden layers
- Nodes / layer
- Batch size