Advanced Topics in Machine Learning - 01/10/2022

Topics covered in this module:

- Recurrent Neural Networks (RNNs)
- Introduction to NLP
    - Assignment #2
- Attention Mechanism and the Transformer
- Self-supervised Learning and Large Language Models (LLMs)

Recurrent Neural Networks (RNNs)

Introduction

So far, the networks were memoryless. In RNNs there is a state $\mathbf{h}$ which enables the network to remember previous states e.g time series processing

This temporal dynamics is made possible by a feedback mechanism: $$ \mathbf{h^{(t)}} = f(\mathbf{x^{(t)}}, \mathbf{h^{(t-1)}}) $$

This function f is the usual affine transformation followed by a *squashing function* (activation function), e.g $\tanh$ or σ $$ \mathbf{h^{(t)}} = \tanh (\mathbf{U}\mathbf{x^{(t)}} + \mathbf{W}\mathbf{h^{(t-1)}} + \mathbf{b}) $$

Notice that the way this is written assumes causality (only previous states affect current state), and it is formulated to depent explicitly only on the previous state (but since the previous state depends on the previos etc., there is a complete temporal relationship implicitly defined).

See Handwritten Notes, figure #1

Training

For training, backpropagation can still be used. Let us see an example:

Considerar input sequence $\mathbf{x} = [\mathbf{x}^{(1)}, \dots,\mathbf{x}^{(T)}]$ and labels $\mathbf{y} = [\mathbf{y}^{(1)}, \dots,\mathbf{y}^{(T)}]$, together with the loss function $$ \mathcal{L} = \sum_{t=1}^{T}\mathcal{L}^{t} = \sum_{t=1}^{T} l(\mathbf{y}^{(t)}, f(\mathbf{x}^{(t)}, \mathbf{h}^{(t-1)}; \mathbf{\theta}) $$

And the optimization problem is

$$ \argmin_{\mathbf{\theta}} \sum_{t=1}^{T} l(\mathbf{y}^{(t)}, f(\mathbf{x}^{(t)}, \mathbf{h}^{(t-1)}; \mathbf{\theta}) $$

with the algorithm using $\mathbf{\theta} \leftarrow \mathbf{\theta} - \eta \nabla_{\mathbf{\theta}} \mathcal{L}}$

The problem of long-term memory

It is hard for the RNN to learn long-term dependencies, i.e. it tends to forget information that is too far away temporally

If we consider a network without non-linearity and without bias, i.e $\mathbf{h}^{(t)} = \mathbf{W}\mathbf{h}^{(t-1)}$, which implies

$$ \mathbf{h}^{(t)} = \mathbf{W}^t\mathbf{h}^{(0)} $$

Imagine that $\mathbf{h}, \mathbf{W}$ are scalars. For t big,there are two possibilities:

1. If $|\mathbf{W}| > 1, $
2. If $|\mathbf{W}| < 1$