1. Calculation on only the gradient involves computing the partial derivatives of the loss function to each parameter. It is less expensive compared to computing the total loss when the dataset is large. Meanwhile, with partial derivatives used in gradient update, the optimization process tends to prioritize features that have higher impact on reducing the loss.

2. The order of words concatenation does matter.

Reason: since in text analysis, the order of words is also a type of information. With different word orders. The sentimental meaning might be different. Therefore, for different task, the output might have big differences.

3. 1

①Vanishing gradient:

In RNN, the same set of weights is applied iteratively across time steps. During BPTT, gradients are multiplied by the same weight matrix at each time step. If the gradient is too small, the gradient propagated will turn smaller, which makes gradient vanish.

② exploding gradient.

exploding gradient means gradient gets bigger while back propagation.

In RNN, if parameters are initialized inappropriately,

gradient might be very large while propagate. This will lead to exploding gradient. Meanwhile, some of activation functions like ReLU, the gradient is unbounded. it will also cause exploding gradient.

3.2.

LSTM has gated mechanisms that can decide how information is transferred. Meanwhile, cell state can keep information for certain state, which can carry information across many time steps. With these architectures, the risk of vanishing gradient will be reduced.