lecture 7 preview

1. vanishing gradient intuition
    a. gradient problem : RNN back propagation 시 그라디언트가 너무 작아지거나, 너무 커져서 학습이 제대로 이루어지지 않는 문제
        - chain rule에 의해 그라디언트가 반복해서 곱해지며 커지는가 작아지는가?
        - Wh matrix의 L2 norm이 결국 Wh의 가장 큰 eigenvalue값이다 → loss에 대한 hidden state의 graident의 L2norm은 Wh의 L2norm의 크기에 의해 결정된다
        - 따라서 Wh의 maximum eigenvalue가 1보다 작다면 그라디언트는 반복해서 곱해지는 과정에서 0에 수렴하게 된다. (vanishing gradient problem)

    (1) Vanishing Gradient

        - Why is vanishing gradient a problem?
            - 가까이 위치한 dependency에 대해서만 학습을 하고 멀리 위치한 dependency에 대해서는 학습을 하지 못한다.
            - 그라디언트는 과거가 미래에 얼마나 영향을 미치는지에 관한 척도인데, 그라디언트 소실 문제가 발생할 경우 파라미터 값의 오류인지/과거가 미래에 영향을 끼치지 않는 것인지 판단을 내리기 어렵다.
        - Effect of vanishing gradient on RNN-LM
            - ex) 멀리 떨어진 단어인 tickets에 대해 학습이 진행되지 않아 printer라는 잘못된 답을 내놓음

    (2) Exploding gradient

        - 솔루션 : Gradient clipping
            - 그라디언트가 일정 범위를 넘어가면 L2 norm으로 나눠준다. (scale down)

2. LTSM(Long Term Short Memory)
    - RNN의 그라디언트 소실 문제로 발생하는 장기 의존성 문제를 해결하기 위한 것
    - 이전 단계의 정보를 메모리셀에 저장한다

- 현재 시점의 정보를 바탕으로 과거의 내용을 얼마나 버릴지 곱해주고, 그 결과에 현재의 정보를 더해 다음 단계로 정보를 전달한다.
- cell state : input, forget, output 세 개의 게이트로 정보의 반영여부를 결정한다.

  (1) Forget gate layer : 과거의 정보 중 어떤 정보를 잊고 어떤 정보를 반영할 것인지 결정한다.

  (2) Input gate layer : 새로운 정보가 들어왔을 때 cell state에 반영할 것인지 결정한다.

  (3) Update cell state : 앞의 두 단계를 바탕으로 업데이트

  (4) Output gate layer : 출력값을 반환한다.

- How does LSTM solve vanishing gradients?

  → forget gate = 1, input gate = 0이면, cell의 정보가 완전하게 보존된다. (장기의존성 문제 해결)
- 하지만 완전히 해결된 것은 아니다.

3. GRU (Gated Recurrent Units) : LTSM의 강점은 취하고 복잡성을 제거한 모델
   - cell state가 존재하지 않고 hidden state에 합쳐짐
   - LTSM과 같이 게이트를 통해 정보의 흐름을 통제한다
   - Reset gate : 과거의 정보를 리셋한다
   - Update gate : forget + input gate

4. More fancy RNN varients

   Bidirectional RNN, Multilayer RNN 등이 있다.