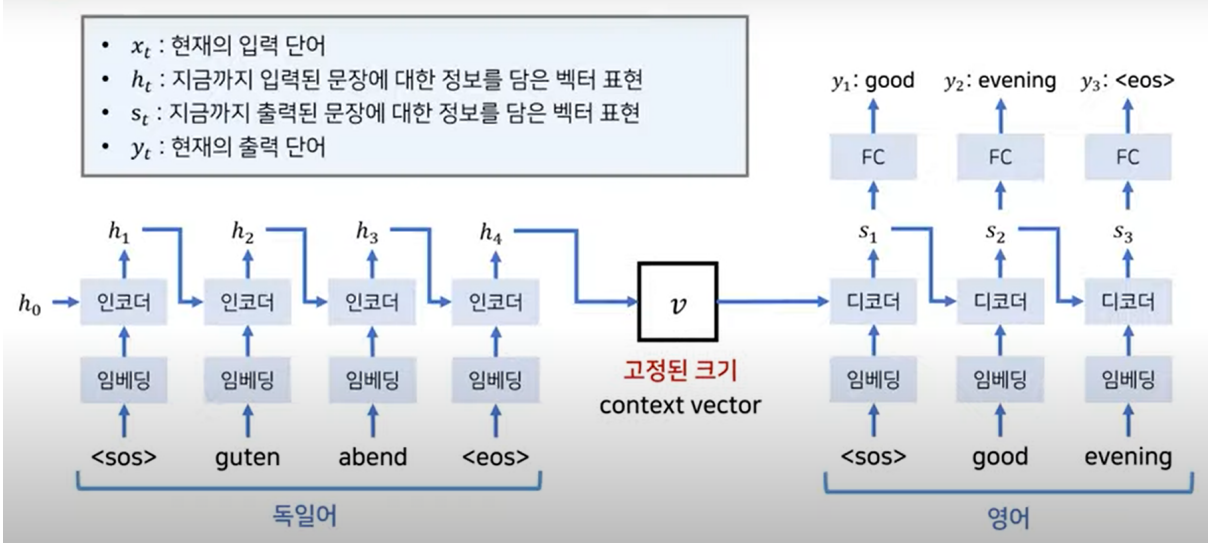
Model 구조



인코더는 RNN or LSTM으로 사용 가능하다.

입력과 출력 size는 독립적이다. (ex) 입력으로 4개의 token이 들어와도 , output은 가변적

대신, Context Vector가 고정되어 있다. ⇒ train시에는 상대적으로 짧은 문장만 들어오다가 test시, 긴 문장이 들어왔을 때, bottlenect의 문제가 발생할 수 있다.

위 그림에서 h4는 이전 token들의 정보를 모두 담고 있다. (단 시점이 t=1에 가까울수록 상대적으로 영향력이 적음) ⇒ 일반적인 언어 체계에서 앞쪽에 위치한 단어끼리 연관성이 높기 때문에, input의 순서를 바꾸면 model 성능이 더 올라감.

token을 한 번에 넣는 것이 아닌, 순차적으로 넣는 구조이기 때문에, 인코더와 디코더는 하나의 network이다.

Encoder의 출력값이 context vector이므로 고정 길이 벡터다.

Encoder의 출력값이 왜 vector인가?? Scalar 아닌가?

여러개의 node를 구성하면 vector이다.

한계점

하나의 context vector가 source 문장의 모든 정보를 가지고 있어야 하므로 성능이 저하됨.

- Seq2Seq 모델에 **어텐션(attention)** 매커니즘을 사용합니다.
  - 디코더는 **인코더의 모든 출력(outputs)을 참고**합니다.