과학기술정보통신부
Ministry of Science and ICT

정보통신기획평가원

medici

AiDA Lab.
Artificial Intelligence & Data Analysis Laboratory

2021년 혁신성장 청년인재 집중양성 추경 사업

빅데이터 분야

자연어 처리

Seq2Seq & Attention

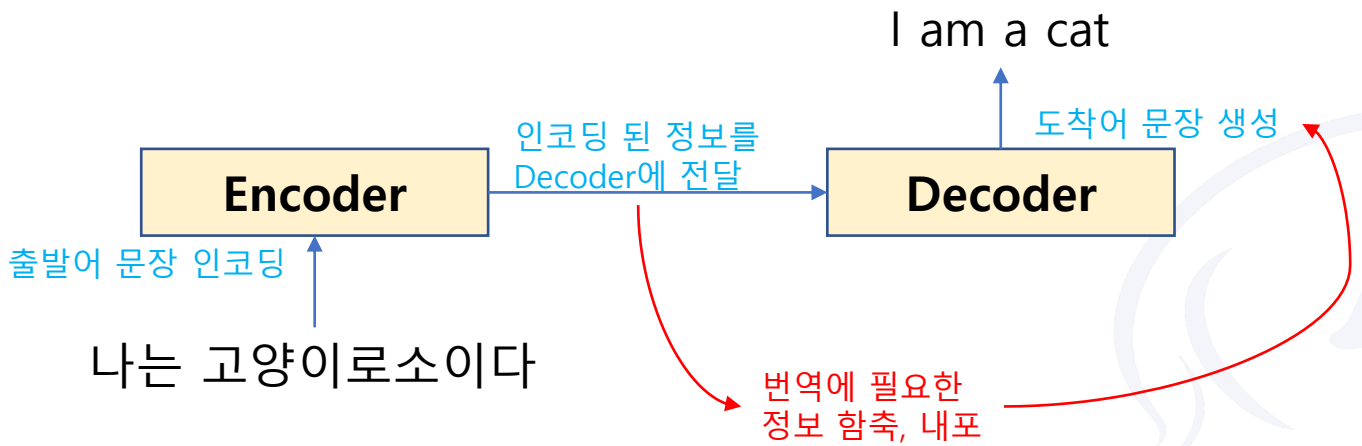
Seokhwan Yang

Seq2Seq

- **Seq2Seq (Sequence to sequence)**
  - 한 도메인(예, 영어 문장)에서 다른 도메인(예, 한국어 문장)으로 시퀀스(Sequence)를 변환하는 모델 학습
  - 주로 시계열 데이터에 대하여 많이 활용됨
  - 2개의 RNN 모델을 이용하여 구성됨
  - Encoder-Decoder 모델이라고도 부름

Seq2Seq

- **Encoder와 Decoder가 번역을 수행하는 예**

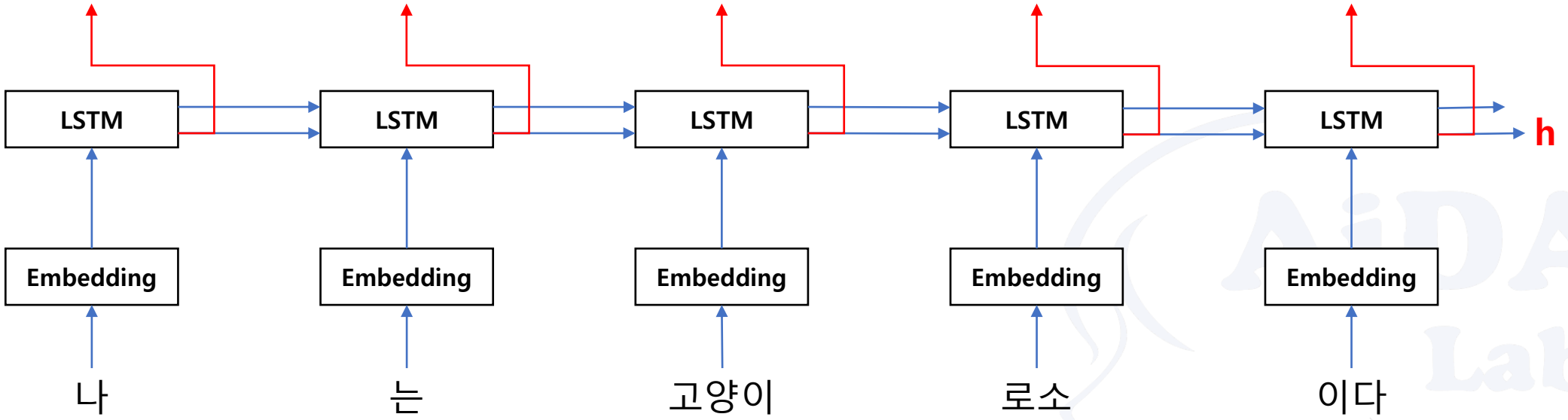


인코딩(부호화) : 정보를 어떤 규칙에 따라 변환하는 것

디코딩(복호화) : 인코딩된 정보를 원래의 정보로 되돌리는 것

Seq2Seq

- **Encoder를 구성하는 계층**

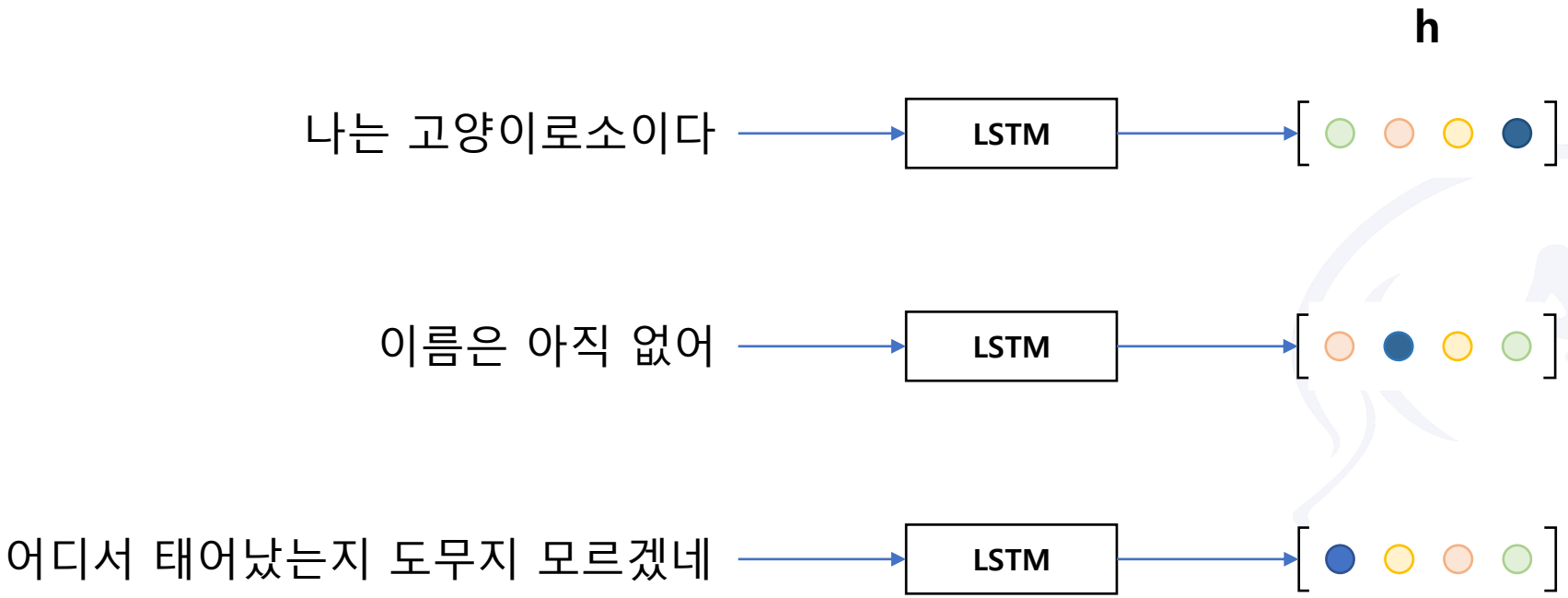


RNN(LSTM)을 이용해서
시계열 데이터를 h라는 은닉상태 벡터로 변환

- 출력벡터 **h**는 LSTM 계층의 마지막 은닉상태
→ 입력 문장을 번역하는데 필요한 정보가 인코딩 됨
- h는 고정길이 벡터
→ 인코딩 작업 = 임의 길이의 문장을 고정길이벡터로 변환하는 작업

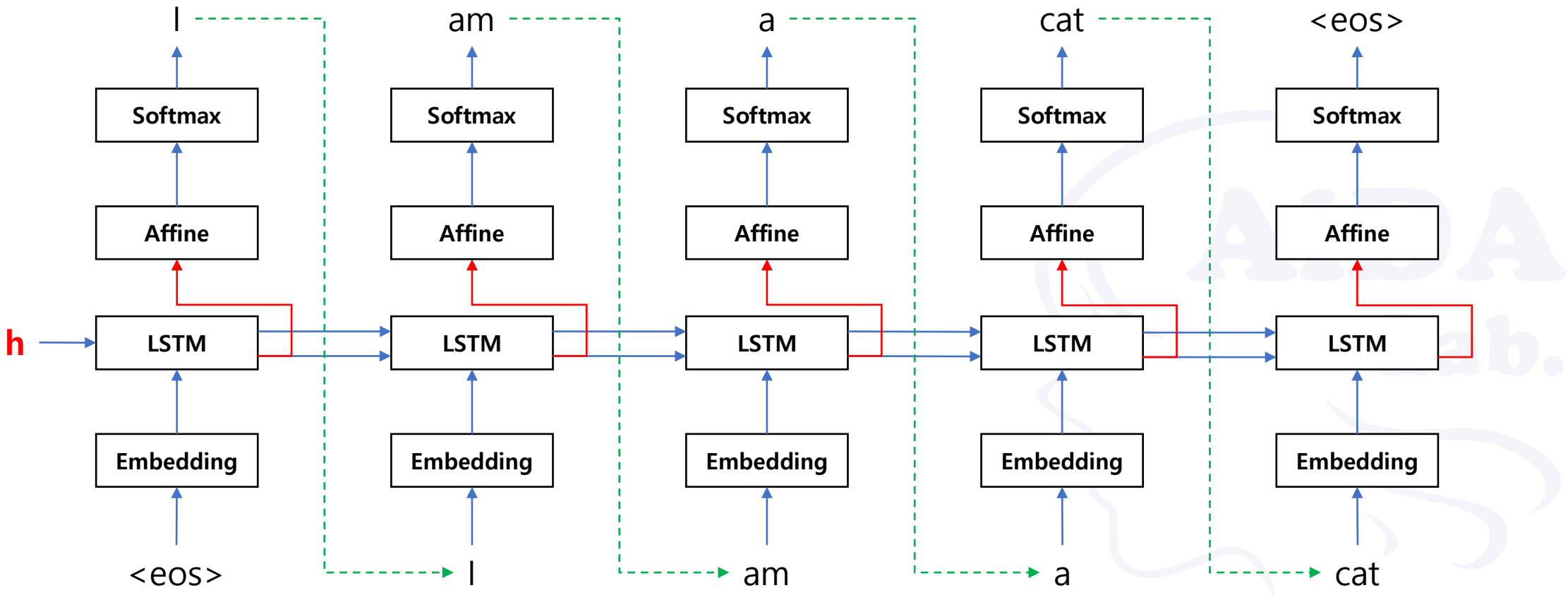
Seq2Seq

- **Encoder는 문장을 고정 길이 벡터로 인코딩한다.**



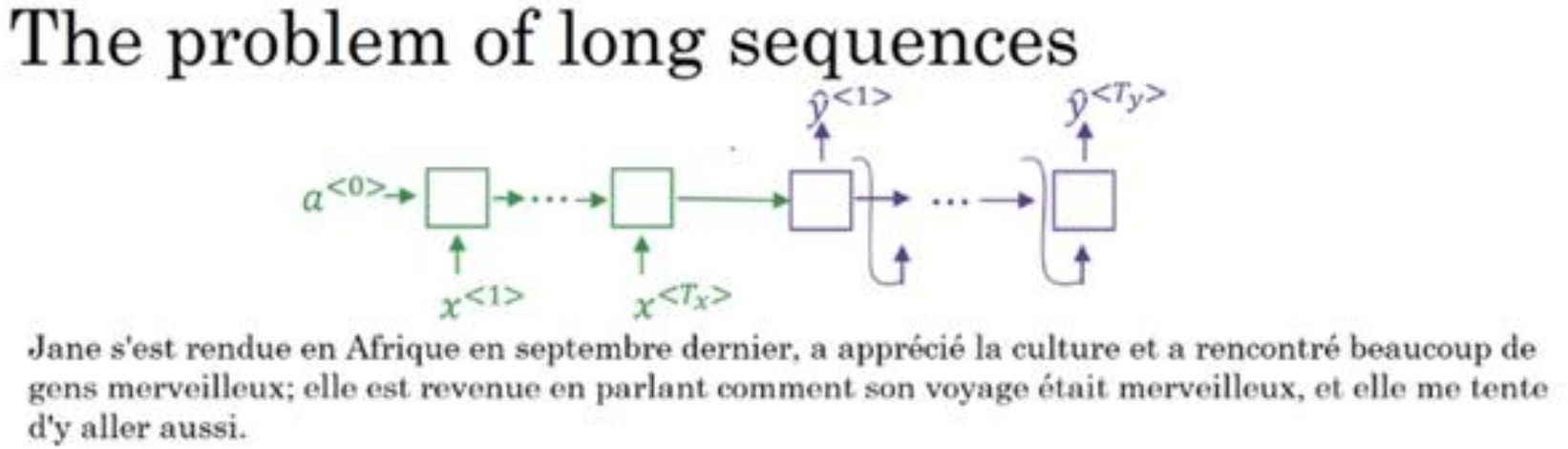
Seq2Seq

- Decoder를 구성하는 계층



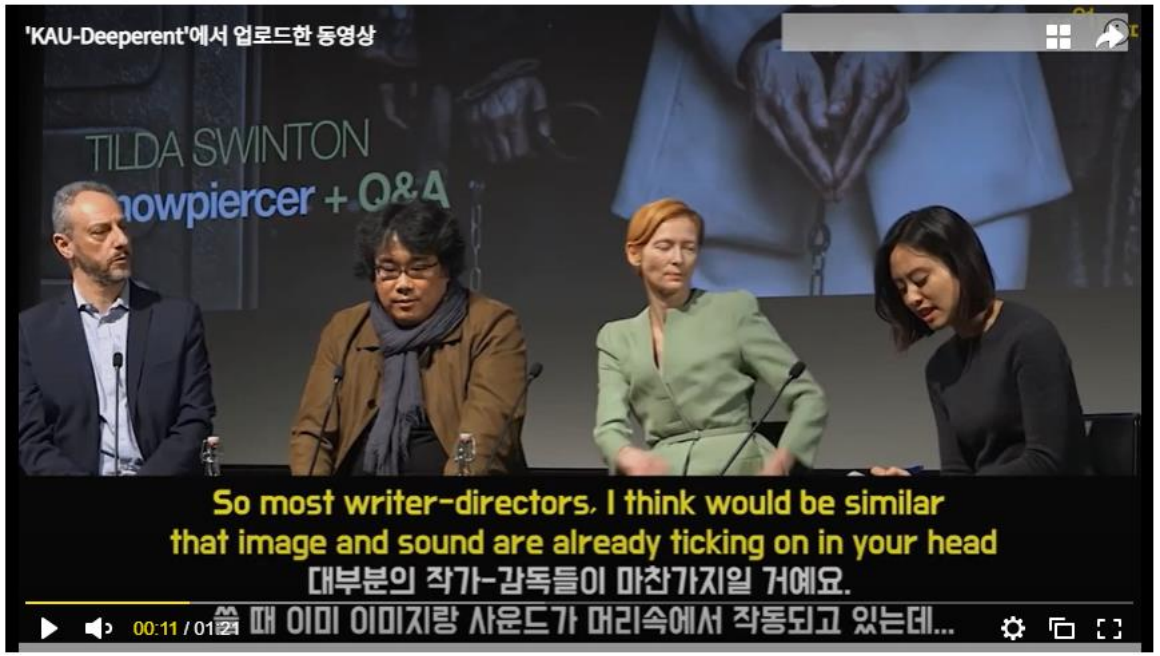
Attention 모델

- 딥러닝 초창기의 기계번역 기술의 주요 방식은 Sequence 방식



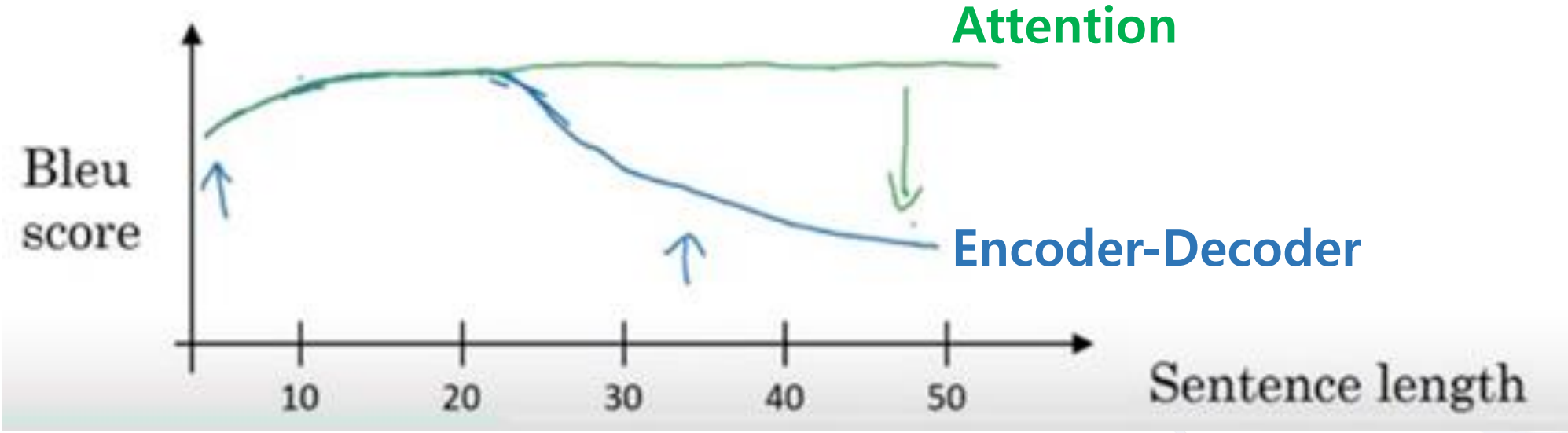
- 데이터를 토큰으로 나눠서 순차적으로 입력 데이터를 준 다음
- 순차적으로 출력을 뽑아내는 방식 → Encoder-Decoder 방식

Attention 모델



- 인간의 경우도 기본적으로 순차 번역을 지향함
- Encoder-Decoder 형 모델
- 그러나 문장이 길어지면(30~40 단어 이상) 성능이 떨어짐

Attention 모델



Attention 모델

- 사람은 번역을 할 때, 모든 문장을 듣지만 각 단어를 번역할 때마다 모든 문장의 정보를 이용하지는 않는다.

- Encoder-Decoder 방식에서는 다음 단어를 번역할 때마다 C 라는 문맥 벡터를 전달하는데, 이 고정된 길이의 벡터에 그 동안 본 모든 단어에 대한 정보가 축약되어 있음 → 문장이 길어지면 효율 저하

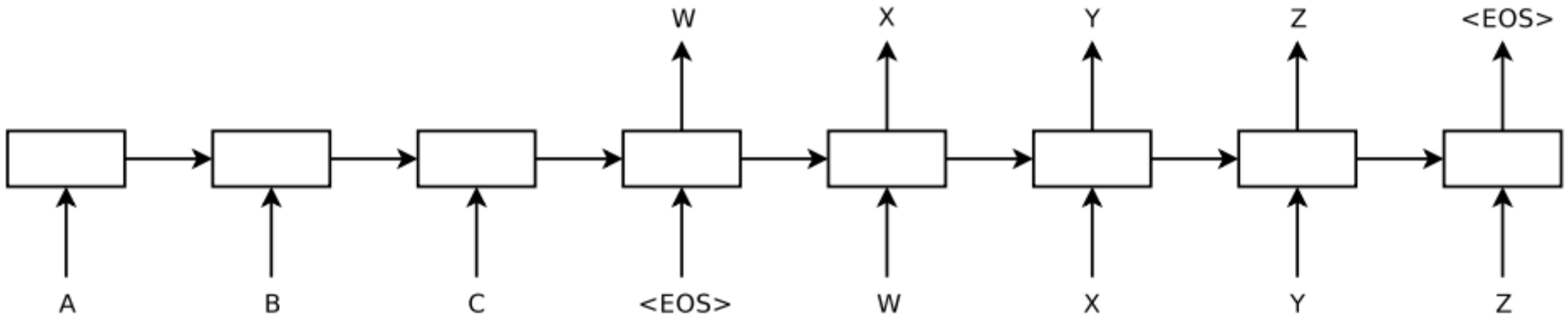
Attention 모델

- 고정된 길이의 C 벡터에 모든 정보를 축약하지 말고, 매 Step마다 필요한 정보를 새로 만들자! → 해결/개선안 제시

→ Attention 모델

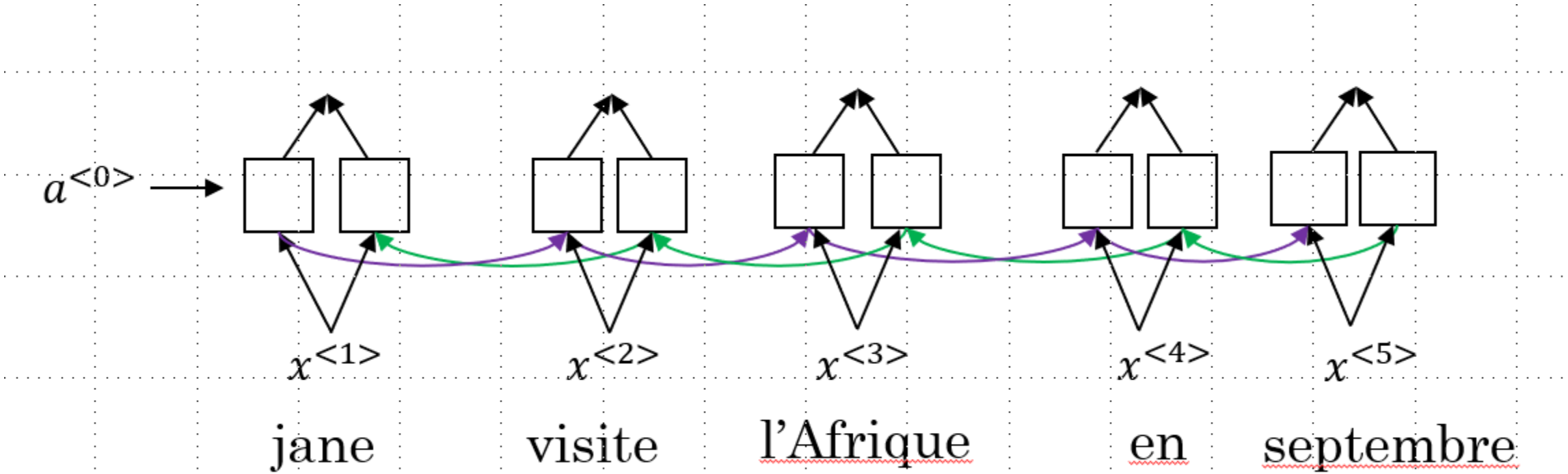
Attention 모델

- RNN 모델에 대한 Attention 기법 적용
  - 기존의 encode는 word-for-word translation. 단어를 하나 번역하면 그 단어를 넘기는 방식



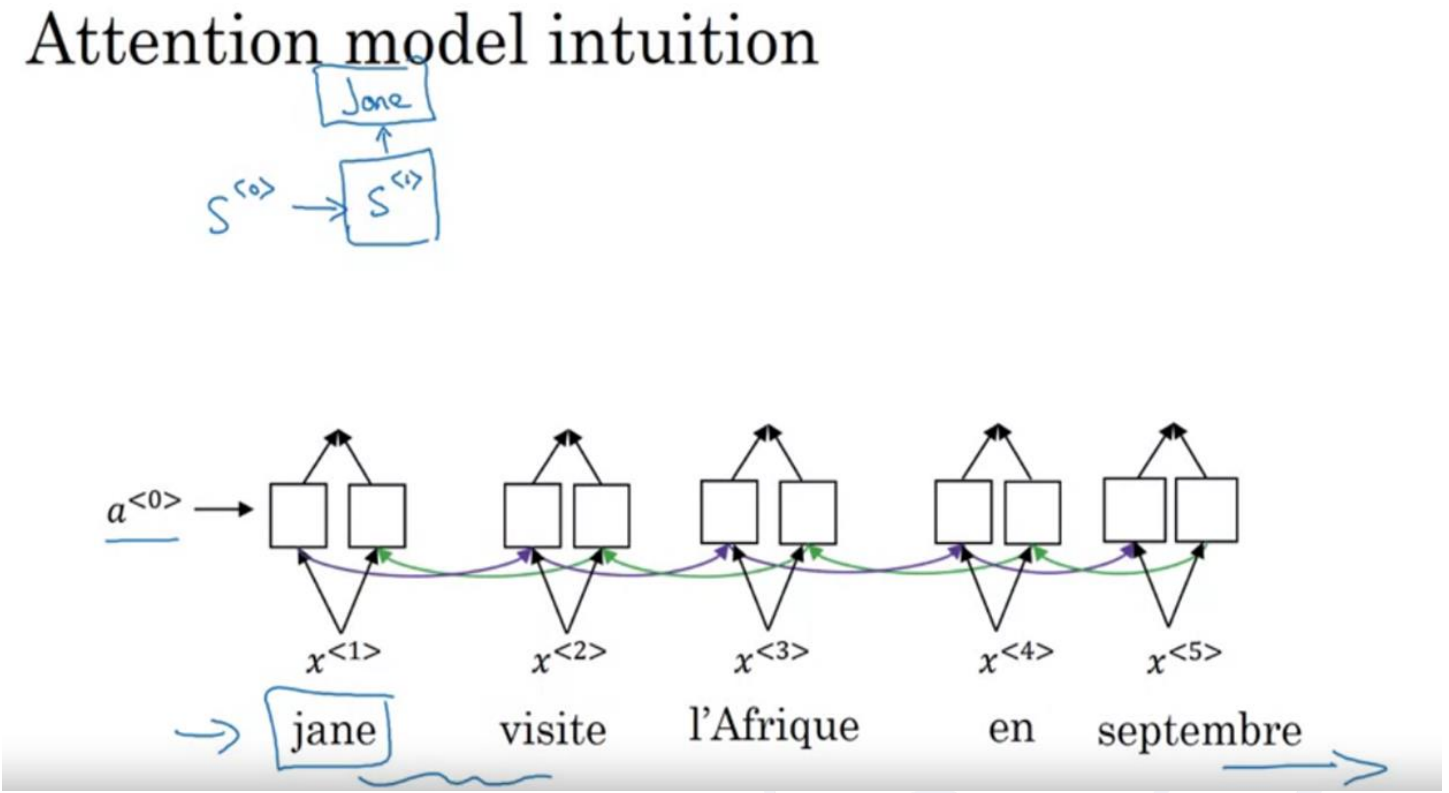
Attention 모델

- RNN 모델에서의 번역



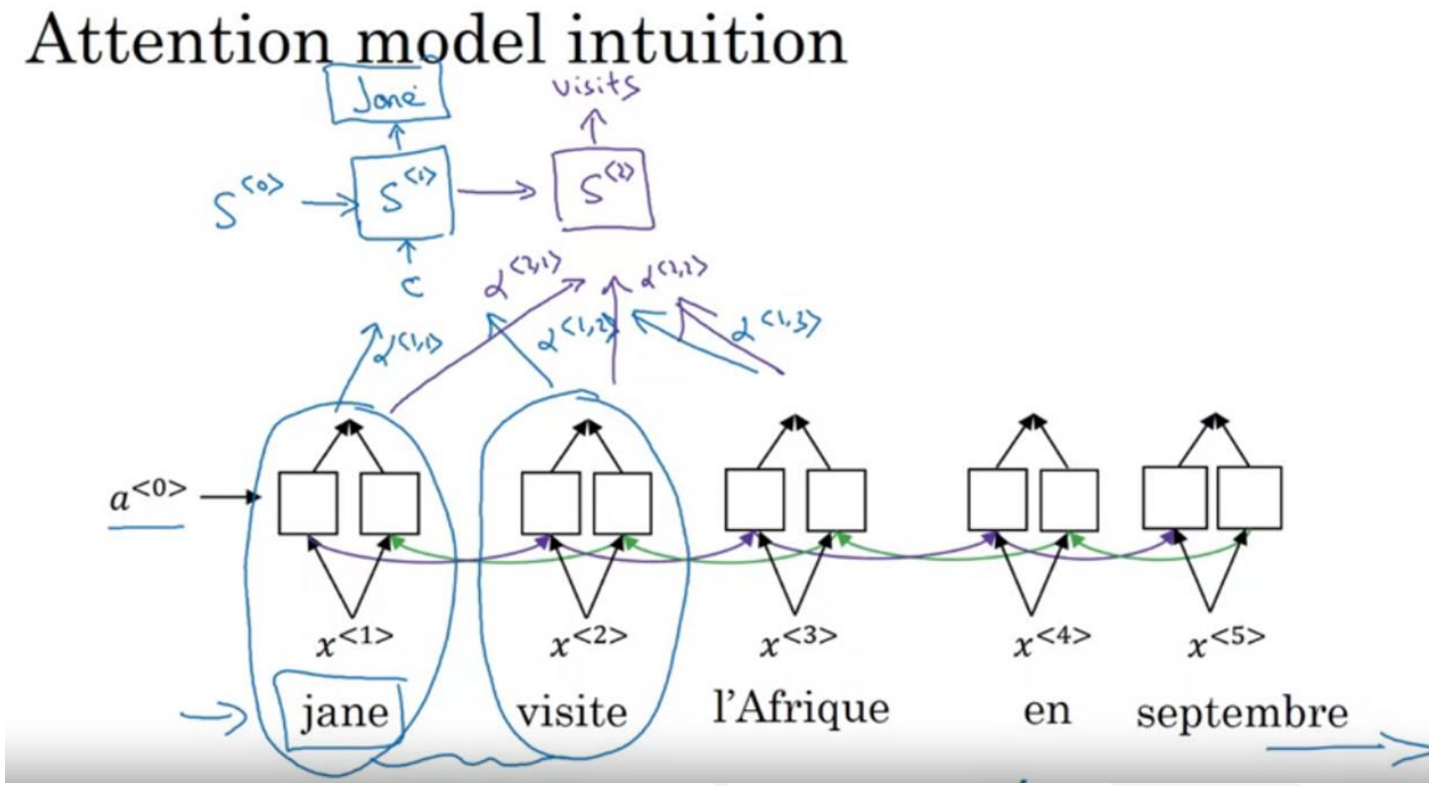
Attention 모델

- **RNN 모델에서의 번역**
  - S: Hidden State
  - 우리는 첫번째 단어가 jane 이 될 것을 기대
  - 목표는 Jane visits to Africa September
  - 여기서 Jane이라는 이름의 결과물을 내려면 몇 개의 프랑스어 단어를 봐야하나?



Attention 모델

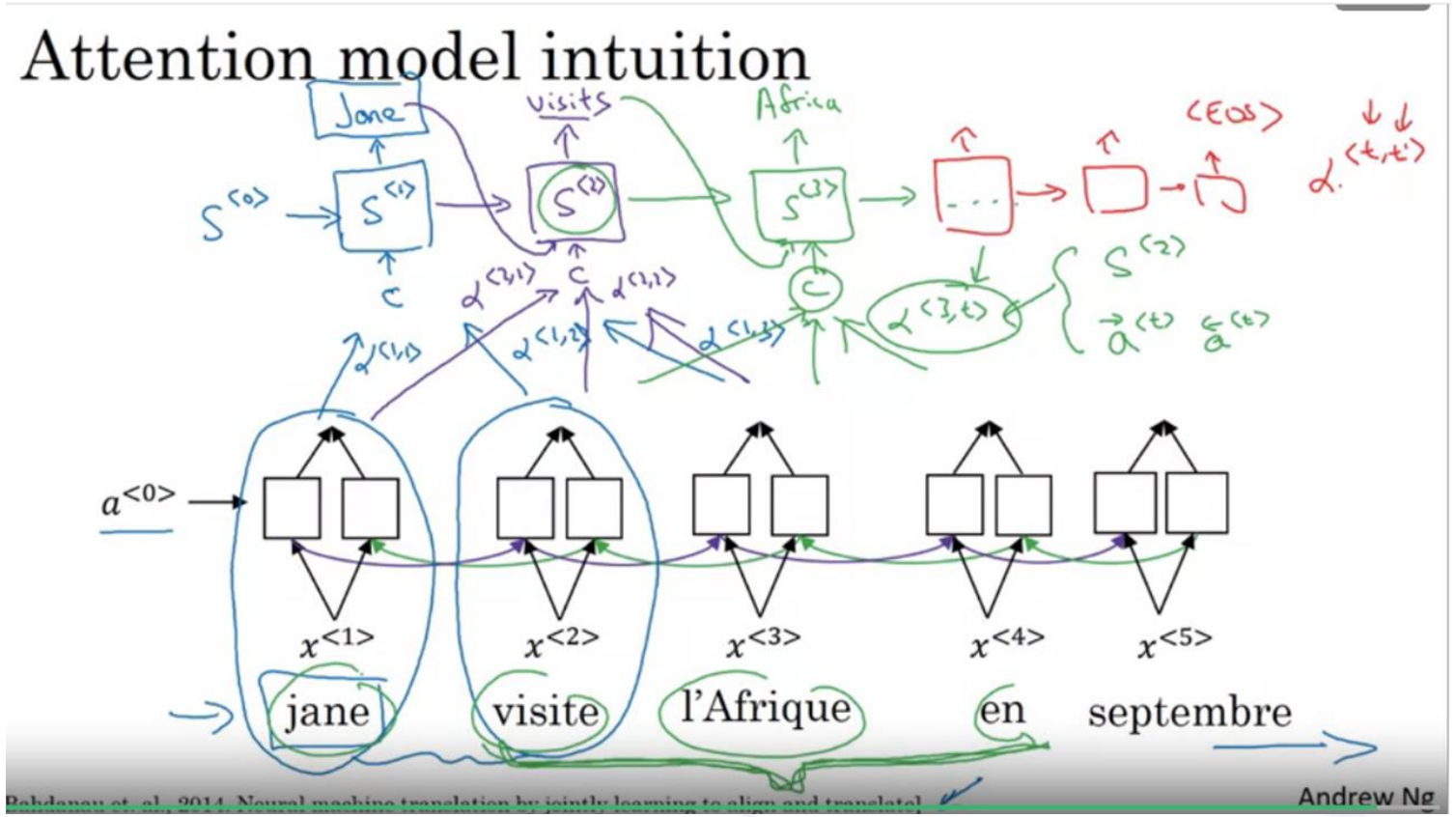
- 문장의 끝까지 볼 필요는 없음
- Attention 모델이 계산해야 할 것은 Attention Weight (A(1,1)로 표기)
- A(1, 1)에서 첫 번째 1은 첫 번째 단어를 생성하는데 얼마나 가중치를 부여할 것인가? 두 번째 1은 첫 번째 정보를 사용하겠다라는 의미



- A(1, 2): 첫번째 단어를 생성하는데 두 번째 정보를 사용하겠다. 의 의미
- 이런 정보들이 합쳐져서 맥락을 나타내는 벡터 C를 계산함

Attention 모델

- 이런 구성을 가짐
- A(3, t)의 경우는,
  S(2), A(2, t), A(4, t)의
  영향을 받음



- A(t, t')는 t번째 단어 번역을 할 때, 원문의 T'번째 단어에 얼마나 Attention을 줄 것이냐?
  를 의미하게 됨 → 전체 원문에서 일부, 즉 Local Window에 Attention을 주겠다는 의미