인공지능을 위한 머신러닝 알고리즘

12. 딥러닝 응용 사례

Tacademy

CONTENTS

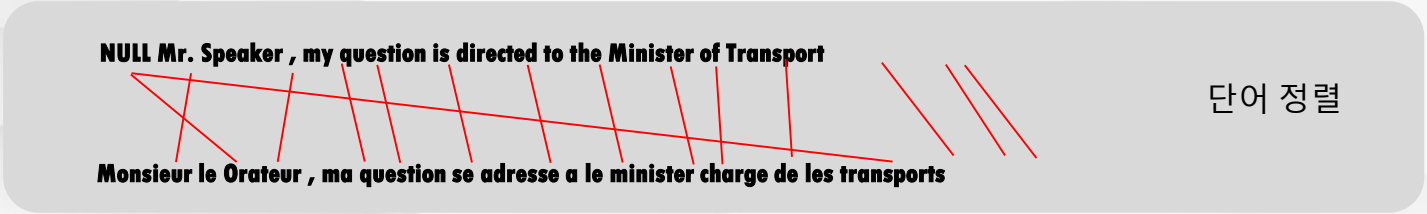
학습 목표

- 딥러닝을 활용한 자동 번역기의 원리를 이해할 수 있다.
- 메모리와 주의집중 기작이 어떻게 딥러닝의 문제를 해결했는지 이해할 수 있다.
- 딥러닝을 활용한 언어와 시각의 결합을 이해할 수 있다.

1. 구글 번역기는 어떤 원리일까?

기존 통계 번역의 문제점

- 과도한 메모리 사용
- 처음 정렬 정보에 따라 번역이 달라짐
- 단어/구문 번역 모델은 의미가 비슷한 단어나 구문을 고려하지 않고, 표면적인 동시 등장 횟수만 고려



한 개의 신경망을 이용한 문장 모델링

◉ 재현 신경망 (**Recurrent Neural Networks**)은 문장 **X**가 나타날 확률 **P(X)**을 모델링

연속된 데이터들의 집합

$$X = (x_1, x_2, ...x_T)$$

X가 나타날 확률

$$p(X) = p(x_1, x_2, ..., x_T) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1, x_2)...p(x_T \mid x_1, ..., x_{T-1}) = \prod_{t=1}^{T} p(x_t \mid x_{<t})$$

- 재현 신경망은 매시간 단위 **t**마다 다음을 계산

$$p(x_t \mid x_{<t}) = g(h_{t-1})$$
$$h_{t-1} = \phi(x_{t-1}, h_{t-2}) \quad \phi \text{ is non-linear activation function}$$

- 은닉 층 = 컨텍스트 층 = 히스토리 층
- 역전파를 사용하여 모델 파라미터 학습

두 개의 신경망을 이용한 기계 번역

- 두 개의 재현 신경망이 문장 **x**가 주어졌을 때, **y**가 나타날 확률을 모델링

$$P(y_1, y_2, ...y_{T_b} \mid x_1, x_2, ..., x_{T_a})$$

❖ 인코더 재현 신경망

- 재현신경망의 은닉층 h는 컨텍스트(히스토리) 정보를 저장하고 있음 $h_i = \phi_\theta(h_{i-1}, x_i)$
- h_{Ta} 는 전체 입력 문장의 정보를 종합적으로 담고 있음

❖ 디코더 재현 신경망

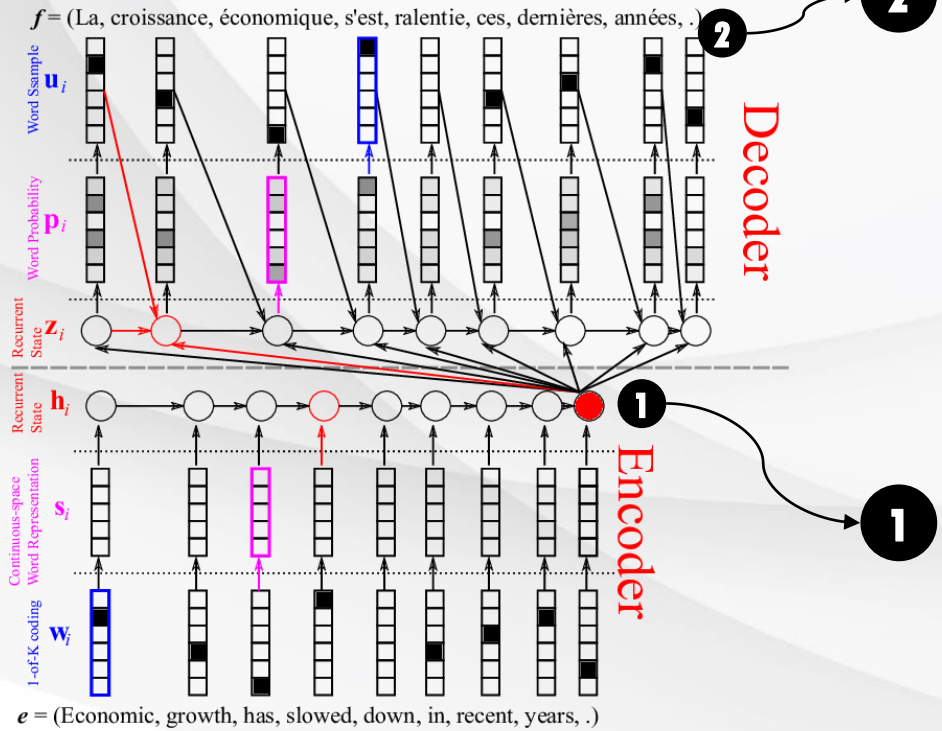
- 각 시간 단위마다, 재현 신경망은 h_{Ta} (입력 문장의 종합적 정보), y_{t-1}(이전 시간 단계에서 생성된 단어), Z_{t-1} (디코더 재현 신경망의 은닉 유닛 정보)를 기반으로 다음 단어 $\mathbf{y_t}$를 예측

$$z_t = \phi_{\theta'}(h_{T_a}, y_{t-1}, z_{t-1})$$
$$p(y_t \mid y_{<t}, X) = g(z_{t-1})$$

1. 구글 번역기는 어떤 원리일까?

두 개의 신경망을 이용한 기계 번역

$$P(y_1, y_2, \ldots y_{T_b} \mid x_1, x_2, \ldots, x_{T_a})$$



디코더 재현 신경망은 매시간 단위마다 h_{Ta}와 이전 시간까지 생성한 단어 정보를 기반으로 새로운 단어 생성

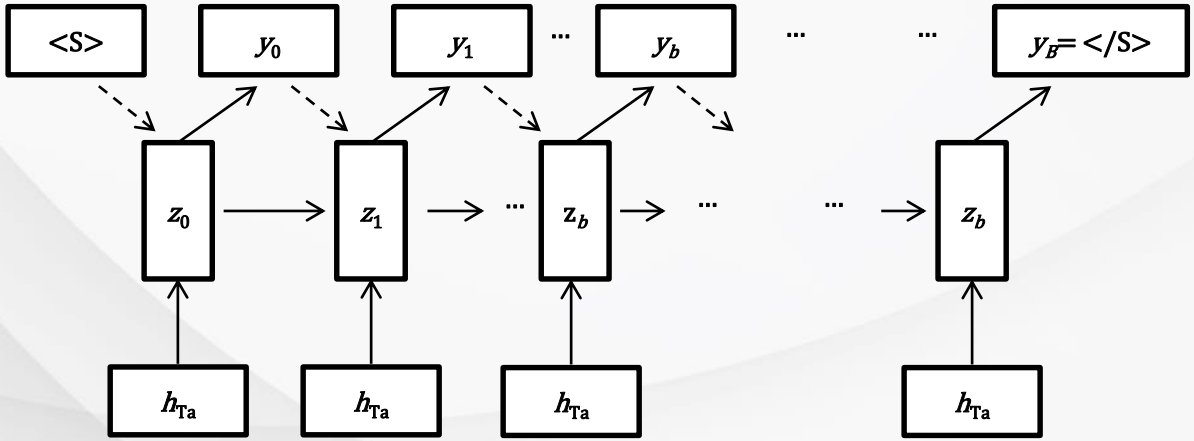
$$z_t = f(U_d y_{t-1} + W_d z_{t-1} + Ch_{Ta})$$
$$y_{t-1} = \text{softmax}(Vz_t)$$

인코더 재현 신경망은 입력 문장의 단어들을 모두 인코딩한 은닉 벡터 h_{Ta}를 디코더 재현 신경망에 넘겨줌

$$h_t = f(U_e x_t + W_e h_{t-1})$$

https://devblogs.nvidia.com/parallelforall/introduction-neural-machine-translation-gpus-part-2/

디코더 재현 신경망의 모습



- y_b는 사전에 있는 전체 단어 중 확률이 가장 높은 단어가 샘플링 됨
  - **Softmax** 함수에 의해 계산
  - 단어 y_b는 **b+1**번째 시간에서 입력이 됨
- 디코더 재현 신경망은 이미 다른 데이터로 학습이 완료된 **(pre-trained)** 재현 신경망의 파라미터를 초기 파라미터로 설정할 수 있음

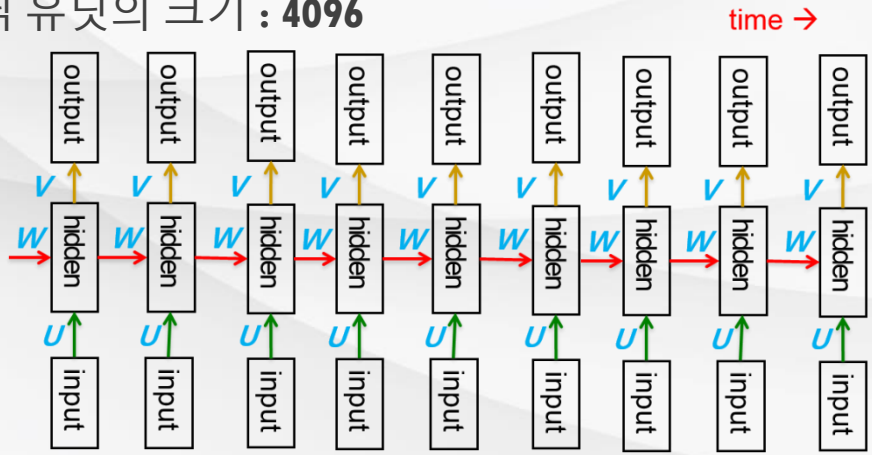
신경망을 이용한 기계 번역의 특징

- 경사하강 법을 이용하여 입력층부터 출력층까지 한꺼번에 종단 학습 가능
- 인코더와 디코더 재현 신경망은 원본 문장과 타겟 문장을 분산 표현으로 나타낼 수 있음
- 번역 문제를 의미적 공간 (은닉 유닛 공간)을 사용하여 학습할 수 있음
- 기존 통계적 기계학습의 방식과 달리 미리 정의된 단어 정렬 방식을 사용하지 않음
- 개념적으로 이해하기 쉬운 디코딩 방식 사용, 음성 인식과 비슷한 크기의 복잡도
- 통계적 기계학습에 비해 적은 모델 파라미터 개수 사용 → 더 적은 메모리 사용

2. 메모리와 주의 집중 기작

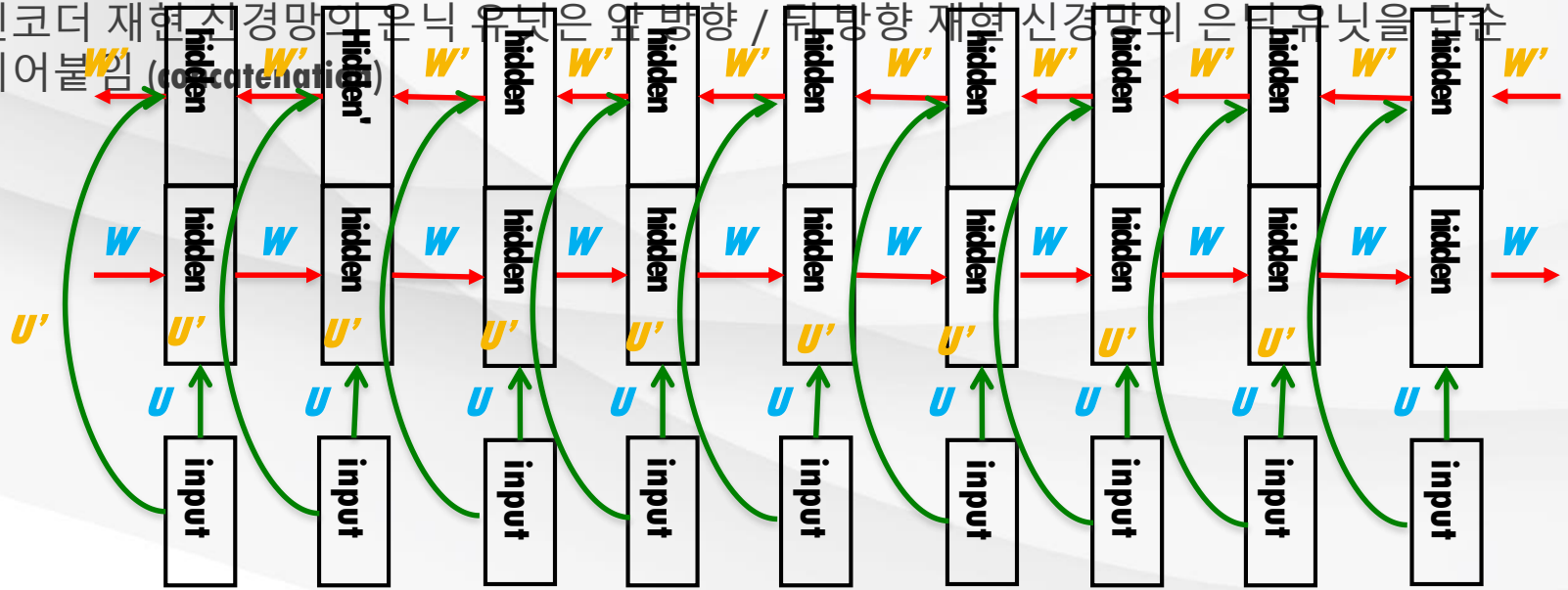
문제점: 제한된 은닉 유닛의 크기

- 인코더 재현 신경망의 가장 마지막 시간의 은닉 유닛 h_{Ta}가 입력 문장의 모든 정보를 담고 있어야함
- 입력의 길이 Ta가 길어질 경우 인코딩해야 할 정보가 많아짐
- 최근에 입력된 정보는 잘 기억하지만, 오래전에 입력된 정보는 손실될 수 있음
  → 번역에서는 큰 문제
- 주로 사용되는 은닉 유닛의 크기 : **4096**



해결책: 양방향 재현 신경망 (bi-directional RNN)

- 앞 방향 재현 신경망: 원본 문장의 가장 앞의 단어부터 인코더 재현 신경망의 입력으로 주어짐
- 뒤 방향 재현 신경망: 원본 문장의 가장 뒤의 단어부터 인코더 재현 신경망의 입력으로 주어짐
- 인코더 재현 신경망의 은닉 유닛은 앞 방향 / 뒤 방향 재현 신경망의 은닉 유닛을 단순 이어붙임 (concatenation)

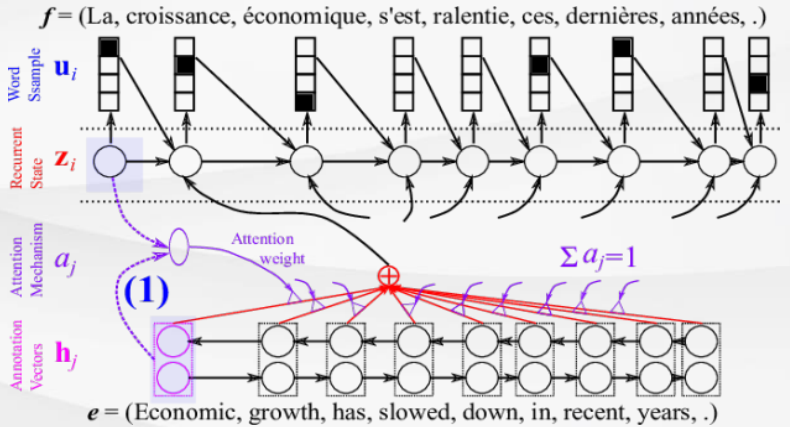


문제점: 메모리와 주의 집중 기작

- 인코더 재현 신경망에서 양방향 재현 신경망에서 계산한 은닉 유닛들을 메모리에 저장
  - 입력 문장의 전체 길이만큼의 은닉 유닛 개수가 저장됨
  - 오래전에 입력으로 주어진 단어의 정보도 갖고 있음

문제점: 메모리와 주의 집중 기작

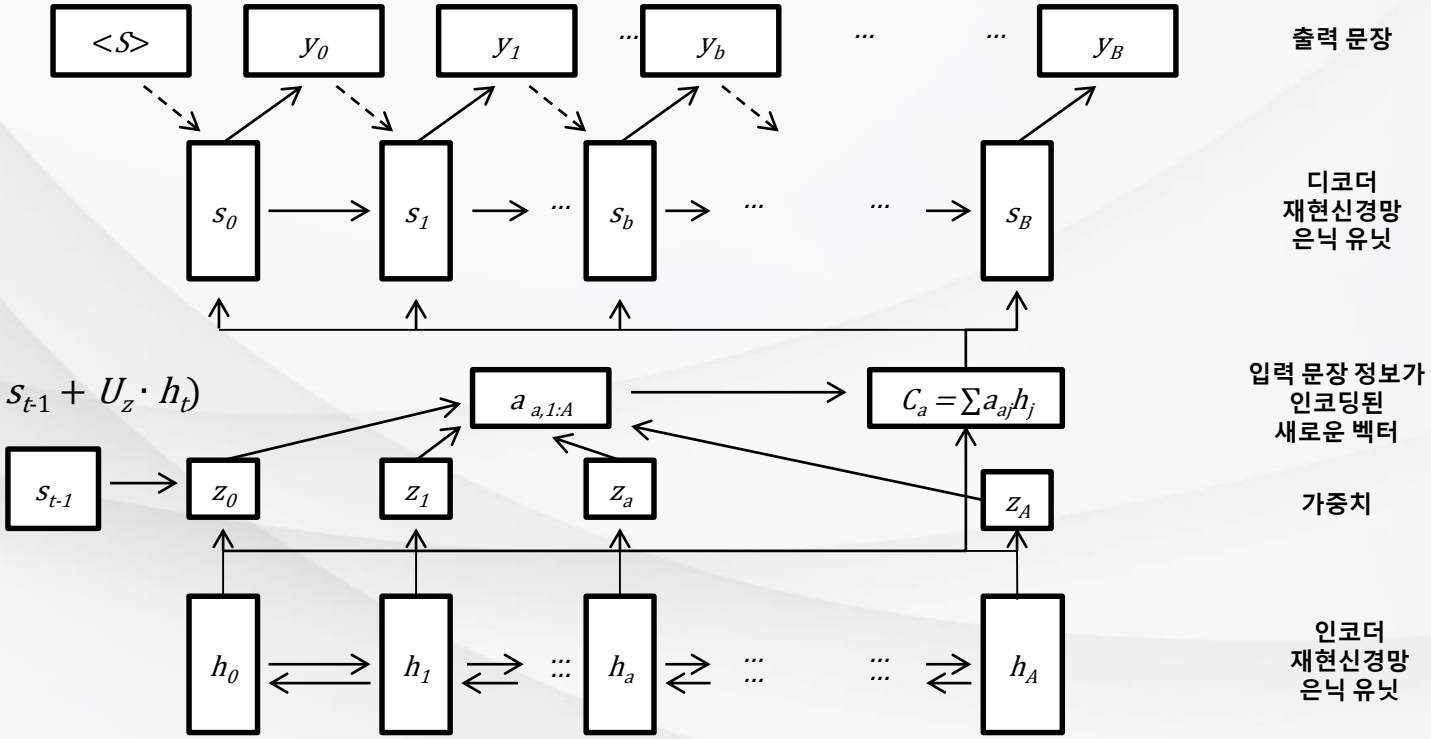
- 디코더 재현 신경망의 입력으로 인코더 재현 신경망 은닉 유닛들의 선형 조합이 사용됨
  - 인코더 신경망의 각 은닉 유닛은 입력 문장에서 한 개의 단어 정보를 인코딩
  - 은닉 유닛들의 가중치의 합: 1
  - 번역기에서 출력되는 단어들은 가중치에 따라 입력으로 주어진 단어들을 선별적으로 고려 (주의 집중)



https://devblogs.nvidia.com/parallelforall/introduction-neural-machine-translation-gpus-part-3/

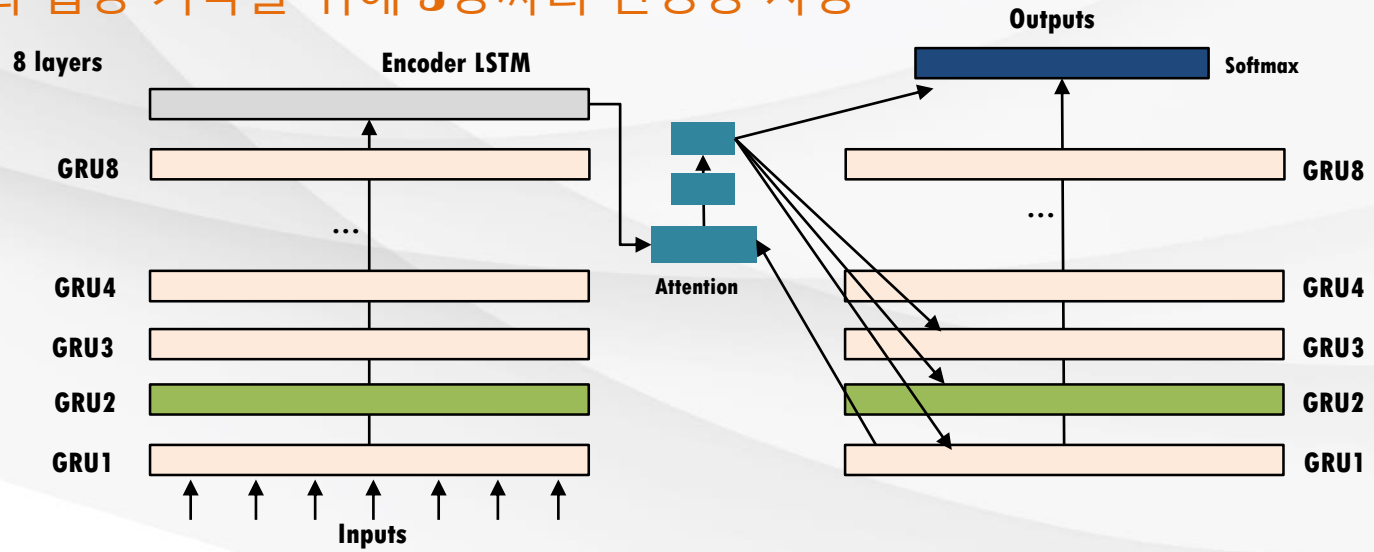
문제점: 메모리와 주의 집중 기작

$$z_t = W \cdot \tanh(V_z \cdot s_{t-1} + U_z \cdot h_t)$$

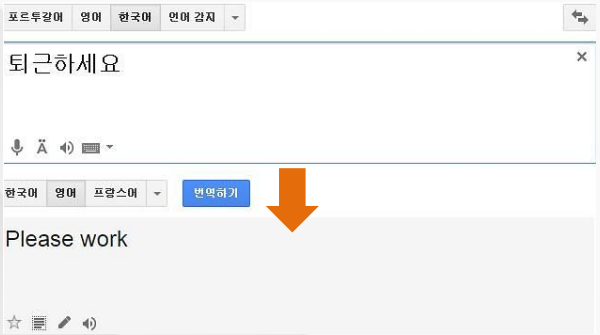


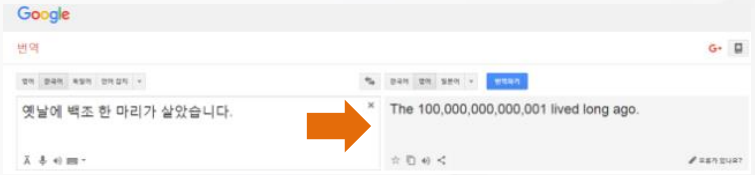
구글의 자동 번역기 (GNMT) (2016.11)

- 재현 신경망 대신 GRU 사용
  - 인코더용 8층 GRU (2층 GRU는 역방향)
  - 디코더용 8층 GRU
- 주의 집중 기작을 위해 3층짜리 신경망 사용



구글 번역기의 발전 (통계 모델 사용 시)





통계 모델 사용 시

딥러닝 사용 시

However, this invention is not only for the Chinese.

Mr.Zander also hopes to sell the charger to countries that have a poor electiricity supply.

For example, farmers in Senegal use cell phones to check on crop prices, and health workers in South Africa use their phones to check patient records.

그러나, 이 발명품은 중국인을 위한 것이 아닙니다.

mr. Zander는 또한 전기 공급이 부족한 국가에 충전기를 판매하기를 희망하고 있습니다.

예를 들어 세네갈의 농민들은 휴대 전화를 사용하여 작물 가격을 확인하고 South Arica의 의료 종사자는 전화기를 사용하여 환자 기록을 확인합니다.

자동 번역기의 성능 측정 (Precision)

- 정답 번역 문장: **The gunman was shot to death by the police.**
- 계산된 번역 문장:

  - **The gunman was shot kill.** (4/5)
  - **Wounded police jaya of** (1/4)
  - **The gunman was shot dead by the police.** (7/8)
  - **The gunman arrested by police kill.** (4/6)
  - **The gunmen were killed.** (1/4)
  - **The gunman was shot to death by the police.** (9/9)
  - **The ringer is killed by the police.** (4/7)
  - **Police killed the gunman.** (3/4)

- 초록색 = **4-gram** 이상 일치 빨강색 = 일치하지 않음

자동 번역기의 성능 측정

- **BiLingual Evaluation Understudy, BLEU (2002년 IBM의 SMT 그룹에서 제안)**
- 기계번역에서 널리 사용됨
- **BLEU** 계산 방식:
  - p_n: 수정된 **n-gram precision** (일치하는 중복 단어 무시)
  - $p_1, p_2, .., p_n$의 기하 평균
  - **BP: Brevity penalty (c=**번역된 문장의 길이, **r=**정답 문장의 길이)
  - 짧게 생성된 번역 문장일수록 높은 점수를 얻게 되는 현상 방지

$$BLEU = BP \left(\prod_{n=1}^{N} p_n \right)^{\frac{1}{N}}$$

$$BP = \begin{cases} 1 & \text{if } c > r \\ r/c & \text{if } c \leq r \end{cases}$$

  - 주로, **N=4**를 사용

자동 번역기의 성능 측정

- 계산된 번역 문장: **The gunman was shot dead by police .**

  정답 1: **The gunman was shot to death by the police .**
  정답 2: **The gunman was shot to death by the police .**
  정답 3: **Police killed the gunman .**
  정답 4: **The gunman was shot dead by the police .**

- **Precision:** $p_1=1.0(8/8)$, $p_2=0.86(6/7)$, $p_3=0.67(4/6)$, $p_4=0.6\ (3/5)$
- **Brevity Penalty:** $c=8$, $r=10$, $BP=0.8$
- 최종 점수: $\sqrt[4]{1 * 0.86 * 0.67 * 0.6} * 0.8 = 0.613$

3. 이미지를 자동으로 설명해주는 기계

신경망을 이용한 이미지 번역

- 한 개의 컨볼루션 신경망과 재현 신경망이 이미지 I가 주어졌을 때, 설명문 Y가 나타날 확률을 모델링

$$P(y_1, y_2, ...y_{T_b} \mid I)$$

❖ 인코더 컨볼루션 신경망

- 이미지를 여러 등분 $I=(i_1, i_2, ..., i_a, ..., i_A)$으로 나눈 뒤 컨볼루션 신경망으로 부분 이미지 i_a를 인코딩
- 인코딩된 벡터들을 인코더 재현 신경망의 은닉 유닛처럼 사용 (매 시간 단위마다 부분 이미지 i_a 가 입력됨)

$$h_a = CNN(i_a)$$

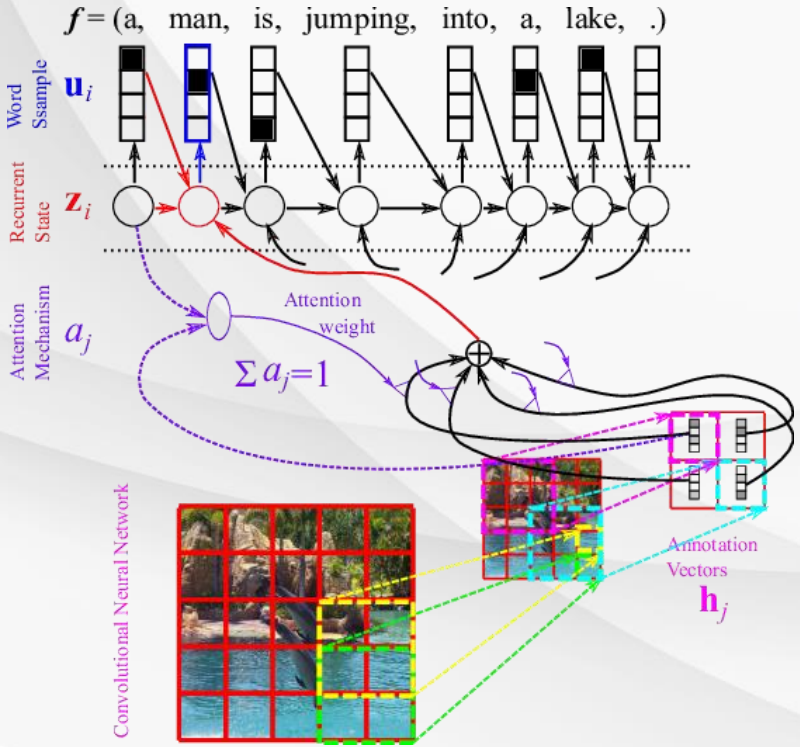
❖ 디코더 재현 신경망

- 기계 번역 세팅과 같이, 각 시간 단위마다, 재현 신경망은 h_a(입력 이미지의 정보), y_{t-1}(이전 시간 단계에서 생성된 단어), z_{t-1}(디코더 재현 신경망의 은닉 유닛 정보)를 기반으로 다음 단어 y_t를 예측

$$z_t = \phi_{\theta'}(h_a, y_{t-1}, z_{t-1})$$

$$p(y_t \mid y_{<t}, X) = g(z_{t-1})$$

신경망을 이용한 이미지 번역



설명문 출력

주의 집중 기작

이미지 인코딩

이미지 분할

https://devblogs.nvidia.com/parallelforall/introduction-neural-machine-translation-gpus-part-3/

이미지 번역의 예시


"little girl is eating piece of cake."
"baseball player is throwing ball in game."
"woman is holding bunch of bananas."
"black cat is sitting on top of suitcase."


Nearest Images
- day + night =
- flying + sailing =
- bowl + box =
- box + bowl =
(Kiros, Salakhutdinov, Zemel, TACL 2015)
Nearest images
- dog + cat =
- cat + dog =
- plane + bird =
- man + woman =

학습정리

지금까지 [딥러닝 응용 사례]에 대해서 살펴보았습니다.

구글 번역기는 어떤 원리일까?

인코더-디코더 모델 사용
인코더 재현 신경망: 원본 문장을 연속된 벡터 공간에 임베딩 시킴
디코더 재현 신경망: 원본 문장의 정보 h_{Ta}와 이전 단계의 출력 단어 y_{t-1},
디코더의 은닉 유닛 s_{t-1}을 기반으로 다음 단어 y_t 예측

메모리와 주의 집중 기작

재현 신경망의 은닉 유닛의 크기가 한정되어 있기 때문에 오래전 입력으로 들어온 단어의 정보가 손실됨
인코더 재현 신경망의 은닉 유닛들을 메모리에 저장하고 선형 조합 (주의 집중 기작)

이미지를 자동으로 설명해주는 기계

이미지를 여러 등분으로 나눈 뒤, 부분 이미지들을 컨볼루션 신경망으로 인코딩 시킴
인코딩된 벡터들은 순서대로 디코더 재현 신경망에 입력됨