CH-12
응용

2023.05.09

12.1 대규모 심층 학습 (Large-Scale Deep Learning)

딥러닝의 바탕은 연결주의 철학이다. 즉, 지능적인 행동이 나타나려면 뉴런들이 아주 많아야 함.

따라서 신경망의 크기가 중요하므로 성공적인 딥러닝을 위해서는 큰 신경망을 감당할 수 있는 고성능 하드웨어와 소프트웨어 기반구조가 필요.

12.1.1 빠른 CPU 구현

예전에는 신경망을 컴퓨터 한 대의 CPU 하나를 이용해서 훈련했지만, 요즘은 그런 접근 방식을 비효율적이라고 간주하는 것이 일반적.

따라서 요즘은 대부분 GPU 컴퓨팅을 이용하거나 컴퓨터를 네트워크로 연결해서 다수의 CPU를 활용

12.1.2 GPU 구현

현세대 신경망 구현들은 대부분 GPU에 의존함. 신경망에는 수많은 값을 담을 큰 버퍼가 쓰이는데, 이런 버퍼들은 전통적인 데스크톱 컴퓨터의 캐시 용량을 완전히 넘어설 정도로 크므로, CPU에 비해 GPU가 용이하게 사용된다.

12.1.3 대규모 분산 구현

컴퓨터 한 대의 계산 자원으로는 큰 신경망을 실행하기에 부족할 때가 많다. 훈련 및 추론 작업을 여러 대의 컴퓨터로 분산하는 것이다. 추론 과정에서는 이런 성질을 자료 병렬성이라 부르고, 하나의 자료점을 여러 대의 컴퓨터가 처리하는 것을 모형 병렬성이라고 한다.

12.1.3 대규모 분산 구현

훈련 과정에서의 해결책은 비동기 확률적 경사 하강법이 있다. 또한 매개변수들을 공유 메모리가 아니라 하나의 매개변수 서버로 관리하는 방법도 있다.

12.1.4 모형 압축

추론 비용을 줄이는 핵심 전략은 모형 압축이다. 모형 압축 전략은 원래의 고비용 모형을 더 작은, 저장 비용과 평가 비용이 낮은 모형으로 대체한다.

모형 압축은 주로 과대적합을 피하려고 크기를 키운 모형들에 적용할 수 있다.

12.1.5 동적 구조

자료 처리 시스템의 속도를 높이는 일반적인 전략 하나는, 하나의 입력을 처리하는데 필요한 계산을 서술하는 계산 그래프에 동적 구조가 존재하는 시스템을 구축하는 것

12.1.6 심층망 전용 하드웨어 구현

CPU나 GPU에서 실행되는 소프트웨어 구현들은 일반적으로 32비트나 64비트의 정밀도로 부동소수점 수를 표현하지만, 적어도 추론 시점에서는 그보다 낮은 정밀도를 사용하는 것이 가능 -> 따라서 NPU나 TPU같은 전용 하드웨어들이 구현되었음

12.2 컴퓨터 비전

컴퓨터 비전을 위한 대부분의 딥러닝 시스템은 어떤 형태든 사물 인식 혹은 물체 검출을 수행한다.

12.2.1 전처리

컴퓨터 비전에 필요한 전처리는 이미지들을 크기로 하여금 통일하는 것뿐이다. 또한 data augmentation을 훈련 집합에만 적용하는 일종의 전처리로 볼 수 있다.

모형의 일반화 오차와 크기를 줄이기 위해서는 모델이 감당해야 할 변동의 양을 줄이는 것이다.

12.2.1.1 명암비 정규화

컴퓨터 비전 과제들에서 안전하게 제거할 수 있는 가장 명백한 변동 요인은 이미지의 명암비이다. 딥러닝에서 흔히 이미지 전체 또는 한 영역의 픽셀들의 표준편차를 명암비로 간주한다. 이미지 전체의 명암비는 다음과 같다

$$\sqrt{\frac{1}{3rc}\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{3}\left(x_{i,j,k}-\overline{x}\right)^2}.$$

이미지 전체의 평균 세기는 다음과 같다.

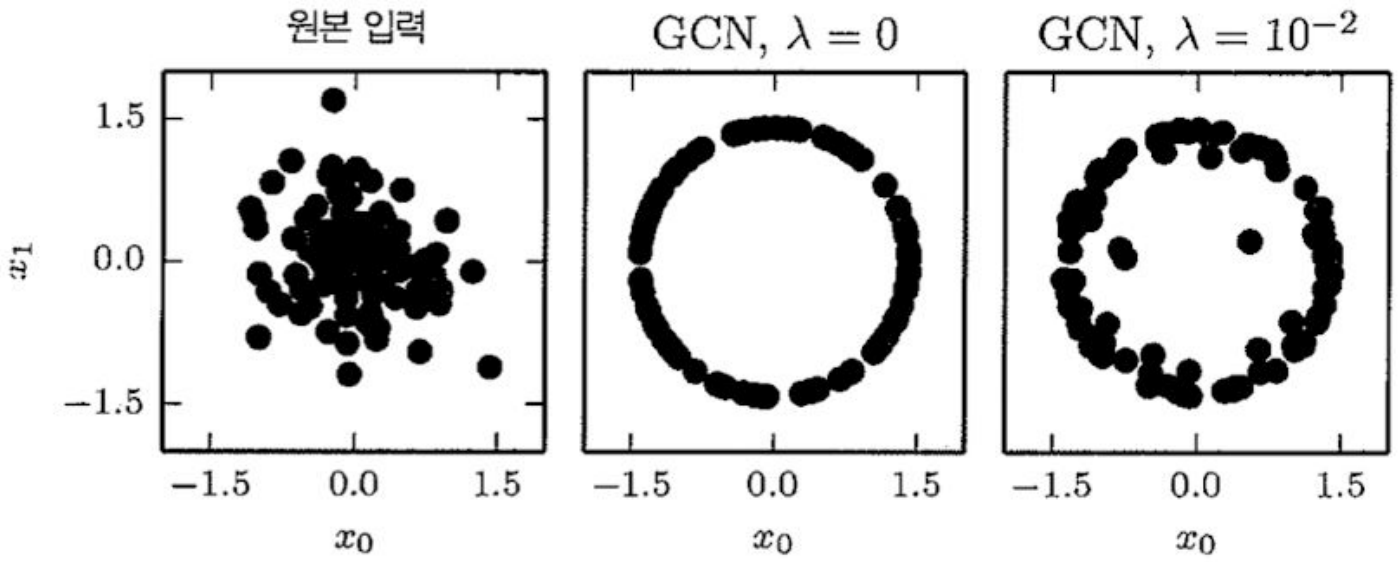
$$\overline{x}=\frac{1}{3rc}\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{3}x_{i,j,k}.$$

12.2.1.1 명암비 정규화

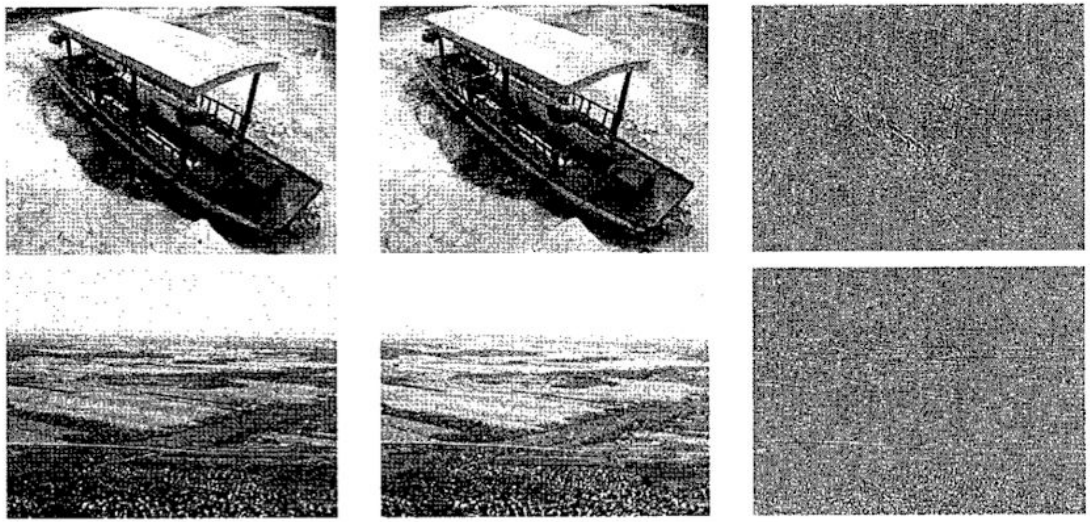
전역 명암비 정규화(global contrast normalization, GCN)는 다수의 이미지들의 명암비 변동을 줄이기 위해, 각 이미지에서 평균 명암비를 뺀 후 이미지 픽셀들의 표준편차가 어떤 축척 상수 s와 같아지도록 적절히 비례한다. 입력 이미지 X에 대해 GCN이 출력하는 X'은 다음과 같다.

$$X'_{i,j,k} = s \frac{X_{i,j,k} - \overline{X}}{\max\left\{\epsilon, \sqrt{\lambda + \frac{1}{3rc}\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{3}\left(X_{i,j,k} - \overline{X}\right)^2}\right\}}.$$

12.2.1.1 명암비 정규화



12.2.1.1 명암비 정규화



12.2.1.2 자료 집합 증강

물체 인식을 위한 분류기 경우에 무작위 이동 변환과 회전 변환, 좌우나 상하를 뒤집어서 새 견본을 만들기도 한다.

12.3 음성인식

음성인식의 과제는 자연어 발화, 즉 입 밖으로 나온 말을 담은 음향 신호를 그에 대응되는, 화자가 의도한 단어들의 순차열로 사상하는 것.

자동 음성 인식(automatic speech recognition, ASR)은 주어진 음향 정보 순차열 X에 대응되는 가장 그럴듯한 언어 정보 순차열 y를 계산하는 f*_ASR를 학습하는 것을 목표로 한다.

$$f^*_{\mathrm{ASR}}(\boldsymbol{X}) = \arg\max_{\boldsymbol{y}} P^*(\mathbf{y}|\mathrm{X} = \mathbf{X}),$$

12.3 음성인식

1980년대부터 약 2009~2012년까지, 최고 수준의 음성 시스템들은 기본적으로 은닉 마르코프 모형(hidden Markov model, HMM)과 가우스 혼합 모형(Gaussian mixture model, GMM)의 조합이었다.

초기에는 신경망을 이용한 ASR 시스템이 많이 있었고, 2009년부터는 비지도 학습에 기초한 일종의 딥러닝 기법을 음성 인식에 적용했다.

2013년 부터는 심층 LSTM 순환 신경망을 훈련하였다. 이후 성능이 대폭 향상하였다.

12.4 자연어 처리

자연어 처리는 컴퓨터가 영어나 한국어 같은 사람의 언어를 사용하게 만드는 것이다. 효율적인 자연어 모형을 구축하려면 순차적인 자료의 처리에 특화된 기법들을 사용해야 한다.

12.4.1 n-그램

언어 모형(language model)은 자연어의 토큰열(sequence of token)들에 관한 하나의 확률 분포를 정의

성공적인 언어 모형들은 n-그램이라고 부르는 고정 길이 토큰열을 사용했다. n-그램은 n개의 토큰들로 이루어진 하나의 순차열이다.

$$P(x_1,\ldots,x_\tau) = P(x_1,\ldots,x_{n-1}) \prod_{t=n}^{\tau} P(x_t \mid x_{t-n+1},\ldots,x_{t-1}).$$

이러한 분해는 확률의 연쇄법칙에 따른 것이며, 초기 순차열 P에 관한 확률 분포를, n 값이 더 작은 또 다른 모형으로 모형화할 수도 있다.

12.4.1 n-그램

보통은 n-그램 모형과 (n-1)-그램 모형을 동시에 훈련한다. 그러면 조건부 확률

$$P(x_t | x_{t-n+1}, \cdots, x_{t-1}) = \frac{P_n(x_{t-n+1}, \cdots, x_t)}{P_{n-1}(x_{t-n+1}, \cdots, x_{t-1})}$$

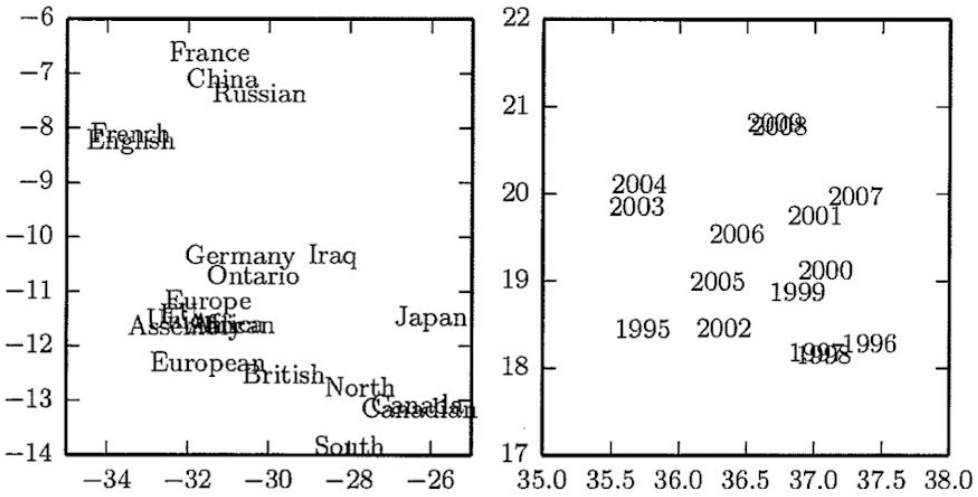
을 계산할 때 그냥 이미 계산해서 저장해 둔 두 확률을 조회해서 나누기만 하면 된다.

$$P(\text{THE DOG RAN AWAY}) = P_3(\text{THE DOG RAN}) P_3(\text{DOG RAN AWAY}) / P_2(\text{DOG RAN}).$$

12.4.2 신경망 언어 모형

신경망 언어 모형(Neural Language model, NLM)은 단어들의 분산 표현을 이용해서 자연어 순차열을 모형화함으로써 차원의 저주를 극복하도록 고안된 일단의 언어 모형들을 아우르는 용어이다.

이는 n-그램 모형과 달리 두 단어가 서로 다른 단어라는 정보를 여전히 유지하는 능력까지 갖추고 있다.



12.4.3 고차원 출력

여러 자연어 처리 응용 프로그램들에서는 문자가 아니라 단어를 출력의 기본 단위로 삼는 것이 바람직하다.

이런 매우 큰 분포를 표현하는 방법은 그냥 은닉 표현에 출력 공간으로의 어파인 변환을 적용한 후 소프트맥스 함수를 적용하는 것이다. 다음과 같은 계산을 수행한다.

$$a_i = b_i + \sum_j W_{ij} h_j \quad \forall i \in \{1, \ldots, |\mathbb{V}|\},$$

$$\hat{y}_i = \frac{e^{a_i}}{\sum_{i'=1}^{|\mathbb{V}|} e^{a_{i'}}}.$$

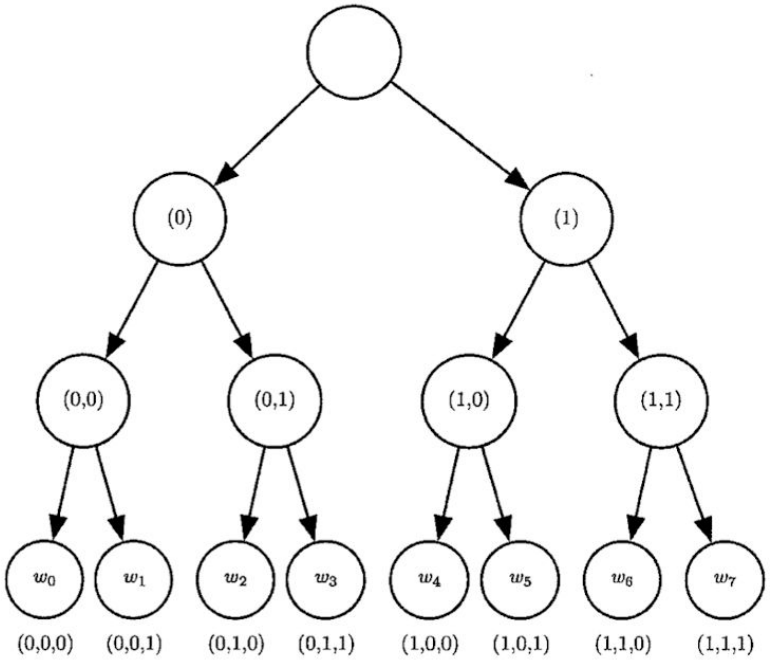
12.4.3.1 후보 목록 활용

어휘 V를 가장 자주 나타난 단어들로 구성된 일종의 후보 목록 L과 그 밖의 단어들로 구성된 '꼬리'목록 T=L/V로 나누고, 전자는 신경망 언어 모형으로, 후자는 n-그램 모형으로 처리한다. 두 모형의 예측과 추가 출력을 이용해서 V의 모든 단어에 관한 확률분포를 추정하는 공식은 다음과 같다.

$$P(y=i|C) = 1_{i\in\mathbb{L}}\ P(y=i|C,i\in\mathbb{L})(1-P(i\in\mathbb{T}|C)) + 1_{i\in\mathbb{T}}\ P(y=i|C,i\in\mathbb{T})P(i\in\mathbb{T}|C).$$

12.4.3.2 계통적 소프트맥스

큰 어휘 V에 대한 고차원 출력층의 계산 부담을 줄이는 고전적인 접근 방식은 확률들을 계통적으로 분해하는 것이다. 그러면 계산횟수를 V에서 log|V|에 비례하는 수준으로 낮출 수 있다.



12.4.3.3 중요도 표집

신경망 언어 모형의 훈련을 가속하는 방법 하나는, 현재 문맥의 다음 위치에 나타날 가망이 없는 모든 단어에 대해서는 기울기 기여도의 계산을 생략하는 것이다. 모든 부정확한 단어에 낮은 확률을 부여해야 마땅하다.

단어들의 한 부분집합만 추출하는 것이 가능하여, 기울기를 다음과 같이 표현할 수 있다.

$$\frac{\partial \log P(y|C)}{\partial \theta} = \frac{\partial \log \mathrm{softmax}_y(\boldsymbol{a})}{\partial \theta}$$

$$= \frac{\partial}{\partial \theta} \log \frac{e^{a_y}}{\sum_i e^{a_i}}$$

$$= \frac{\partial}{\partial \theta} (a_y - \log \sum_i e^{a_i})$$

$$= \frac{\partial a_y}{\partial \theta} - \sum_i P(y=i|C) \frac{\partial a_i}{\partial \theta}.$$

12.4.3.3 중요도 표집

편향 중요도 표집에서는 중요도 가중치들을 그 합이 1이 되도록 정규화한다. 부정적 단어 n_i가 추출되었을 때, 그에 해당하는 기울기에 적용되는 가중치는 다음과 같다.

$$w_i = \frac{p_{n_i} / q_{n_i}}{\sum_{j=1}^{N} p_{n_j} / q_{n_j}}.$$

$$\sum_{i=1}^{|\mathbb{V}|} P(i \mid C) \frac{\partial a_i}{\partial \theta} \approx \frac{1}{m} \sum_{i=1}^{m} w_i \frac{\partial a_{n_i}}{\partial \theta}.$$

12.4.3.4 순위 손실과 잡음 대비 추정

순위 손실 방법은 각 단어에 대한 신경망 언어 모형의 출력을 하나의 점수로 간주해서, 정확한 단어가 다른 단어들보다 더 높은 순위가 되도록 그 단어의 점수를 다른 단어들의 점수들보다 높게 책정하는 것이다. 최근은 잡음 대비 추정이 사용된다.

$$L = \sum_i \max(0, 1 - a_y + a_i).$$

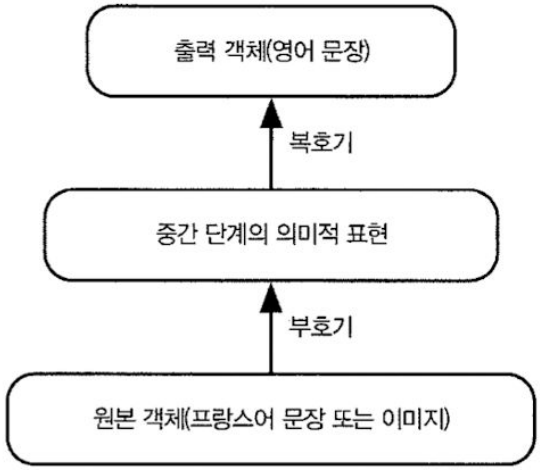
12.4.4 신경망 언어 모형과 n-그램의 조합

n-그램의 장점은 모형의 수용력이 높으면서도 견본 하나의 처리에 드는 계산 비용은 아주 적다는 것이다. 반면 신경망 언어 모형에서는 매개변수의 개수가 두 배가 되면 계산 시간도 대략 두 배가 된다. 이를 결합해 계산량을 크게 증가하지 않고 수용력을 높이는 것이 가능하다. 앙상블 기법을 이용하여 결합하였을 때 좋은 결과가 나왔다.

12.4.5 신경망 기계 번역

기계 번역은 어떤 자연어로 된 문장을 읽어서 다른 자연어로 된 같은 뜻의 문장을 산출하는 과제이다. 주로 n-그램 언어 모형을 사용하다 신경망 언어 모형으로 업그레이드 되면서 경쟁력을 갖추게 되었다.

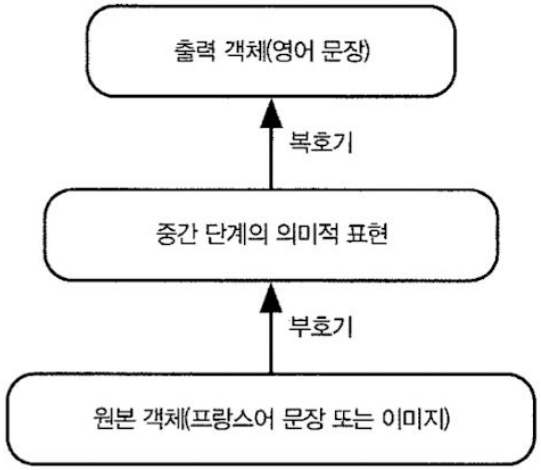
기존에는 그냥 자연어 문장의 확률을 보고하였다.



12.4.5 신경망 기계 번역

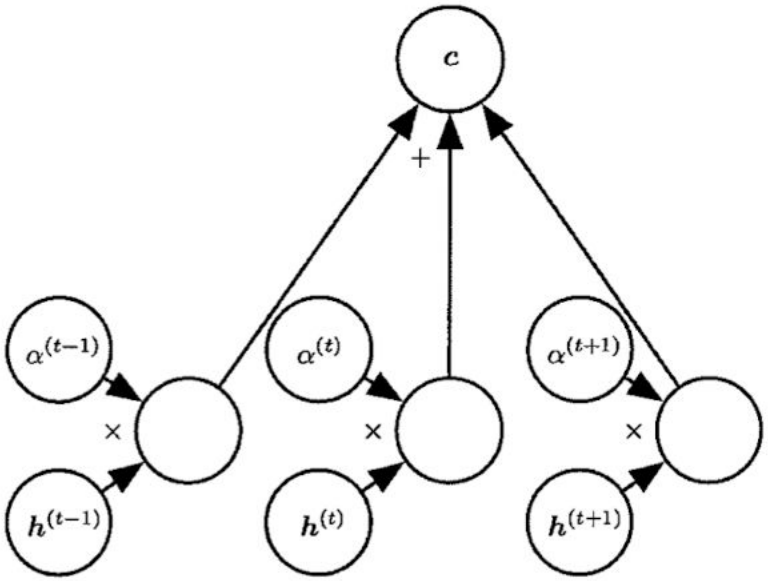
기계 번역은 어떤 자연어로 된 문장을 읽어서 다른 자연어로 된 같은 뜻의 문장을 산출하는 과제이다. 주로 n-그램 언어 모형을 사용하다 신경망 언어 모형으로 업그레이드 되면서 경쟁력을 갖추게 되었다.

기존에는 그냥 자연어 문장의 확률을 보고하였다.



12.4.5 신경망 기계 번역

위의 MLP 기반 접근 방식의 단점은 순차열들을 고정된 길이로 만드는 전처리 과정이 필요하다는 점이다. 그러나, RNN은 가변 길이 출력을 지원해준다.



12.4.5.1 주의 메커니즘과 자료 조각 정렬의 활용

아주 긴 문장을 처리하기 위한 효율적인 접근 방식은 일단 문장 또는 문단 전체를 읽은 후, 각 단예에서 입력 문장의 서로 다른 부분에 주목해서 다음 출력 단어를 산출하는 데 필요한 의미론적 세부사항을 수집하는 식으로 진행하면서 한 번에 한 단어씩 출력해나가는 것이다.

12.4.5.1 주의 메커니즘과 자료 조각 정렬의 활용

주의 기반 시스템은 크게 다음 세 가지 요소로 구성된다.

1. 원본 자료를 읽어서 분산 표현으로 변환하는 판독 과정. 이때 분산 표현은 단어 위치당 하나씩의 특징 벡터들로 구성된다.
2. 판독 과정이 출력한 특징 벡터들을 저장한 목록. 이를 사실들의 순차열을 담은 하나의 기억으로 생각하면 될 것이다. 이후 시스템은 이 기억에 담긴 특정 요소를 임의로 조회할 수 있다.
3. 기억 내용을 활용해서 과제를 순차적으로 처리하는 과정. 각 단계에서 하나의 기억 요소의 내용에 집중할 수 있다.

이 중 번역문을 생성하는 것은 셋째 구성요소이다.

12.4.5.1 주의 메커니즘과 자료 조각 정렬의 활용

한 언어의 문장에 있는 단어들이 다른 언어의 번역문에 있는 단어들과 정렬된다면, 대응되는 단어 내장들을 연관시키는 것이 가능하다.

12.5.1 추천 시스템

추천 시스템은 크게 두 가지로 나뉘는데, 하나는 온라인 광고이고, 다른 하나는 항목 추천이다. 두 응용 모두 사용자와 항목 사이의 연관 관계를 파악하는 능력에 의존한다. 이때 모형은 사용자에게 어떤 광고를 표시하거나 어떤 제품을 추천했을 때 어떤 행동이 발생할 확률을 예측하거나 그에 따른 기대 수익을 예측한다.

12.5.1.1 탐험 대 활용

사용자에게 뭔가를 추천할 때는 보통의 지도 학습의 영역을 벗어나서 강화 학습의 영역으로 넘어가는 문제점 하나가 제기된다. 여기서 문제점이란, 추천 시스템을 이용해서 자료를 수집할 때 편향이 생겨서 사용자의 선호도를 제대로 파악하지 못하게 된다는 것이다.

따라서 강화 학습에서는 탐험과 활용의 적절한 절충점을 찾아야 한다. 활용은 현재까지 학습한 최고의 정책에 기초해서 보상이 가장 클 것이라고 예상하는 행동을 취하는 것을 말하고, 탐험은 훈련 자료를 좀 더 수집하는 행동을 취하는 것을 말한다.

12.5.2 지식 표현, 추론, 질의응답

딥러닝 접근 방식들은 언어 모형화, 기계 번역, 자연어 처리에서 아주 성공적이었는데, 이는 기호들에 대한 내장과 단어들에 대한 내장을 사용한 덕분이다. 그런 내장들은 개별 단어와 개념에 관한 의미론적 지식을 표현한다.

12.5.2.1 지식, 관계 질의응답

한 가지 흥미로운 연구 방향은 두 개체(entity) 사이의 관계(relation)를 포착하도록 분산 표현을 훈련하려면 어떻게 해야 하는가이다. 그런 관계를 표현할 수 있으면, 객체들에 관한 사실이나 개체들 사이의 상호작용 방식을 공식화할 수 있다.

수학에서 이항관계는 두 객체로 이루어진 순서쌍들의 집합으로 정의된다. 어떤 두 객체의 순서쌍이 그 집합에 존재한다면 그 두 객체 사이에는 집합이 나타내는 관계가 있는 것이고, 집합에 없다면 둘은 그런 관계가 없는 것이다.

집합 {1, 2, 3}을 생각해보자 “~은 ~보다 작다” 관계를 순서쌍 집합

S = {(1,2),(1,3),(2,3)}으로 정의할 수 있다.

12.5.2.1 지식, 관계 질의응답

AI의 맥락에서는 관계라는 것을, 구문이 간단하고 고도로 구조화된 언어의 문장으로 생각할 수 있다. 관계는 동사의 역할을 하고, 관계에 대한 두 인수는 주어와 목적어의 역할을 한다.

(주어, 동사, 목적어)

이러한 삼중항은 다음과 같은 값들을 가진다.

(개체i, 관계j, 개체k)

또한 관계와 비슷하되 인수가 하나인 개념은 특성도 정의할 수 있다.

(개체i, 특성j)

예를 들어 '털이 있음' 특성을 '개' 같은 개체들에 적용할 수 있다.