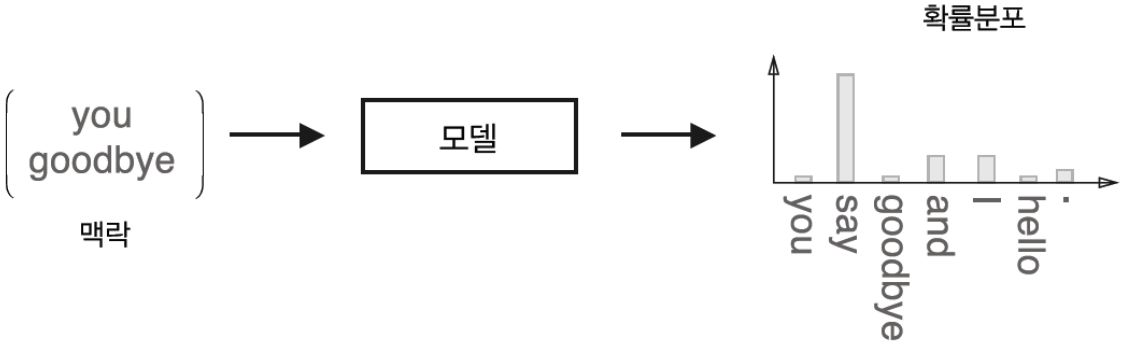
밑바닥부터 시작하는 딥러닝2 CHAPTER 3

word2vec

추론 기반 기법과 신경망

- 분포 가설(단어의 의미는 주변 단어에 의해 형성된다) -> 통계 기반 기법, **추론 기반 기법** / 단어의 동시발생 가능성을 얼마나 잘 모델링하는가
- 통계 기반 기법의 문제점말뭉치 전체의 통계를 이용해 1회의 처리로 단어의 분산 표현 획득말뭉치가 대규모 크기일 경우 행렬의 크기도 대규모가 되고 여기에 O(n^3) 비용을 갖는 SVD를 적용해서 차원 감소를 하는 것도 비효율적
- 추론 기반 기법은 미니배치 학습을 한다대규모 말뭉치 처리 가능여러 머신 & 여러 GPU 병렬 계산 -> 학습 속도 up
- 추론 기반 기법 개요추론 문제 반복 풀기 -> 단어의 출현 패턴 학습타깃이 되는 단어를 추론하기 위해 그 주변 단어들인 맥락을 윈도우 크기에 맞춰서 넣은면 모델, 즉 신경망이 지금까지 학습된 것을 바탕으로 어떤 타깃 단어가 나와야 하는지 확률 분포를 도출한다

그림 3-3 추론 기반 기법: 맥락을 입력하면 모델은 각 단어의 출현 확률을 출력한다.



- 신경망(모델)에서의 단어 처리단어 -> 원핫 벡터입력층과 가중치를 곱하고 편향은 생략하여 matmul만으로 완전연결계층 계산 구현이 가능하다여기서 중요한 점 c를 입력의 원핫 벡터, W를 가중치 행렬이라고 했을 때 matmul(c,W)는 결국 가중치에서의 해당 위치 행벡터 뒤에 나오는 CBOW 모델에서 이 가중치 W는 해당 단어의 분산 표현 > 학습을 진행할수록 이 가중치 W값이 더 적절한 값들로 갱신될 것이고 정확하게 단어들을 잘 추론하는 W값을 찾는 것이 목표!

그림 3-6 완전연결계층에 의한 변환: 입력층의 각 뉴런은 7개의 단어 각각에 대응(은닉층 뉴런은 3개를 준비함)

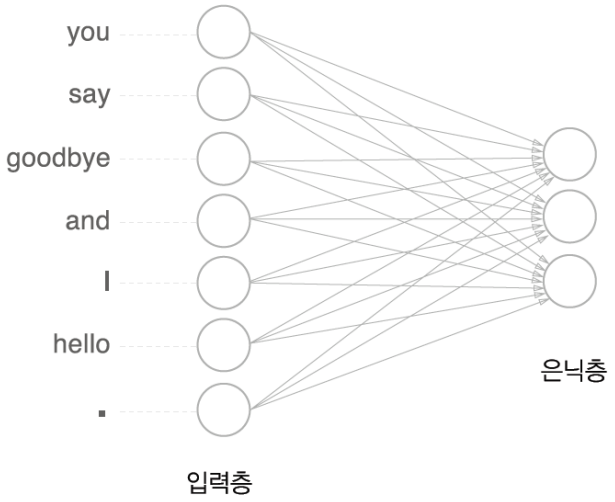
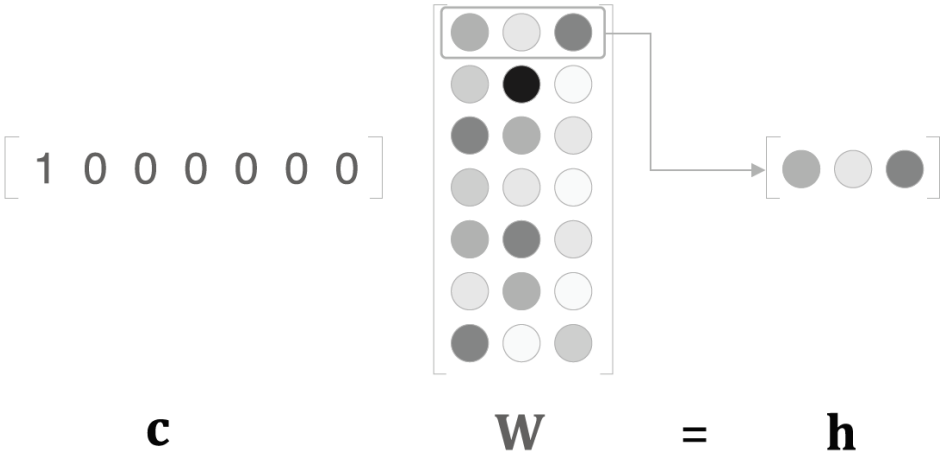


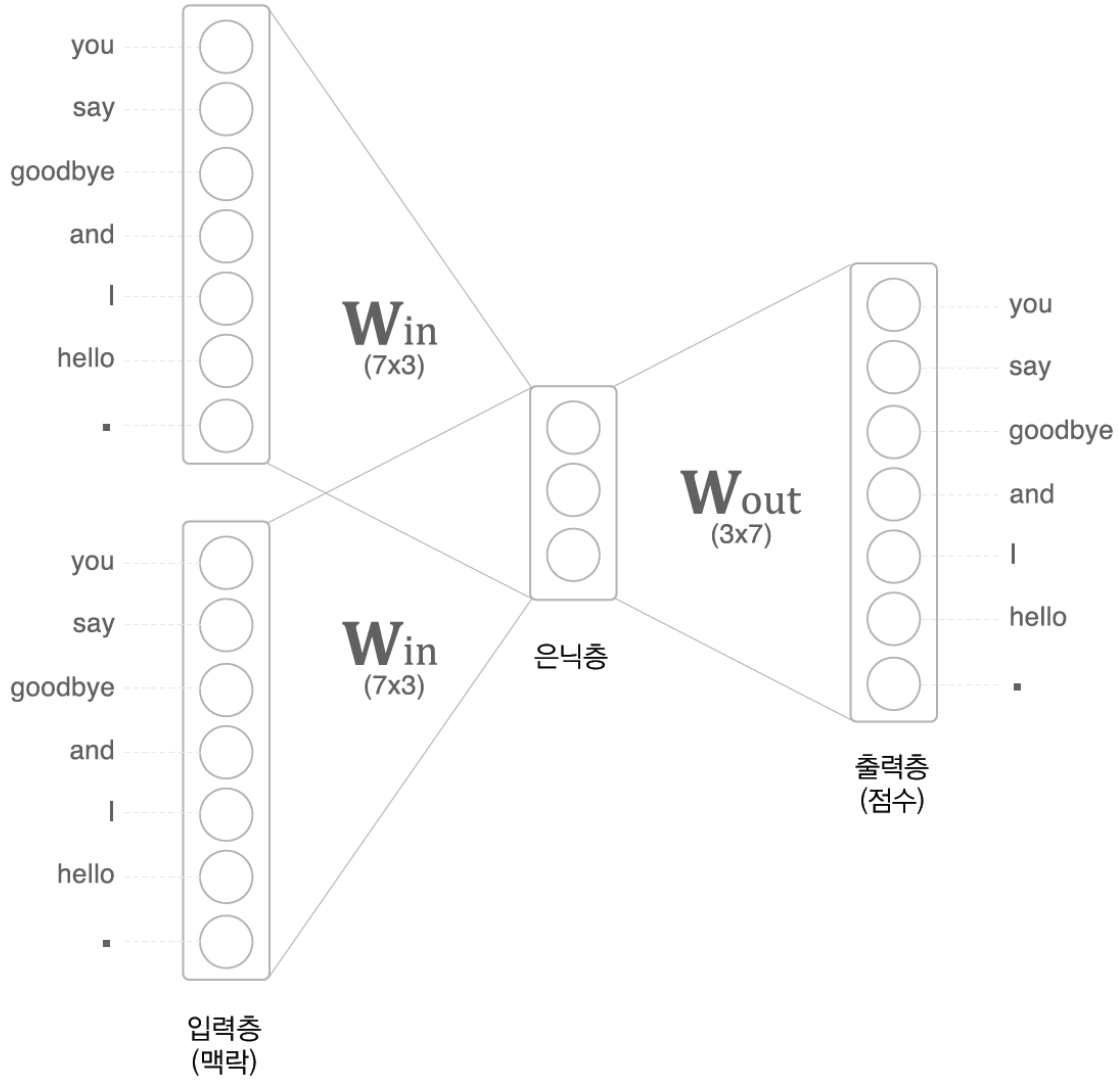
그림 3-8 맥락 $\mathbf{c}$와 가중치 $\mathbf{W}$의 곱으로 해당 위치의 행벡터가 추출된다(각 요소의 가중치 크기는 흑백의 진하기로 표현).



단순한 word2vec

- word2vec에서 사용되는 신경망에는 **CBOW(continuous bag-of-words)** 모델, skip-gram 모델
- CBOW 모델 개요맥락으로부터 타깃을 추론은닉층의 뉴런값은 입력층들의 완전연결계층으로 변환된 값의 평균입력: 맥락, 출력: 각 단어의 점수(점수가 높을수록 확률이 높다)입력층의 뉴런수 > 은닉층의 뉴런수 > 결과적으로 밀집벡터를 얻을 수 있다은닉층의 정보는 인코딩, 디코딩 과정 필요활성화 함수를 사용 X

그림 3-9 CBOW 모델의 신경망 구조



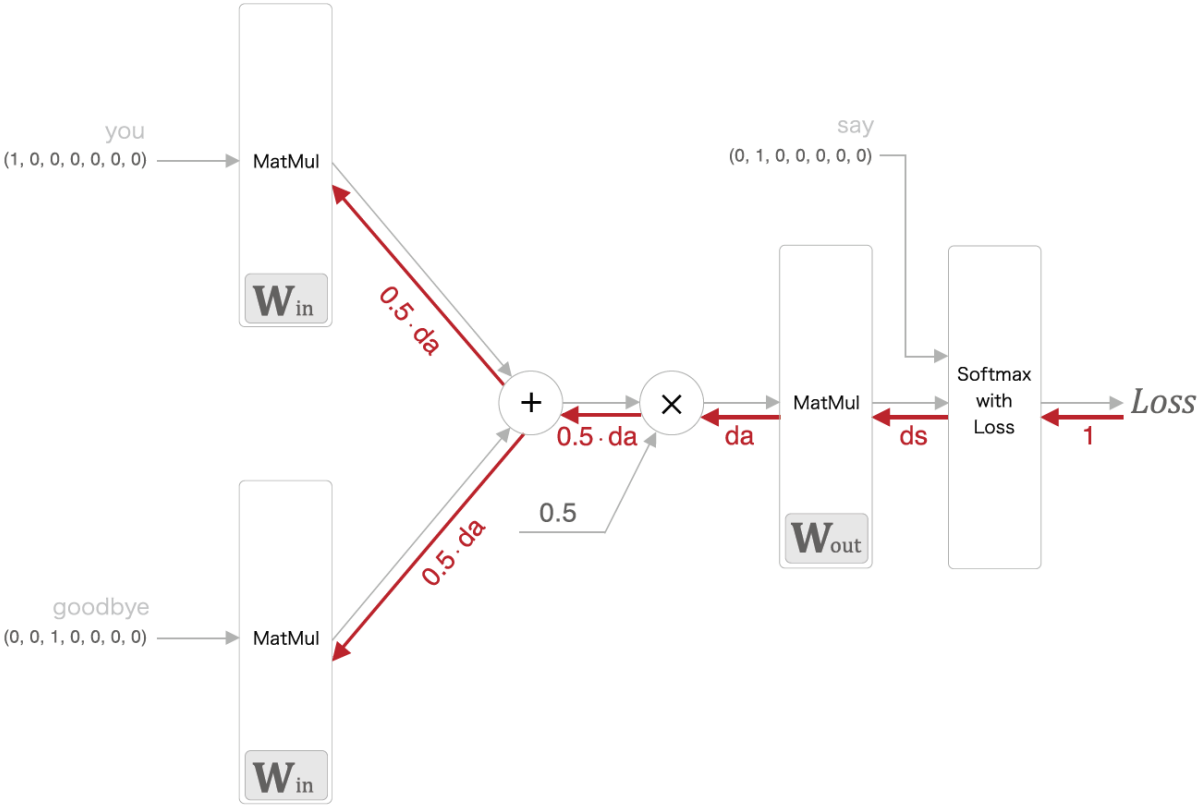
- CBOW 모델 학습가중치 조절 필요softmax 함수를 이용해 확률값으로 바꾸고 이를 정답 레이블과 함께 cross entorpy error를 거쳐 손실을 구해낸다
- CBOW 모델 가중치W_in(입력 측 완전연결계층) 가중치 & W_out(출력 측 완전연결계층) 가중치W_in 가중치: 각 행이 각 단어의 분산 표현(수평 방향)W_out 가중치: 단어의 의미가 인코딩된 벡터값 저장(수직 방향)최종적으로 사용하는 가중치?
  - 입력 층의 가중치만 사용: word2vec, 특히 skip-gram 모델에서 가장 대중적
  - 출력 층의 가중치만 사용
  - 양쪽 가중치 모두 이용: GloVe(통계 기반 + 추론 기반) 기법에서는 이 때 가장 좋은 결과

학습 데이터 준비

- 말뭉치 -> 맥락, 타깃(최종적으로 원핫 표현으로 변환)

CBOW 모델 구현

그림 3-20 CBOW 모델의 역전파(역전파의 흐름은 두꺼운(붉은) 화살표로 표시)



- 이 때 입력측 matmul 계층들은 모두 같은 가중치 공유
- X의 역전파는 순전파 시의 입력을 서로 바꿔 기울기에 곱함
- 미니배치 선택 -> 신경망 입력 -> 기울기 구함 -> 매개변수 갱신 -> 반복

word2vec 보충

- CBOW 모델과 확률P(타깃 단어 | 맥락 단어들): 맥락 단어들이 주어졌을 경우 타깃 단어가 나타날 확률말뭉치 전체에 대한 손실 값은 다음과 같이 구할 수 있다(음의 로그 가능도)(윈도우 크기가 1인 경우)

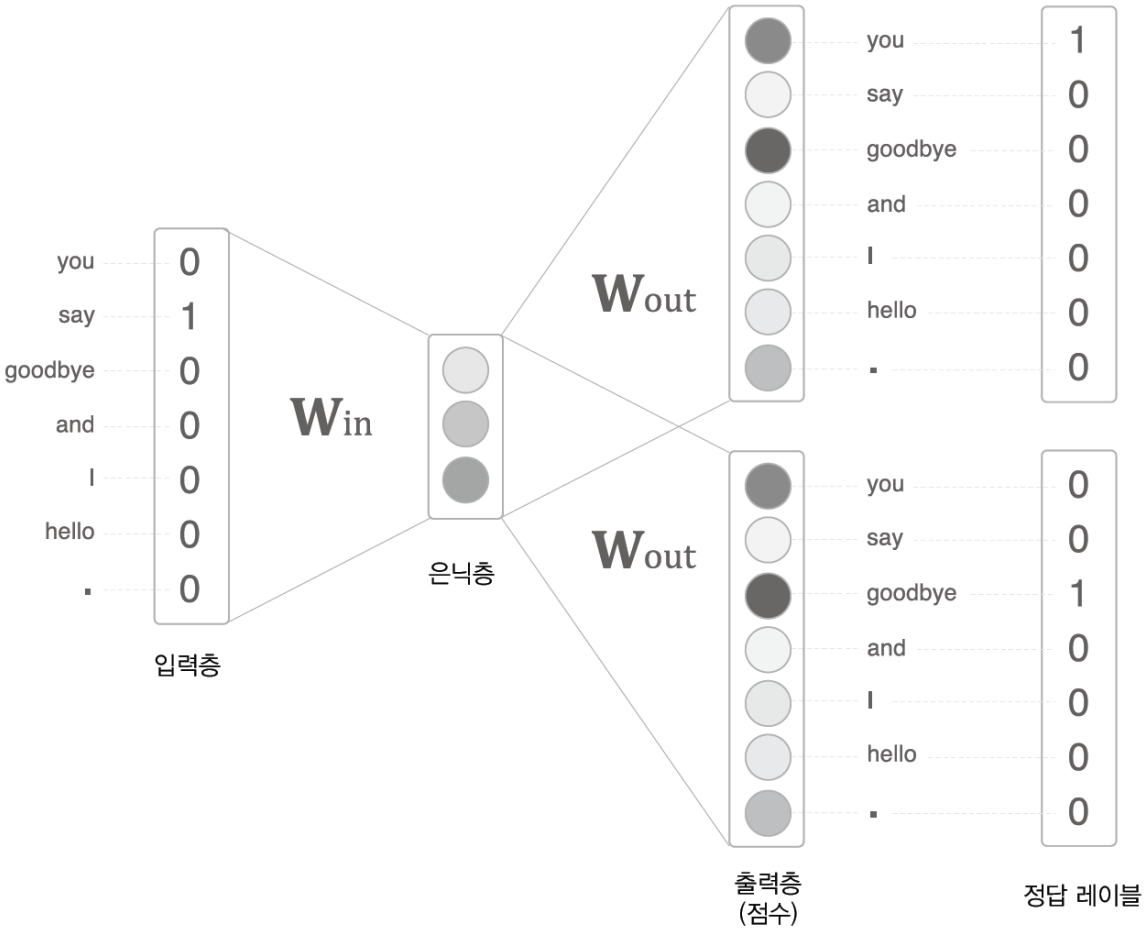
$$L = -\frac{1}{T}\sum_{t=1}^{T}\log P(w_t \mid w_{t-1}, w_{t+1})$$

- skip-gram 모델타깃 단어로부터 맥락 단어들 추론최종 손실값 = 맥락의 수만큼 있는 각 출력층의 개별 손실값들의 합> 단어 분산 표현의 정밀도 측면에서 skip-gram >

cbow 경우가 많음> 학습 속도 측면에서 cbow > skip-gram

그림 3-23 CBOW 모델과 skip-gram 모델이 다루는 문제

그림 3-24 skip-gram 모델의 신경망 구성 예



통계 기반 vs. 추론 기반

통계 기반 기법에서는 주로 단어의 유사성이 인코딩된다.

한편 word2vec, 특히 skip-gram 모델에서는 단어의 유사성은 물론, 한층 복잡한 단어 사이의 패턴까지도 파악되어 인코딩된다.

추론 기반 기법이 통계 기반 기법보다 정확하다고 흔히들 오해하곤 한다.

하지만 단어의 유사성을 정량 평가해본 결과, 추론 기반과 통계 기반 기법의 유열을 가릴 수 없었다고 한다.

추론 기반 기법과 통계 기반 기법은 서로 관련되어 있다.

word2vec 이후 추론 기반 기법과 통계 기반 기법을 융합한 GloVe 기법이 등장했다.

GloVe 의 기본 아이디어는, 말뭉치 전체의 통계 정보를 손실 함수에 도입해 미니매치 학습을 하는 것이다.