

삼성전기 AI전문가 양성과정 - 프로젝트 실습 (비영상)

자연어처리를 위한 One-Hot Encoding

현청천

2021.06.21

Discrete Value vs Continuous Value

- Discrete Value
 - Vocabulary의 일련번호, class ([긍정, 부정], [명사, 동사, 목적어, ...])
 - Finite numbers
- Continuous
 - 높이, 거리, 무게
 - Any value over a continuous range

One-Hot (감정분류)

부정	0
긍정	1

1	0
0	1
부정	긍정

문장의 긍정 부정을 예측하는 문제의 정답 표현

One-Hot (MNIST)

0	0
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9

1	0	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0
0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	0	1
0	1	2	3	4	5	6	7	8	9

0 ~ 9 사이의 숫자를 예측하는 문제의 정답 표현

One-Hot (단어)

나는	0
학생	1
입니다	2
당신은	3
수학	4
선생님	5
만나서	6
대한민국	7
...	...
만세	7999

1	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0
0	0	0	0	1	0	0	0	0
0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	1	0
...
0	0	0	0	0	0	0	0	1
0	1	2	3	4	5	6	7	... 7999

단어를 의미하는 일련번호 표현

One-Hot (단어)

나는	0
학생	1
입니다	2
당신은	3
수학	4
선생님	5
만나서	6
대한민국	7
...	...
만세	7,999

1	0	0
0	1	0
0	0	1
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
...
0	0	0

나는 학생 입니다

단어를 의미하는 일련번호 표현

One-Hot (단어)

나는	0
학생	1
입니다	2
당신은	3
수학	4
선생님	5
만나서	6
대한민국	7
...	...
만세	7,999

0	0	0	0
0	0	0	0
0	0	0	1
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	0
0	0	0	0
...
0	0	0	0

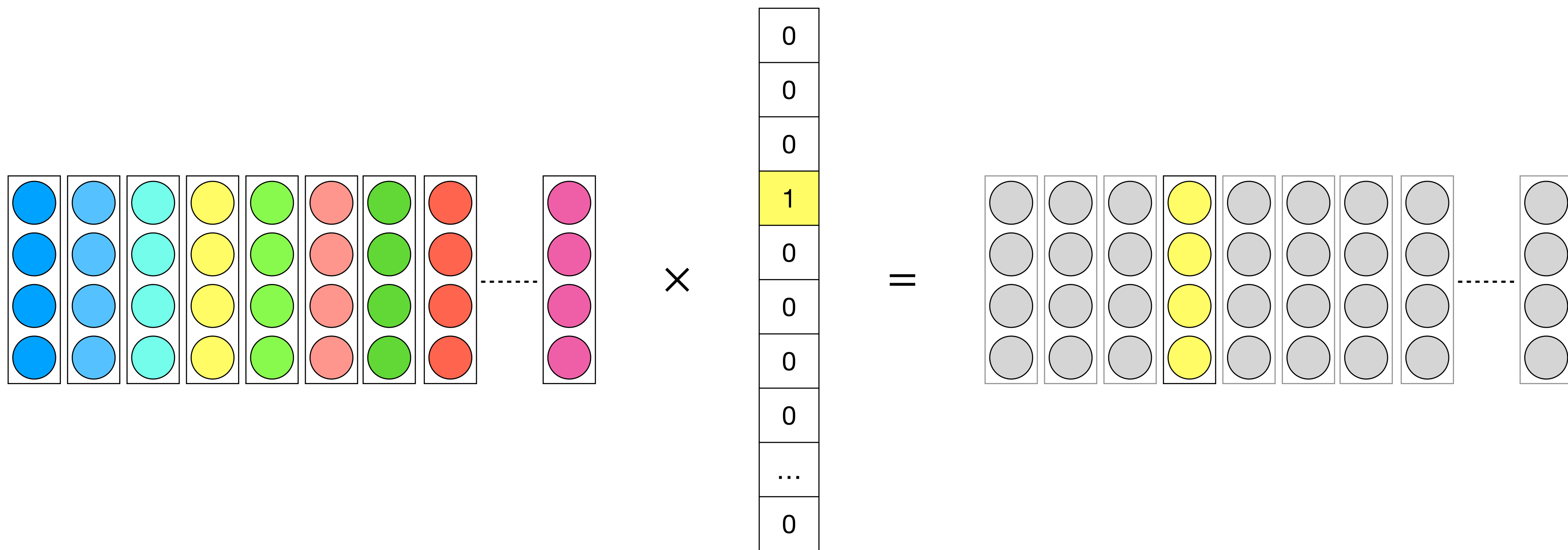
당신은 수학 선생님 입니다

단어를 의미하는 일련번호 표현

One-Hot Encoding

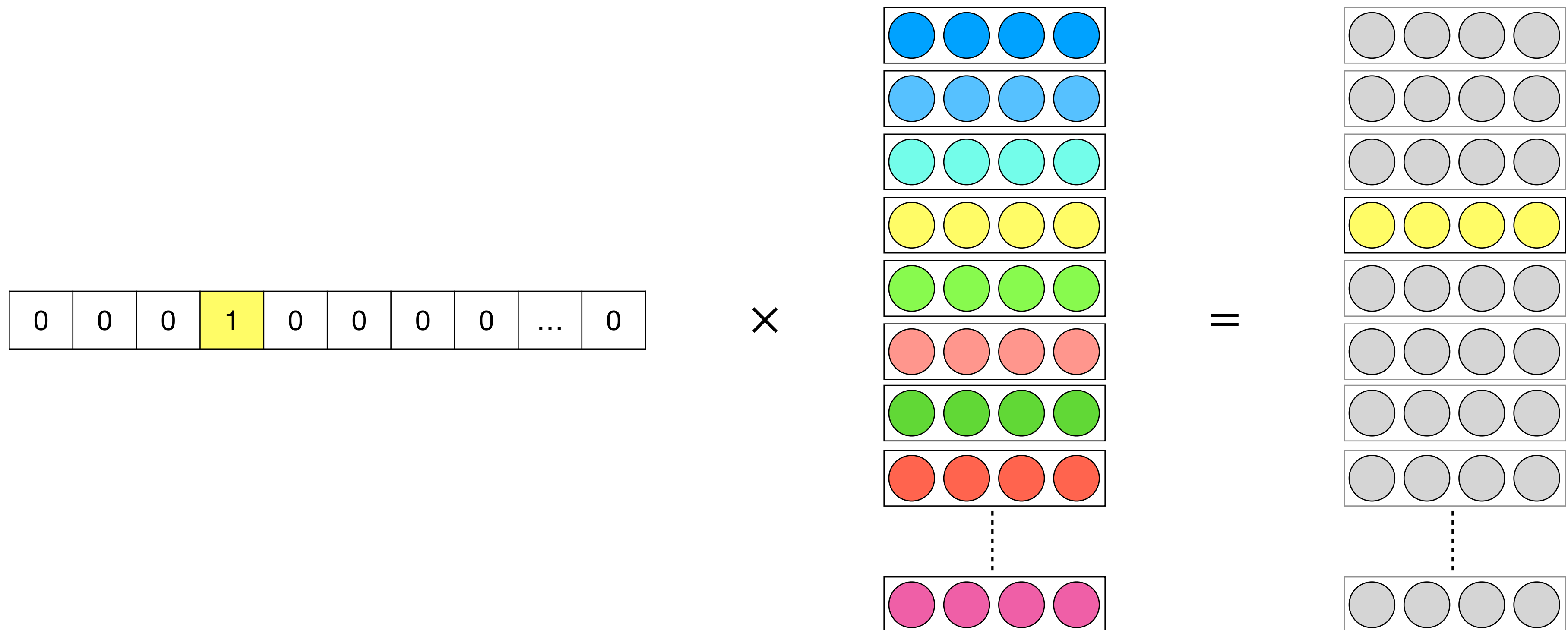
모든 클래스를 표현하는 벡터의 내적은 모두 0
하나의 값만 1 나머지는 0
매우 sparse 하고 비효율적인 벡터

One-Hot to Word Vector



$$v = Wx$$

One-Hot to Word Vector



$$v^T = x^T W^T$$

감사합니다.