1. OOV 처리에 대한 질문

AI의 basic abilities에 관하여 understanding ability를 말씀하셨는데, 기존의 학습 데이터와 다른 새로운 input data가 주어졌을 때에는 어떻게 understanding을 하게 되나요? ex) NLP model에게 신조어가 주어졌을 때

❖ 지식 표현 방법의 차이

Symbol

alphabet, **word**, sentence

This is my life now.

Number

number, **vector**, matrix

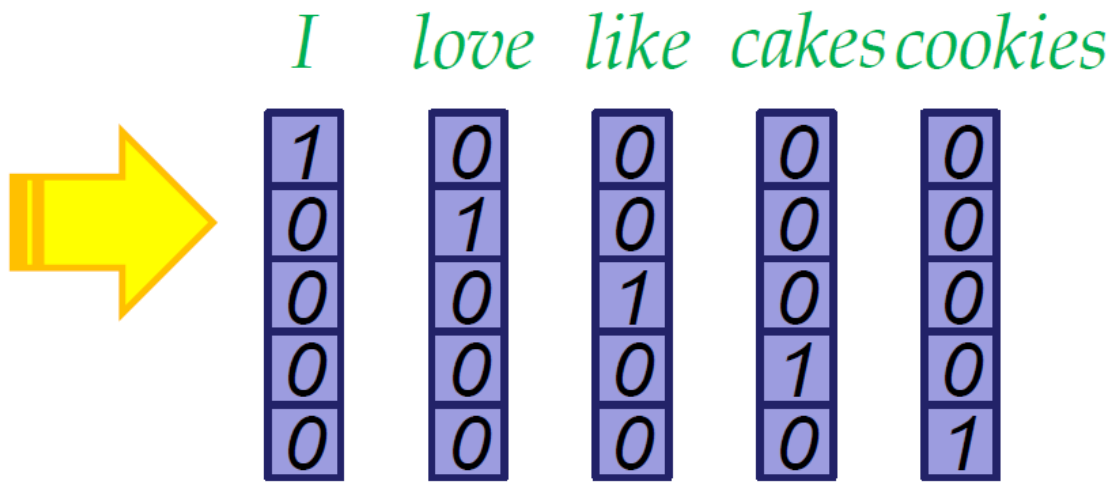
$$3 \quad \begin{bmatrix} 3 \\ 5 \\ 1 \end{bmatrix} \quad \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$$

❖ One-hot *representation in* Vector Space Model

- ***Document Examples (Bag of Words)***
  - **D1: "I love cakes."**
  - ***D2: "I like cookies."***

- **Bag-of-*words* model**

	D1	D2
I	1	1
love	1	0
like	0	1
cakes	1	0
cookies	0	1



One-Hot Representation

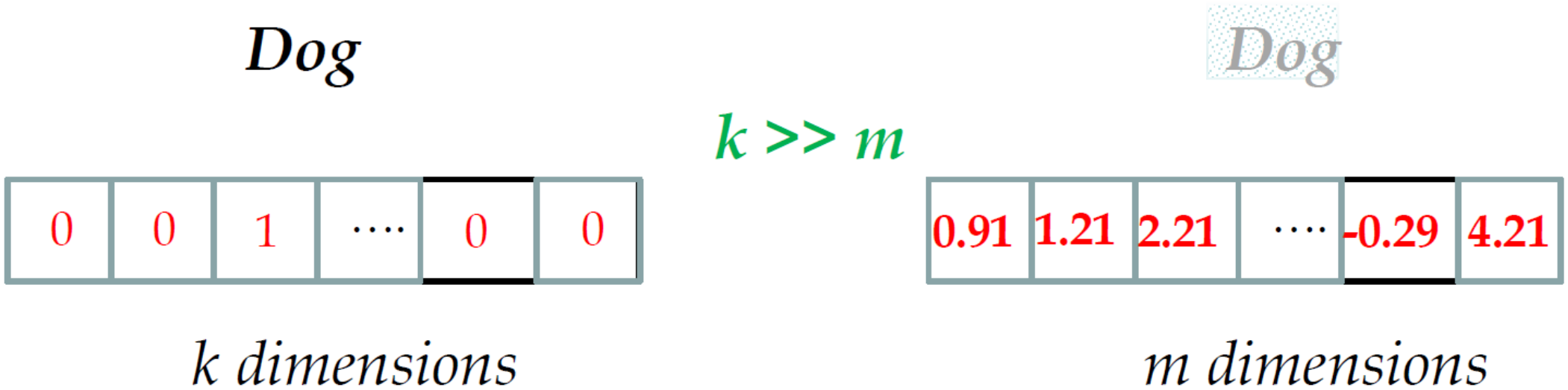
Only one neuron is active!

- Local representation
- Integer space
- Very sparse
- Very high dimensionality

Neural Word Embedding

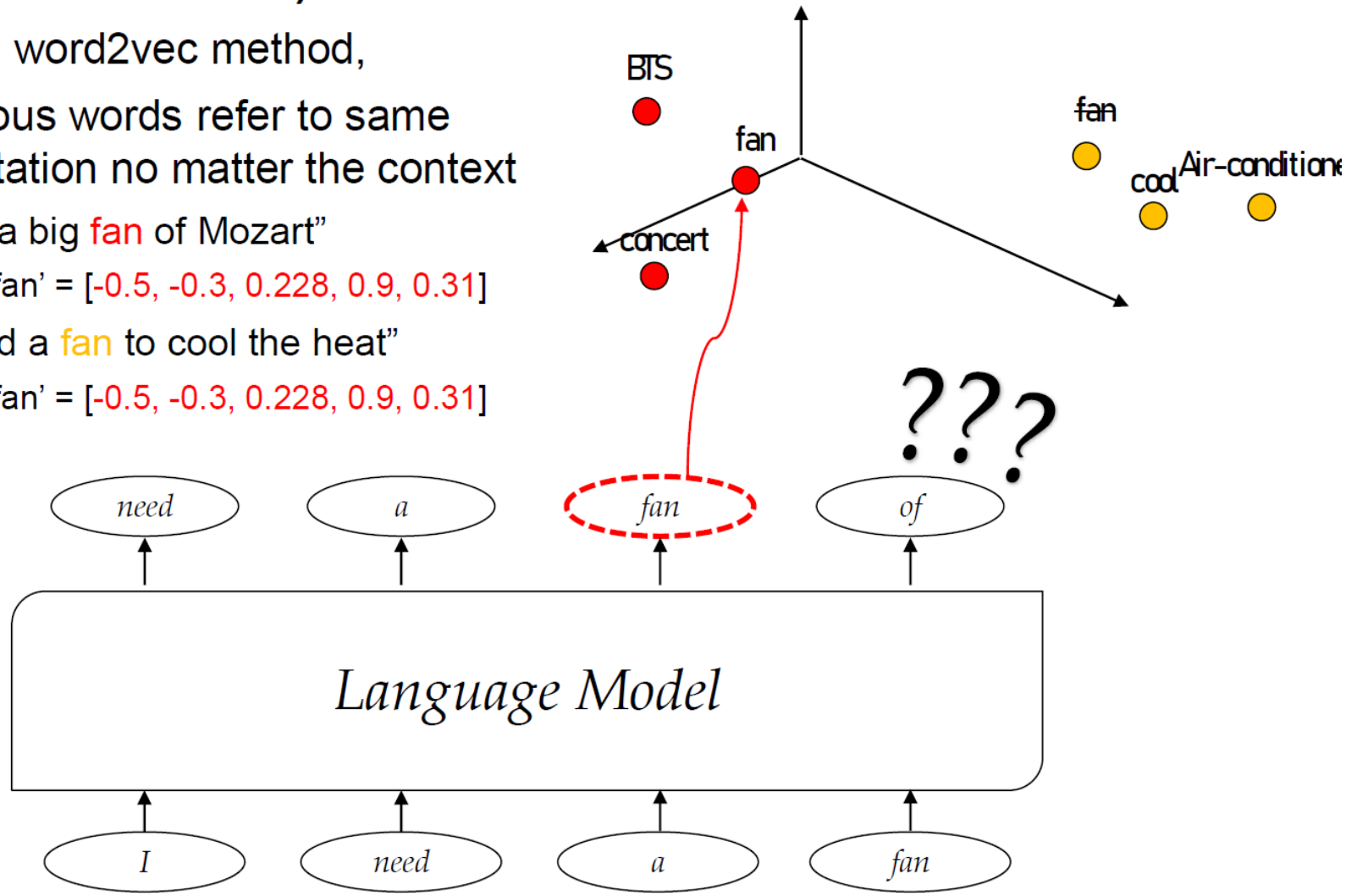
Each feature can separately be active or inactive with a different degree

- Distributed representation
- Real value space
- Dense
- Low dimensionality



❖ ELMo (Peters et al. 2018)

- In GloVe, word2vec method,
- Polysemous words refer to same representation no matter the context
  - "I am a big fan of Mozart"
    - ✓ 'fan' = [-0.5, -0.3, 0.228, 0.9, 0.31]
  - "I need a fan to cool the heat"
    - ✓ 'fan' = [-0.5, -0.3, 0.228, 0.9, 0.31]



2. 한국어 LLM 관련

AI의 기본 능력 세가지로 understanding, learning, reasoning이 있다고 말씀해주셨습니다.

그렇다면 현재 ChatGPT와 같은 여러 LLM은 상대적으로 한국어 능력이 떨어지는 모습을 보이는데 세가지 능력 중 어느 측면이 이러한 약점에 가장 큰 기여를 하고 있다고 볼 수 있을까요?

한국어 LLM 성능이 영어 LLM에 비해 낮은 이유

1. 데이터 양이 적음
2. 한국어 특징
3. 투입 가능 자원 부족