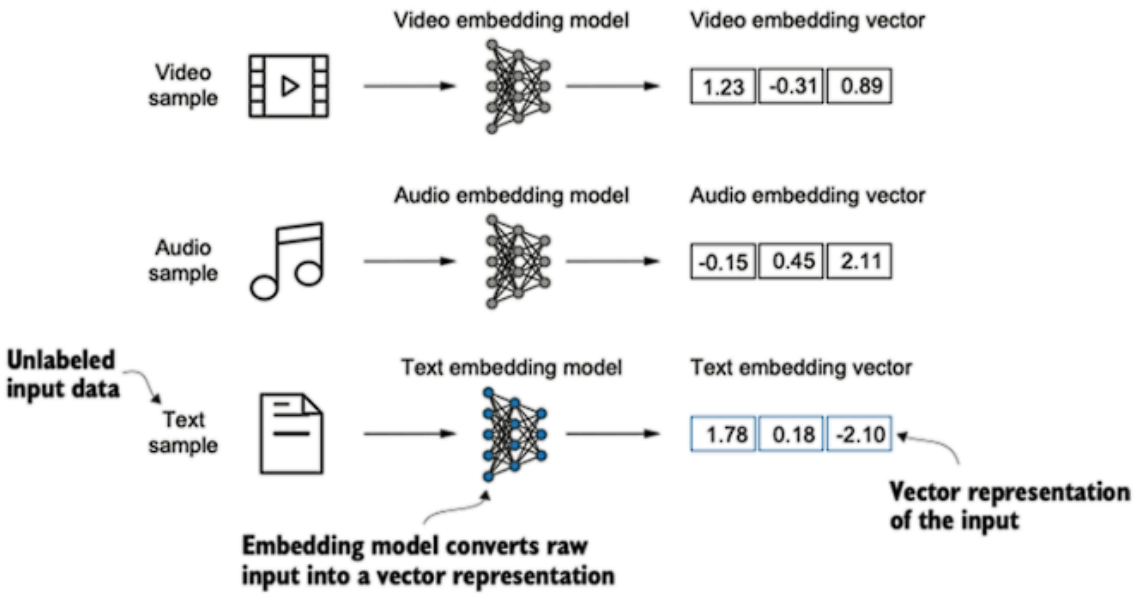
2️⃣

텍스트 데이터로 작업하기

2. 훈련 데이터셋 준비

2-1. 단어 임베딩 이해하기

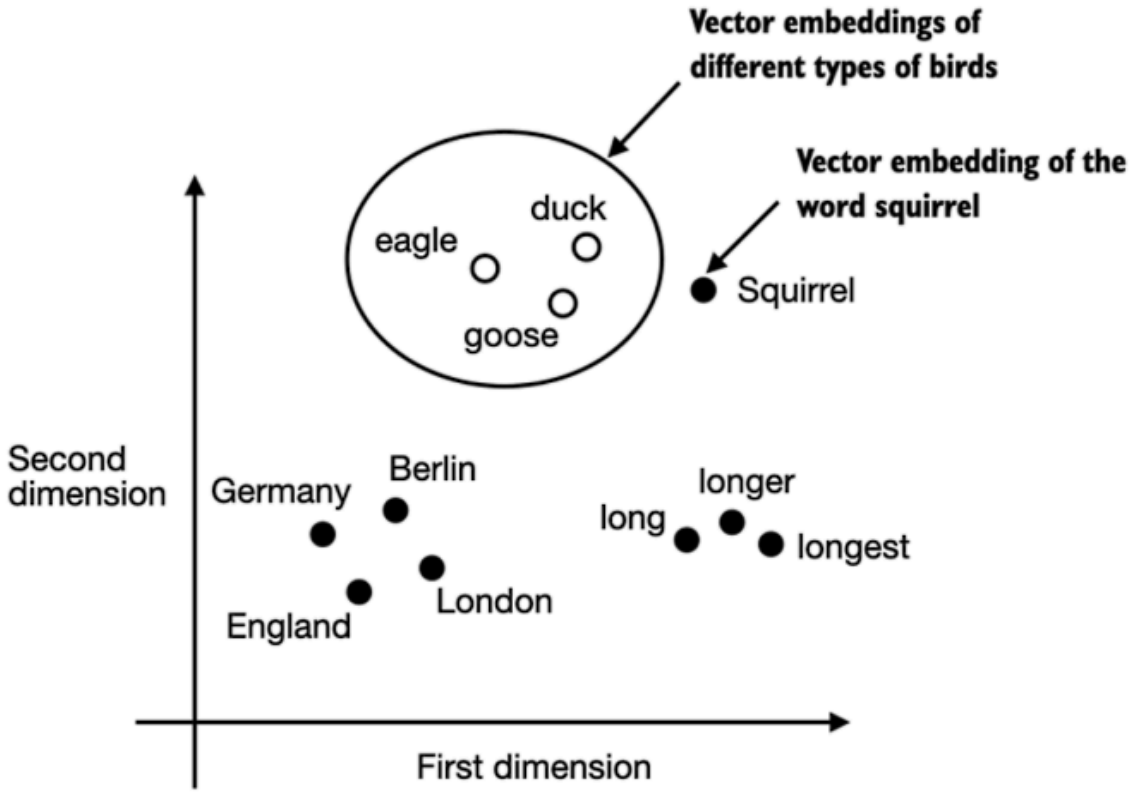
- LLM을 포함한 딥러닝 모델은 원본 텍스트를 직접 처리할 수 없다.
  - 텍스트는 범주형이기 때문에 신경망을 구현하고 훈련하는 데 사용되는 수학적 연산과 호환되지 않음
- 데이터를 벡터 형식으로 변환하는 개념을 **임베딩(embedding)**이라 한다.
  - 특정 신경망 레이어 혹은 다른 사전 훈련된 신경망 모델을 사용하여 다양한 데이터 유형을 임베딩할 수 있음



- 위 그림처럼 딥러닝 모델은 비디오, 오디오 및 텍스트와 같은 데이터 형식을 원시 형태로 처리할 수 없다.
- 따라서 임베딩 모델을 사용하여 이러한 원시 데이터를 딥러닝 아키텍처가 쉽게 이해하고 처리할 수 있는 고밀도 벡터 표현으로 변환한다.

- 즉, 임베딩의 핵심은 **단어, 이미지 또는 전체 문서와 같은 이산 객체를 연속 벡터 공간의 점으로 매핑하는 것**이다!
  - 비수치 데이터를 신경망이 처리할 수 있는 형식으로 변환하는 것이 주요 목적임
- 텍스트 임베딩의 가장 일반적인 형태는 단어 임베딩이지만, 문장, 단락 또는 전체 문서에 대한 임베딩도 존재한다.
  - 우리의 목표는 GPT 유사 LLM을 훈련하는 것이므로, 단어 임베딩에 중점을 둘 것임

- 단어 임베딩을 생성하기 위한 여러 알고리즘과 프레임워크가 개발되었지만, 초기에 인기 있었던 예로 Word2Vec 접근 방식에 대해 알아보자.
- **Word2Vec**
  - 타겟 단어를 기준으로 단어의 문맥을 예측하거나
  - 문맥을 기반으로 단어를 예측하게 하는 단어 임베딩 생성 신경망 아키텍처를 훈련함
- Word2Vec의 주요 아이디어는 **비슷한 문맥에서 나타나는 단어들이 비슷한 의미를 갖는다**는 것이다!
  - 시각화를 위해 2차원 단어 임베딩으로 투영했을 때, 유사한 용어들이 서로 가까이 클러스터링 되는 것을 볼 수 있음
  - 단어 임베딩의 차원은 1차원에서 수천 차원까지 다양할 수 있음
    - 더 높은 차원은 더 미묘한 관계를 포착할 수 있지만, 계산 효율성이 저하됨



- Word2Vec와 같은 사전 훈련된 모델을 사용하여 머신 러닝 모델을 위한 임베딩을 생성할 수 있지만
- LLM은 일반적으로 입력 계층의 일부인 자체 임베딩을 생성하고 훈련 중에 업데이트 한다.
    - Word2Vec를 사용하는 대신 LLM 훈련의 일부로 임베딩을 최적화하는 이점은 임베딩이 특정 작업과 데이터에 맞게 최적화된다는 것이다

2.2 텍스트 토큰화

- LLM을 위한 임베딩을 생성하는 데 필수적인 전처리 단계인 **토큰화**에 대해 알아보자.
    - 토큰화는 입력 텍스트를 개별 토큰으로 분할하는 것이다.

- 이후 내용은 주피터 노트북 파일로...