CAT2VEC

LEARNING DISTRIBUTED REPRESENTATION OF MULTI-FIELD CATEGORICAL DATA

REFERENCE

- Wen, Ying, et al.
  "Cat2Vec: Learning Distributed Representation of Multi-field Categorical Data." (2016).

ABSTRACT

ABSTRACT

- 멀티-필드 카테고리컬 데이터의 분포된 표현
- 카테고리 필드 간 상호작용 탐구가 중요, 필요
  - Inter-field

ABSTRACT

- 저자들은 NLP 씬의 Word2Vec에서 영감을 받아
  - Cat2Vec을 제안
  - Categories to vectors
- Cat2Vec
  - 저차원 연속 벡터가 각 카테고리에 대해 학습

ABSTRACT

- 신경 게이트(neural gate)
  - Inter-field 카테고리 간 상호작용 탐색
- 풀링 레이어(Pooling layer)
  - 정보가 많은 조합을 선택

ABSTRACT

- 실험에서
  - Cat2Vec을 적용해 큰 성능 향상을 얻음

INTRODUCTION

INTRODUCTION

- 데이터는 서로 다른 추상 레벨로 구성
  - 저-추상: 이미지, 비디오, 오디오 등
  - 고-추상: 자연어, 로그 데이터 등
- 고-추상 데이터들은
  - 통상 이산적이고
  - 아토믹 심볼을 포함함

INTRODUCTION

- 고-추상 데이터를 다루기 위해
  - 잘 알려진 기법은 임베딩(enbedding)
  - 이산적 토큰을 저-차원 연속 공간으로 임베딩
    - 신경망의 입력으로 넣을 수 있도록
  - 그 후 신경망이 잠재 패턴을 학습하도록 진행

INTRODUCTION

- 멀티-필드 카테고리컬 데이터
  - 고-추상 데이터의 한 부류
  - 각 필드의 카테고리들이 이종적(hetero-)
- 이러한 데이터는
  - 추천 시스템, 소셜 링크 예측, 계산적 광고 등
  - 트랜잭션 로그에서 잘 발생

INTRODUCTION

- 멀티-필드 카테고리컬 데이터의 예
  - iPinYou 데이터셋:

TARGET	GENDER	WEEKDAY	CITY	BROWSER
1	MALE	TUESDAY	BEIJING	CHROME
0	FEMALE	MONDAY	SHANGHAI	IE
1	FEMALE	TUESDAY	HONGKONG	IE
0	MALE	TUESDAY	BEIJING	CHROME
NUMBER OF CATEGORY	2	7	351	6

INTRODUCTION

- 인터-필드(inter-field) 카테고리들의
  - 의존성에 대한 명시가 없음
  - 두 솔루션이 주로 사용됨

INTRODUCTION

- 1. 필드 간 특징들을 사람이 직접 조합하기
  - 비용이 높음
  - 특징/파라미터 공간이 커지면 불가능
- 2. 함수 또는 신경망으로 특징 임베딩하기
  - aimless하게 브루트포스로 찾아
  - 효율이 떨어짐

INTRODUCTION

- 본 페이퍼에서 제안하는 방식
  - 비지도
  - 멀티-필드 카테고리컬 데이터의 분포된 표현을 학습

INTRODUCTION

- 신경 게이트로부터 탐색
- K-max 풀링 레이어로 선택
  - 전통적인 아프리오리(Apriori) 알고리즘과 유사하게 동작

INTRODUCTION

- 효율적 학습을 위해
  - 카테고리 벡터를 추정하기 위한
  - 판별기(discriminant) 학습 방법을 제안

RELATED WORK

RELATED WORK

- One-hot 표현

$$\underbrace{[0,1]}_{\text{GENDER:MALE}},\underbrace{[0,1,0,0,0,0,0]}_{\text{WEEKDAY:TUESDAY}},\underbrace{[0,\ldots,0,1,0,\ldots,0]_{351}}_{\text{CITY:BEIJING}},\underbrace{[1,0,0,0,0,0]}_{\text{BROWSER:CHROME}}$$

- 문제
  - 차원의 저주
  - 관계성을 담지 못함 (필드 간, 카테고리 간)

RELATED WORK

- 분포된 표현 (Distributed representation)
  - d-차원 벡터로 매핑
  - 일반적으로 d는 하이퍼파라미터
    - 원래 카테고리 수 보다 (현저히) 작은 값
- 의미적 유사성이 거리로써 표현됨
  - 코사인 유사도
  - 유클리디언 거리 등

RELATED WORK

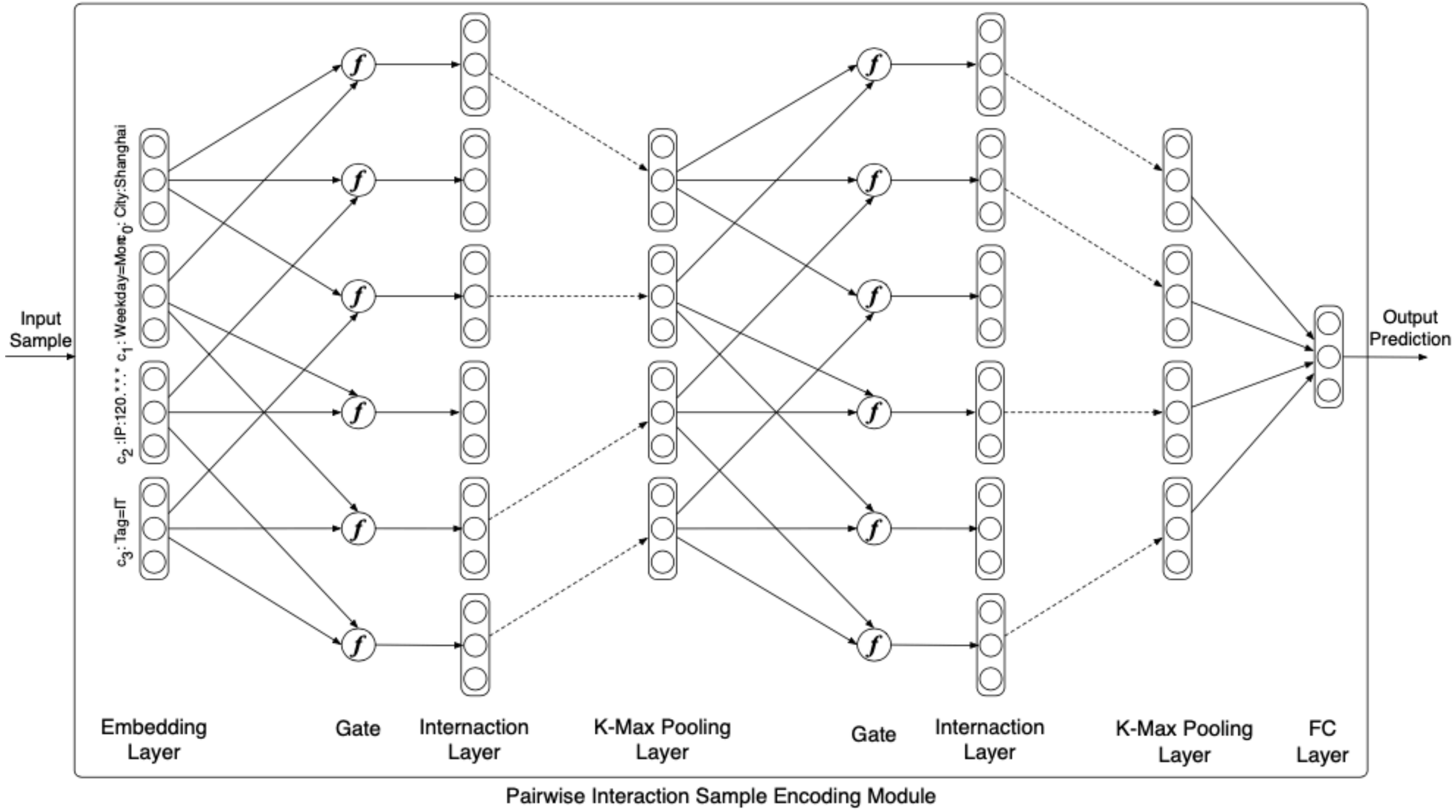
- 분포된 표현 (Distributed representation)
- Word2Vec
  - 단어를 벡터로 표현
  - 의미적 유사도 판단 가능
  - (+) 연산이 가능

CAT2VEC

CAT2VEC

- 신경 게이트
  - 카테고리 쌍의 상호 작용을 파악하기 위함
- K-max 풀링 레이어
  - 가장 중요한 상호 작용을 선별하기 위함
- 위 두 구조를 반복
  - 높은 레벨의 상호 작용을 탐색하기 위함

CAT2VEC


Input Sample
c_0: City:Shanghai
c_1: Weekday=Mon
c_2: IP:120.*.*.*
c_3: Tag=IT
Embedding Layer
Gate
Internaction Layer
K-Max Pooling Layer
Gate
Internaction Layer
K-Max Pooling Layer
FC Layer
Output Prediction
Pairwise Interaction Sample Encoding Module

CAT2VEC: INTERACTION LAYER

- 상호 작용 레이어
  - 카테고리 쌍의 상호 작용을 평가하기 위해 게이트 사용
  - 게이트는 상호 작용 결과를 반환

CAT2VEC: INTERACTION LAYER

- 수학적으로, 게이트는 함수
  - $f : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}^d$
  - 카테고리 벡터 c_i, c_j를 입력으로 벡터 $c'_{i,j}$를 반환
    - $c'_{i,j} = f(c_i, c_j)$

CAT2VEC: INTERACTION LAYER

- 게이트 f 는 다양하게 구성할 수 있음
  - $f^{\text{sum}}(\boldsymbol{c}_i, \boldsymbol{c}_j) = \boldsymbol{c}_i + \boldsymbol{c}_j,$

    $f^{\text{mul}}(\boldsymbol{c}_i, \boldsymbol{c}_j) = \boldsymbol{c}_i \odot \boldsymbol{c}_j,$
- 더 복잡하게는:
  - $f^{\text{highway}}(\boldsymbol{c}_i, \boldsymbol{c}_j) = \boldsymbol{\tau} \odot g(\mathbf{W}_H(\boldsymbol{c}_i + \boldsymbol{c}_j) + \boldsymbol{b}_H) + (1 - \boldsymbol{\tau}) \odot (\boldsymbol{c}_i + \boldsymbol{c}_j),$
  - g는 비선형 함수
  - $\tau = \sigma(W_\tau(c_i + c_j) + b_\tau)$ 인 transform gate

CAT2VEC: K-MAX POOLING LAYER

- K 개의 최대 상호 작용 아웃풋 벡터 $c'_{i,j}$를 선택
  - K는 학습 샘플의 원래 카테고리 수

CAT2VEC

- 상효 작용 및 K-max 풀링 연산을 반복
  - 수 차례 반복
  - 다른 필드 간 높은 레벨의 상호 작용을 포착하기 위함
- 이후 최종 상호 작용 벡터 표현을 예측
  - 전 연결 레이어를 통해

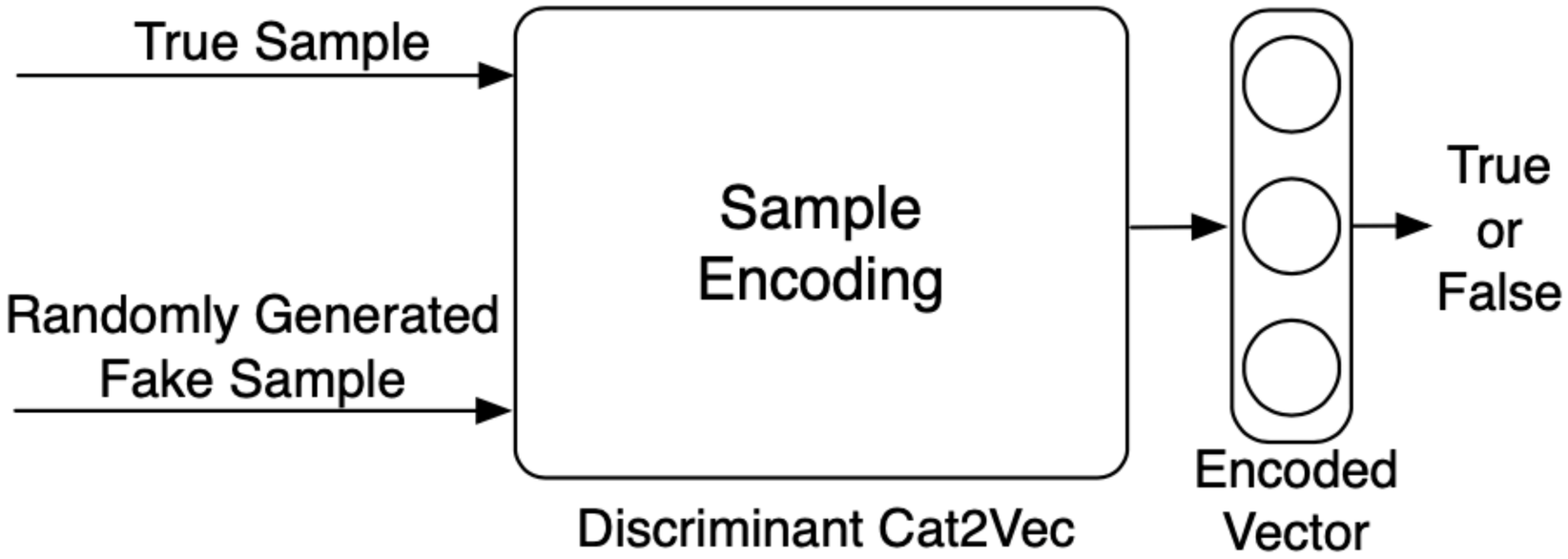
DISCRIMINANT
CAT2VEC

DISCRIMINANT CAT2VEC

- Cat2Vec의 학습 방법
  - 데이터의 비지도 학습
  - 모델의 지도 학습

DISCRIMINANT CAT2VEC

- 판별기 Cat2Vec 모델



DISCRIMINANT CAT2VEC

- 샘플 인코딩 모듈에
  - 참 또는 거짓 샘플을 입력
- 인코딩된 샘플 벡터는
  - MLP를 통해 참일 확률 p를 예측

DISCRIMINANT CAT2VEC

- 거짓 샘플이 카테고리 벡터 학습에 영향을 끼침
- 거짓 샘플의 제작 과정
  - 학습 셋에서 무작위로 샘플을 하나 선택
  - 무작위로 여러 카테고리들을 선택
  - 같은 필드에 속하도록 무작위로 교체

DISCRIMINANT CAT2VEC

- 판별기 네트워크는 새 샘플이 참인지 아닌지를 학습
- 판별기 네트워크 손실 함수
  - 평균 크로스 엔트로피
  - 올바른 예측의 우도를 최대화

DISCRIMINANT CAT2VEC

- 판별기 네트워크 손실 함수
  - $$L = \frac{1}{M}\sum_{i=1}^{M} -y_i \log(p_i) - (1-y_i)\log(1-p_i)$$
  - M : 샘플의 수
  - i번째 샘플의 레이블 $y_i \in \{1,0\}$: {참, 거짓}
  - p_i : 참일 확률의 예측값

EXPERIMENT

CLICK-THROUGH RATE PREDICTION

- 클릭률 (Click-through rate, CTR)
  - 광고를 본 사용자가 해당 광고를 클릭하는 빈도의 비율
- iPinYou 데이터셋
  - 23개의 필드
  - 이 중 카테고리가 10개 이상인 18개 필드를 선택해 사용

CLICK-THROUGH RATE PREDICTION

- CTR 추정을 위한 기존 방법
  - 원-핫 데이터 표현에 기반한 LR (Logistic Regression)
  - Factorisation-Machine Supported Neural Networks (FNN)
  - Convolutional Click Prediction Model (CCPM)
- 이 방법들은 예측 성능 향상에만 힘씀
  - 멀티-필드 카테고리컬 데이터의 표현 학습이나
  - 더 나은 표현에 대해 고심하지 않음

CLICK-THROUGH RATE PREDICTION

- Cat2Vec-FNN-1
  - K-max 풀링 결과를 concatenate 해서 최종 벡터 표현을 형성, 예측
- Cat2Vec-FNN-2
  - K-max 풀링 결과와 카테고리 임베딩으로 최종 벡터 표현을 형성, 예측
  - 카테고리 임베딩의 영향력이 조금 더 있을 것으로 예상

CLICK-THROUGH RATE PREDICTION

- 실험 결과:

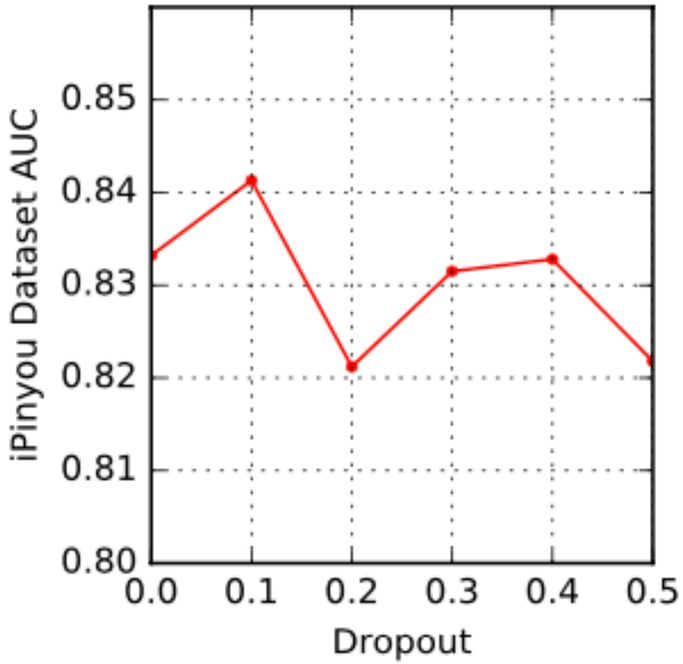
Table 3: AUC of CTR prediction on iPinYou dataset.

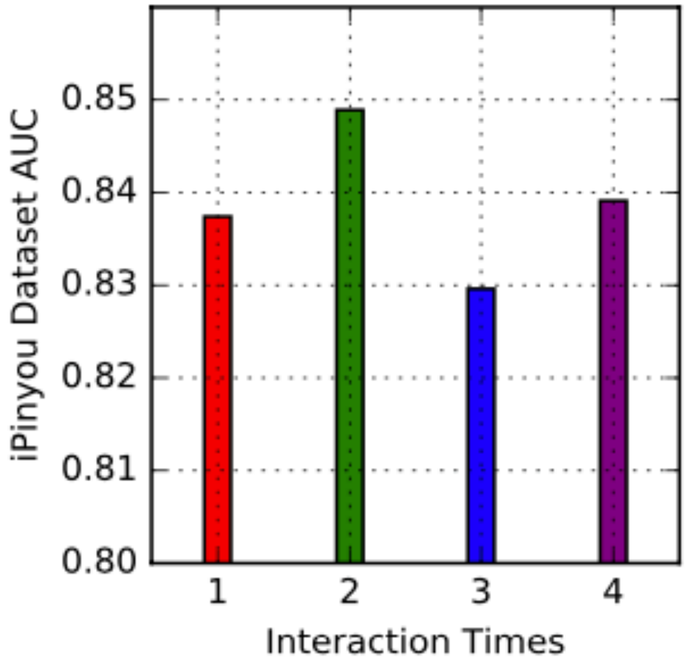
MODEL	LR	FM	CCPM	FNN	Cat2Vec-FNN-1	Cat2Vec-FNN-2
AUC	0.8323	0.8349	0.8364	0.8453	0.8599	**0.8640**

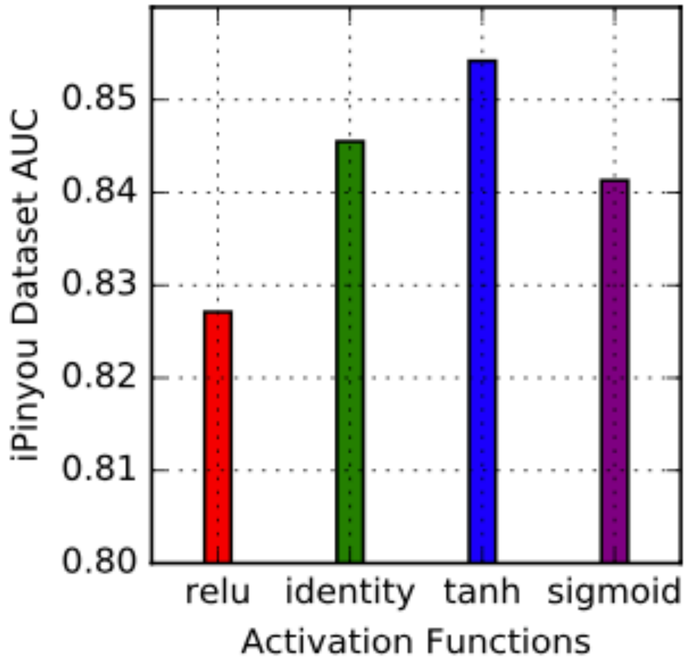
  - 다른 기존의 방법들을 상회

CLICK-THROUGH RATE PREDICTION

- 하이퍼 파라미터의 변경:







- 드롭아웃 0.1, interaction 2, tanh에서 제일 좋은 성능을 보임

CONCLUSION

CONCLUSION

- Cat2Vec
  - 멀티-필드 카테고리컬 데이터를 효과적으로 임베딩
- 다른 방법들과는 달리, Cat2Vec은
  - 높은 레벨의 상호 작용을 탐색할 수 있도록 함

CONCLUSION

- 추후 더 복잡하고 정교한 게이트를 설계해
  - 인터-필드 카테고리의 상이한 상호 작용 패턴에
  - 관련된 탐색도 진행할 예정

CAT2VEC

LEARNING DISTRIBUTED REPRESENTATION OF MULTI-FIELD CATEGORICAL DATA