Samsung Data Challenge 2018

교통사망사고정보 Data Completion

Timmy
YeongTaek Oh

Samsung Data Challenge 2018

Index

- Problem & Data Overview
- Data Split & Augmentation
- Model Structure
- Result

Problem & Data Overview

Problem Overview (Test Data 기준)

- **Imputation Missing Values** based on other values
- Missing Variables can be any variable in test data (Figure 1)
- 〉Self Data Completion

How to Solve?

- AutoEncoder with Missing Value
  - ➢ Baseline on Multiple Imputation using Denoising AutoEncoder (MIDA) (Figure 2)

주야	요일	사망자수	사상자수	중상자수	경상자수	부상신고자수	발생지시도	발생지시군구	사고유형_대분류	사고유형_중분류	법규위반	도로형태_대분류	도로형태	당사자종별_1당_대분류	당사자종별_2당_대분류
야간	금			0		0	경기	화성시	차대차	측면충돌	중앙선침범	단일로	기타단일로	승용차	승합차
야간	화	1	1	0	0	0	대구	북구				단일로	기타단일로	승용차	화물차
주간		1	1	0	0	0	서울	동작구			신호위반	교차로	교차로내	화물차	원동기장치자전거
	일			0	0	0	전북	고창군	차대사람	기타	부당한회전	교차로	교차로내		보행자

Figure 1. Sample of Test Data

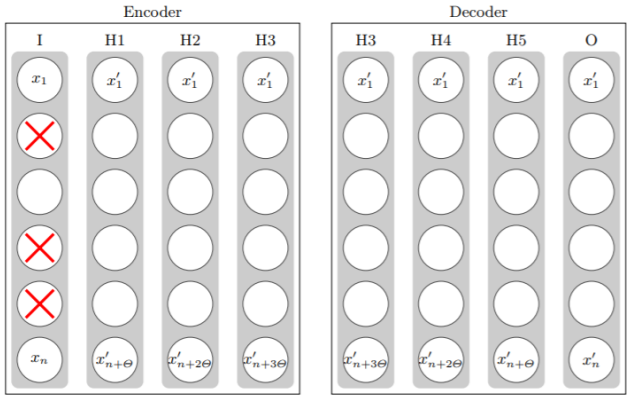


Figure 2. MIDA

Problem & Data Overview

- # of data : Train 25037, Test : ?
  - Train Data : 12.01 ~ 17.06 기간의 교통사망사고정보 데이터
  - Test Data : 17.07 이후의 교통사망사고정보 데이터 (Not provided)

- # of Variables : Train(27) ⊃ Test(16)
  - Categorical Variables : 11
  - Numerical Variables : 5

- Metric for Variables
  - Categorical Variable : $C \times \sum_{i=1}^{k_2} \delta_{c_i d_i}$
  - Categorical Loss : Categorical Cross Entropy
  - Numerical Variable : $B \times \sum_{i=1}^{k_1} \exp\left\{-\left(\frac{n_i - m_i}{s_j}\right)^2\right\}$
  - Numerical Loss : Mean Squared Error

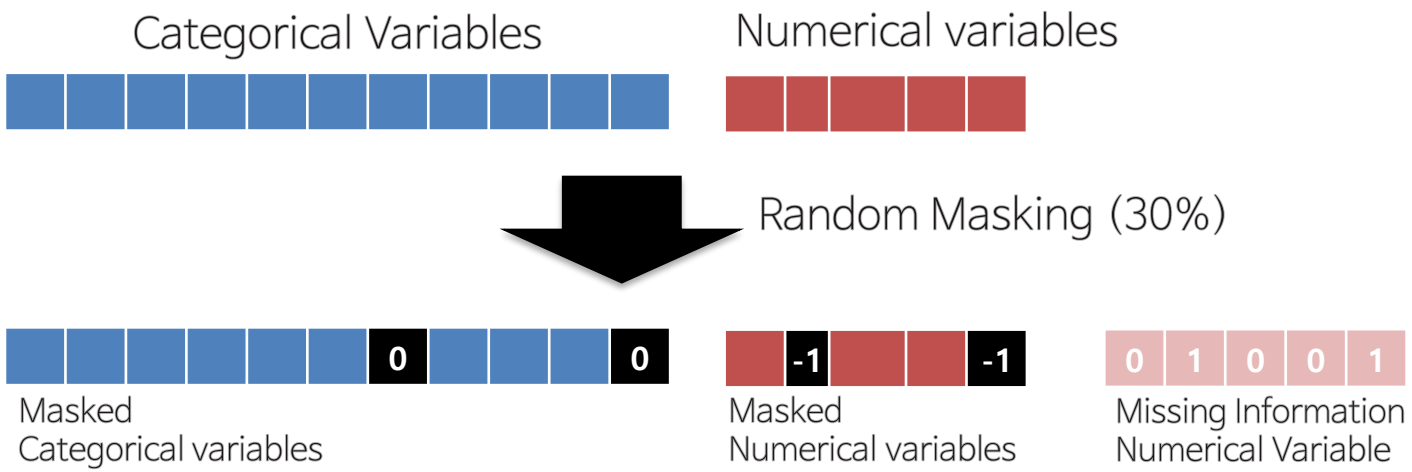
발생년	사고유형
발생년월일시	법규위반_대분류
발생분	법규위반
주야	도로형태_대분류
요일	도로형태
사망자수	당사자종별_1당_대분류
사상자수	당사자종별_1당
중상자수	당사자종별_2당_대분류
경상자수	당사자종별_2당
부상신고자수	발생위치X_UTMK
발생지시도	발생위치Y_UTMK
발생지시군구	경도
사고유형_대분류	위도
사고유형_중분류	

Table 1. Variables in data

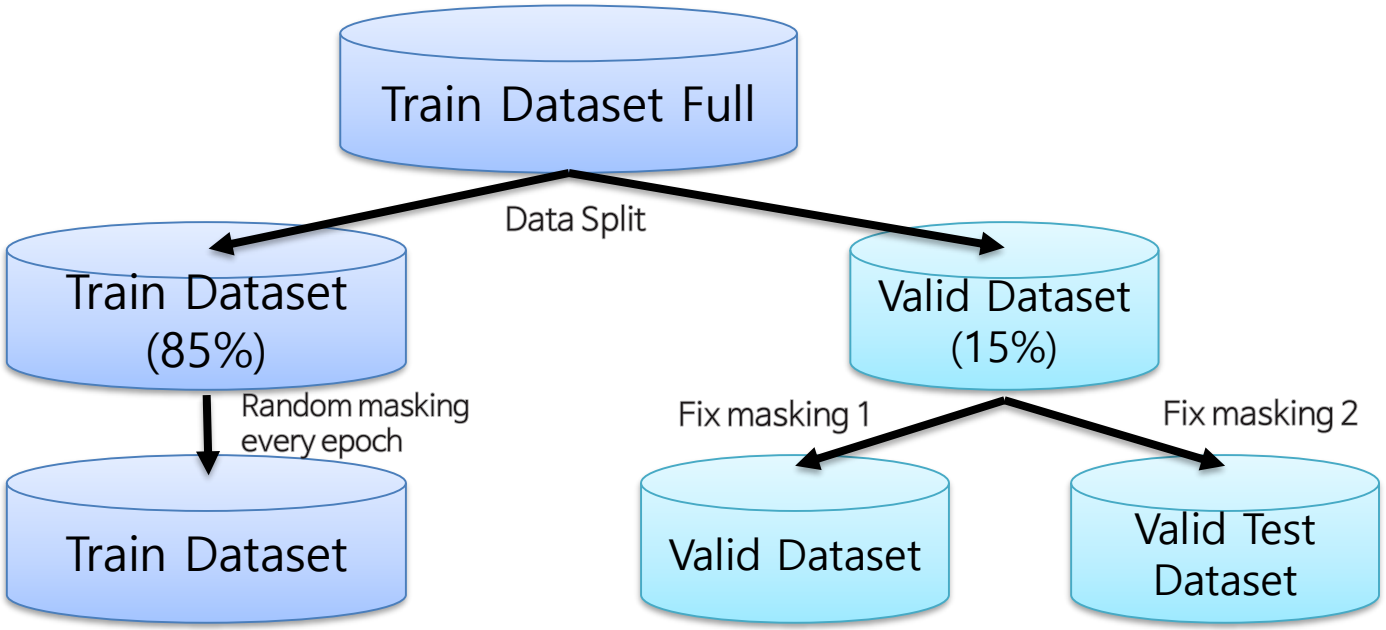
Data Split & Augmentation

- Training Set / Validation Set (85% / 15%)
  1. 15% Random Sampling from Training Data (Time Shuffled) * 2
  2. 15% Latest Data in Training Data (Time Not Shuffled)

- Masking missing values
  - Categorical variables : mask '0' for missing categorical value
  - Numerical variables : mask '-1' value on missing and create missing table (n x 5)



Data Split & Augmentation



- Random masking on Training Data on every step
  - gives **data augmentation effects**

- Fixed masking on Validation Data **for early stopping**
- Different Fixed masking on Validation Data **for parameter tuning (self-test)**
  - **to prevent overfitting on validation set**

Model Structure

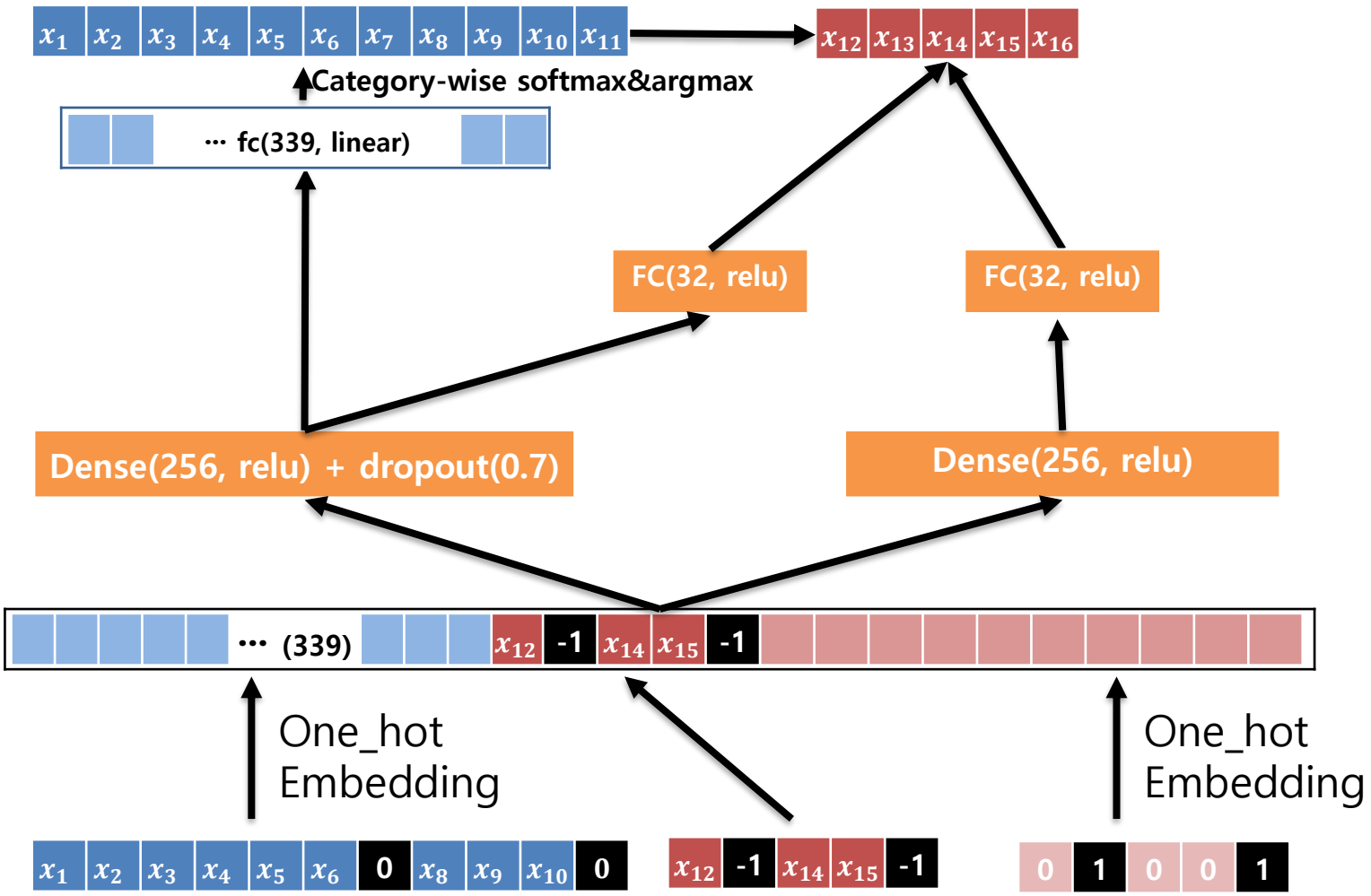
Projection Layer

Hidden Layer2

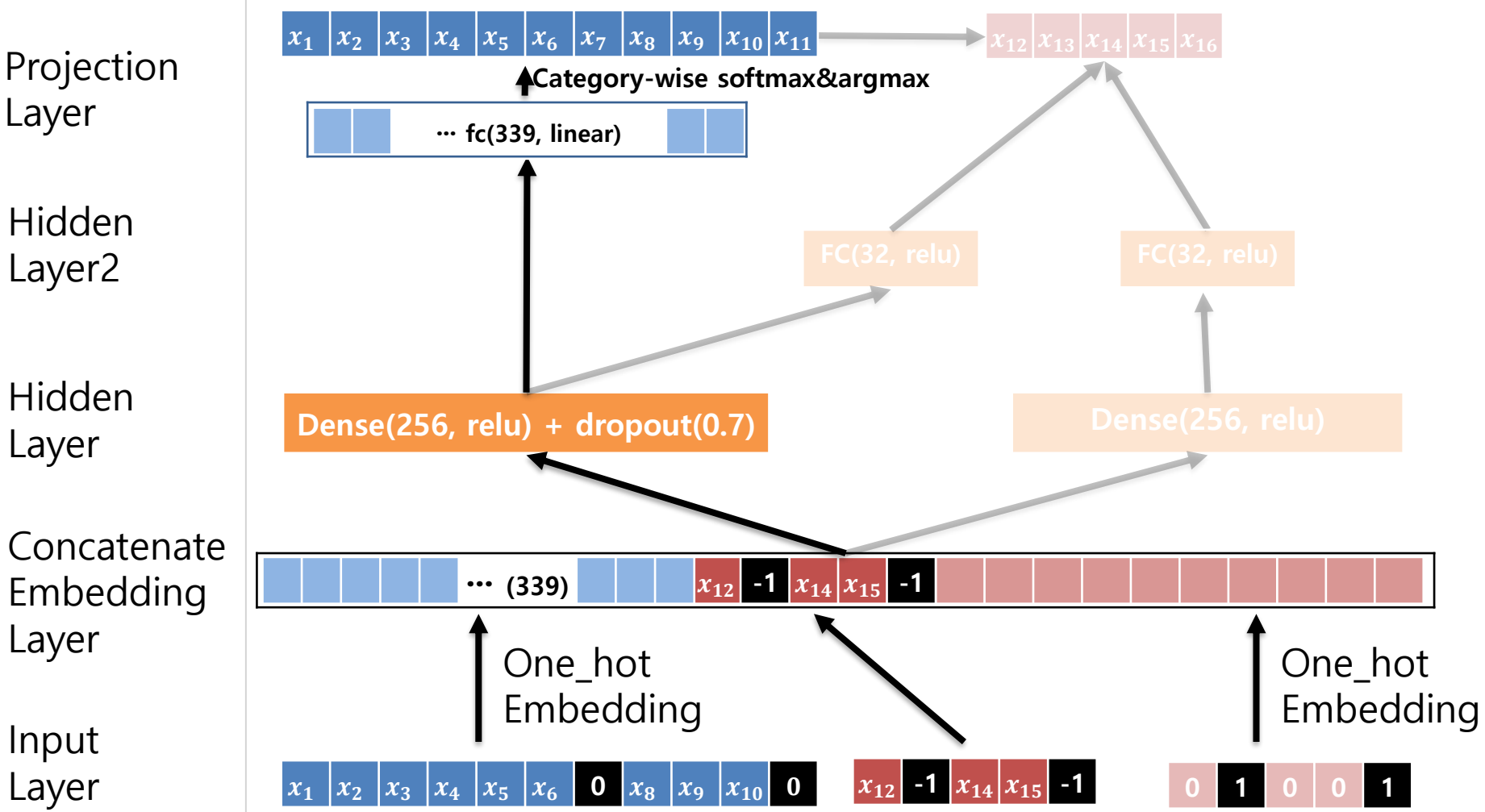
Hidden Layer

Concatenate Embedding Layer

Input Layer

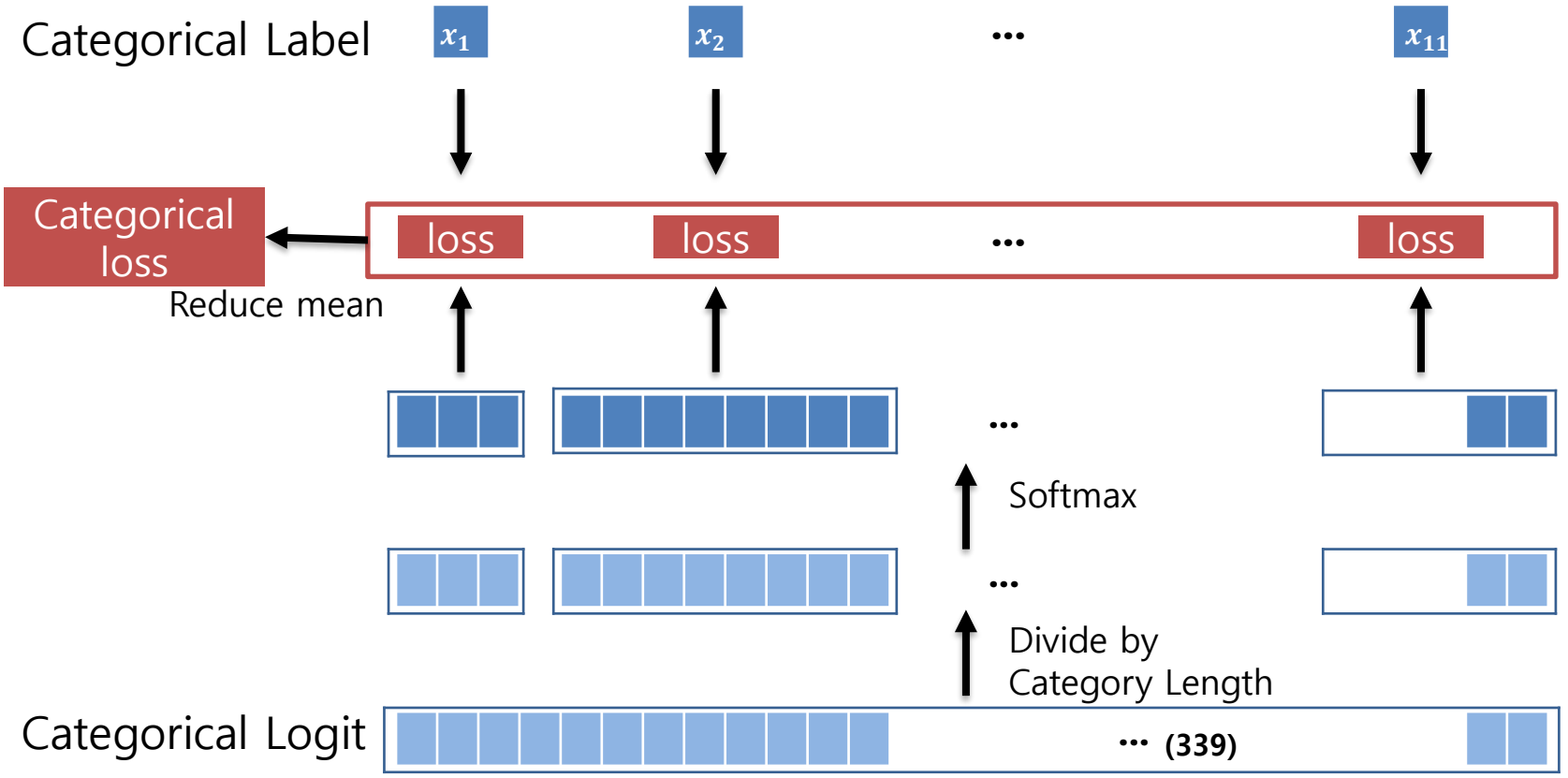


Model Structure (Categorical Label)

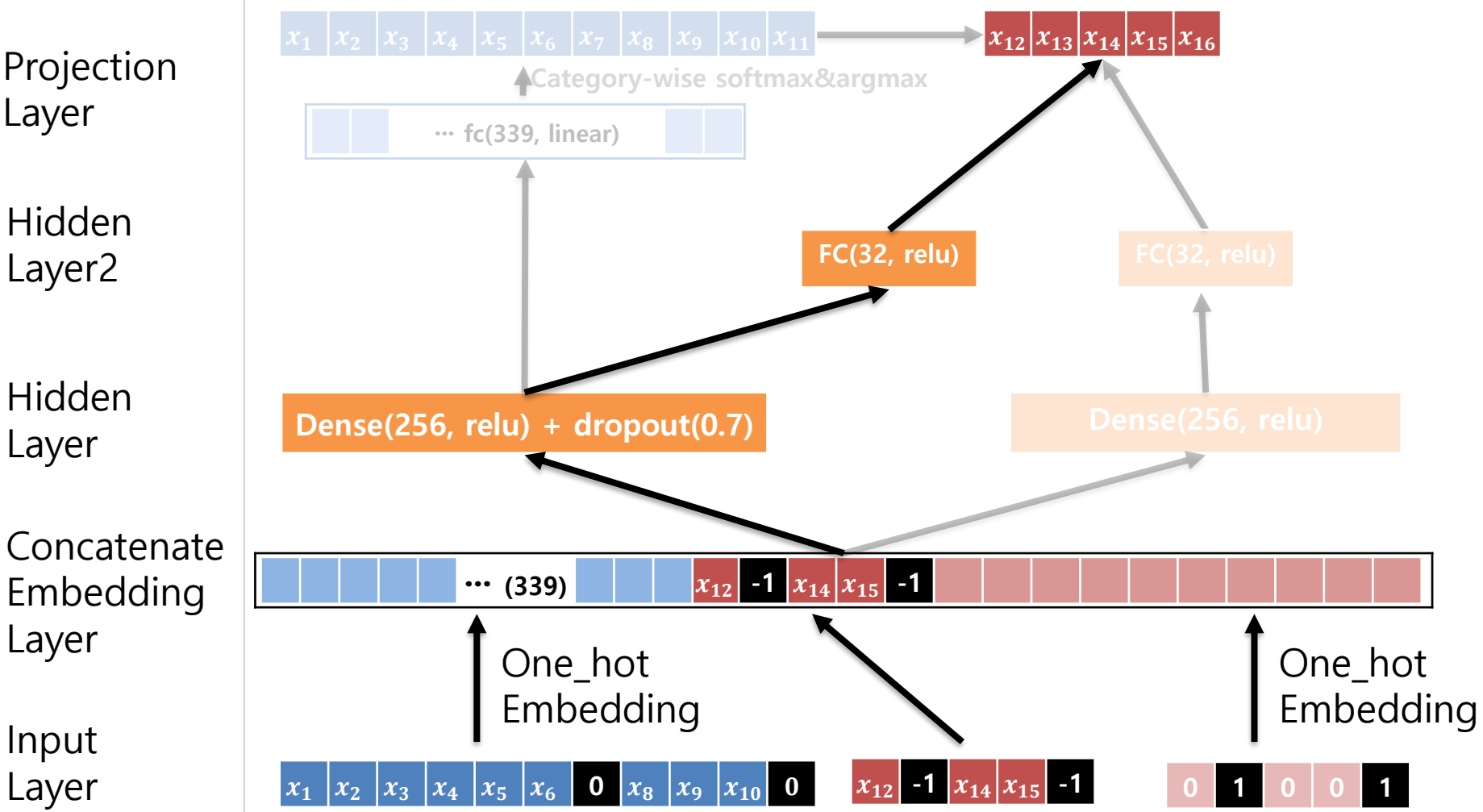


Model Structure (Categorical Label)

Category-wise softmax&argmax, Categorical loss



Model Structure (Numerical Label Path1)



Model Structure (Numerical Label Path2)

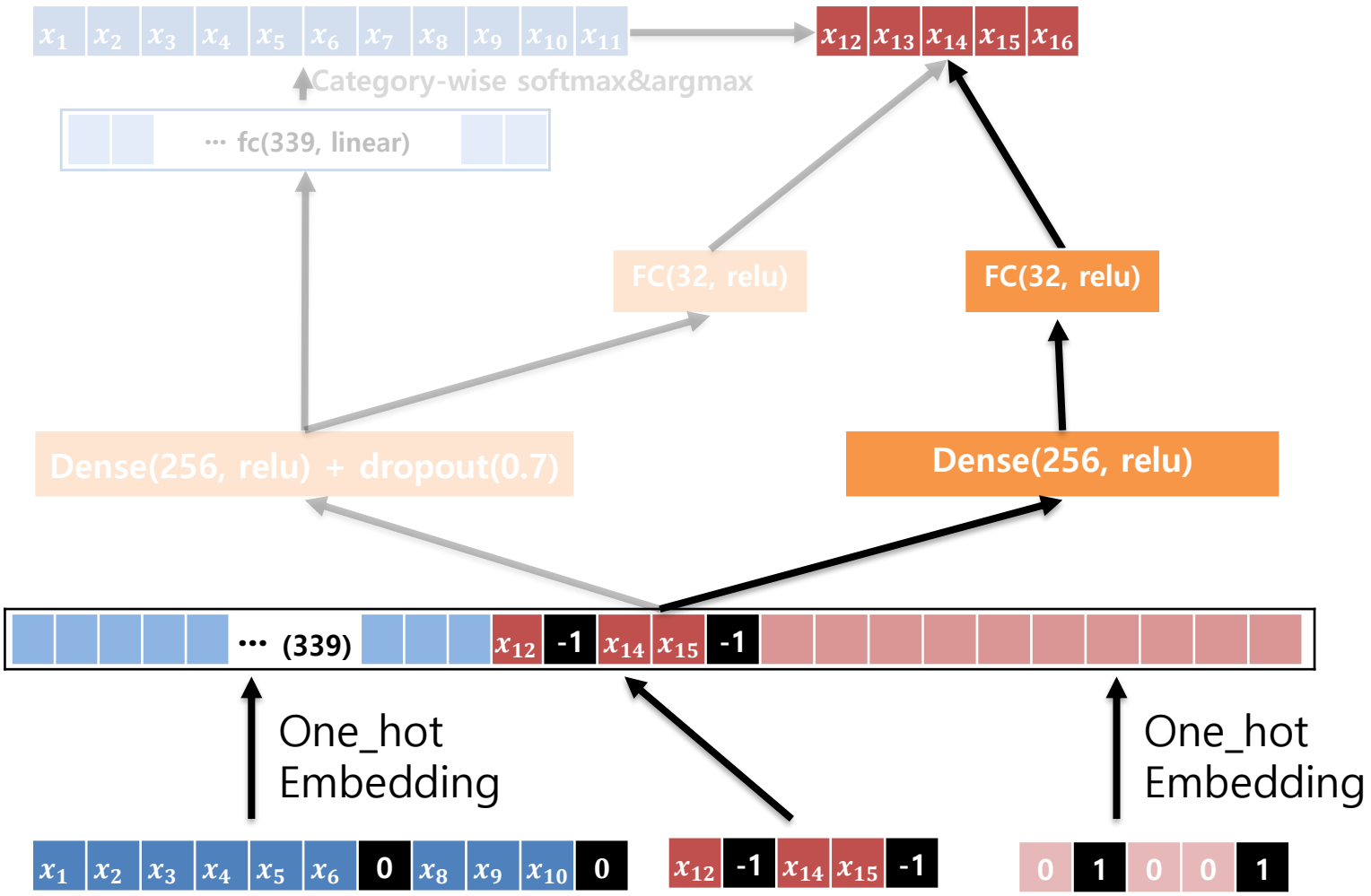
Projection Layer

Hidden Layer2

Hidden Layer

Concatenate Embedding Layer

Input Layer



Model Structure (Numerical Label Path3)

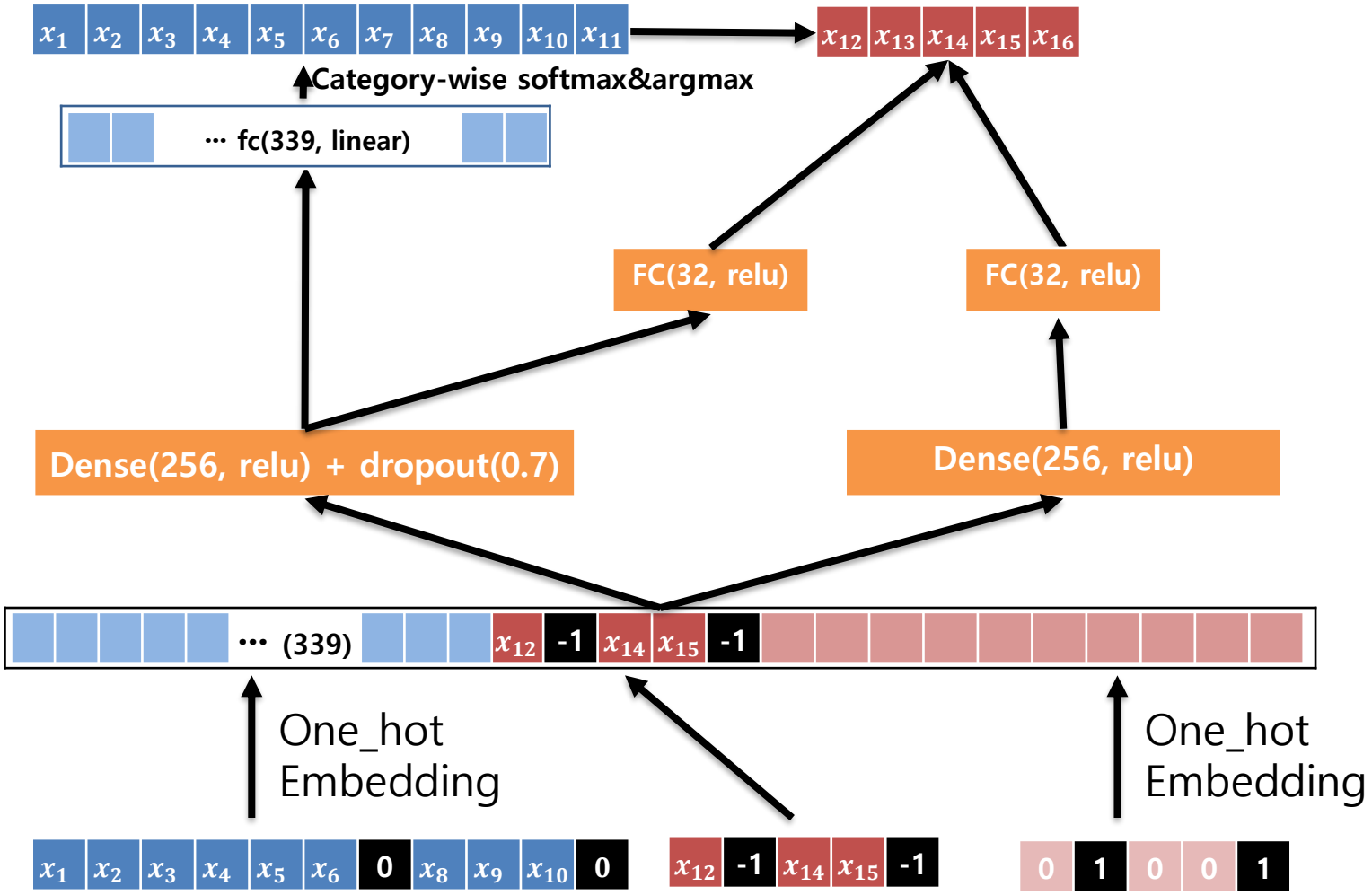
Projection Layer

Hidden Layer2

Hidden Layer

Concatenate Embedding Layer

Input Layer



Result on Validation Test

Total Loss = Categorical Loss + 2* Numerical Loss

- ✓ Categorical Loss : Categorical Cross Entropy per categorical variable
- ✓ Numerical Loss : Average 3 Path MSE on Continuous Label

- Categorical score : $C \times \sum_{i=1}^{k_2} \delta_{c_i d_i}$
- Numerical score : $B \times \sum_{i=1}^{k_1} \exp\left\{-\left(\frac{n_i - m_i}{s_j}\right)^2\right\}$

3 different validation-data split Models

- ✓ 2 Random Shuffle split
- ✓ 1 Latest Split

Final Model is ensemble 3 models by averaging scores

Model No.	Split Type(seed)	Category Score	Numerical Score	Total Score
1	Random (1000)	0.6417	0.9479	1.5896
2	Random (2000)	0.6470	0.9468	1.5938
3	Latest	0.6103	0.9440	1.5544

Conclusion

- DAE 를 이용하여 미래 교통사고사망정보의 데이터를 예측(복원)하는 모델을 개발

- Numeric 변수에 대해서는 스코어 기준 평균 0.94점의 높은 성능을 보임
  Categorical 변수에 대해서는 스코어 기준 평균 0.63점의 상대적으로 낮은 성능을 보임

- 해당 모델은 변수 수의 확장에 따른 Flexible한 모델
  데이터별 변수가 많아지면 전체 성능을 향상할 수 있음
  즉, 현재는 테스트셋에 존재하는 변수만 사용하여 학습을 하였으나 트레이닝 셋에 존재하는 기타 변수들에 대해서도 추가 수집이 가능하다면 성능이 높아질 것으로 보임

- 또한 현재 모델은 해당 도메인 이외에도 의료 레코드, 금융정보 등 타 도메인의 결측치가 많은 데이터에 대해서도 호환 가능한 모델

Thank you