지식 그래프 임베딩을 활용한 시각정보 기반 질의응답

발표자 : 김민준

김민준, 송승우, 나현우, 김은경, 임경태
한밭대학교, 서울과학기술대학교

Contents

01. Introduction

Visual Question Answering(VQA)?

- 주어지는 이미지와 질의에 대해 적절한 답변을 도출하는 시스템
- 일반적인 VQA task는 이미지와 질의에 대한 얕은 이해만으로 답변이 가능

Q : What time is it?

A : 11:55

01. Introduction

Knowledge-Based VQA

- 답변 도출을 위해 이미지와 질의에 대한 정보 뿐만 아니라, 외부지식의 정보가 필요

- 외부지식을 활용하여 넓고 복잡한 질문에 대한 답변 가능

- 외부지식으로 위키피디아의 caption 정보를 활용하거나 (OK-VQA, KVQA 등), 지식그래프 활용(FVQA 등)

01. Introduction

Knowledge-Based VQA

KVQA

Q : What can the red object on the ground be used for?
KB : Khan with United States Secretary of State Hillary Clinton in 2009.
A : Aamilr Khan

OK-VQA

Q : What sort of vehicle used this item?
KB : A fire truck is an emergency road vehicle for firefighters…
A : fire truck

02. Dataset

BOK-VQA
(Bilingual Outside-Knowledge VQA)

- 이미지-질문 쌍과 지식그래프가 주어짐.
- 지식 그래프 : [Head, Relation, Tail]

Q : What is the range of the instrument in the image?
(이미지 속 악기의 음역은 몇이야?)
KB : [violin, range, 130]
A : 130

Q : To which superorder does the animal in the image belong?
(이미지 속 동물은 어느 상목에 속해?)
KB : [shrimp, superorder, eucarida]
A : eucarida

02. Dataset

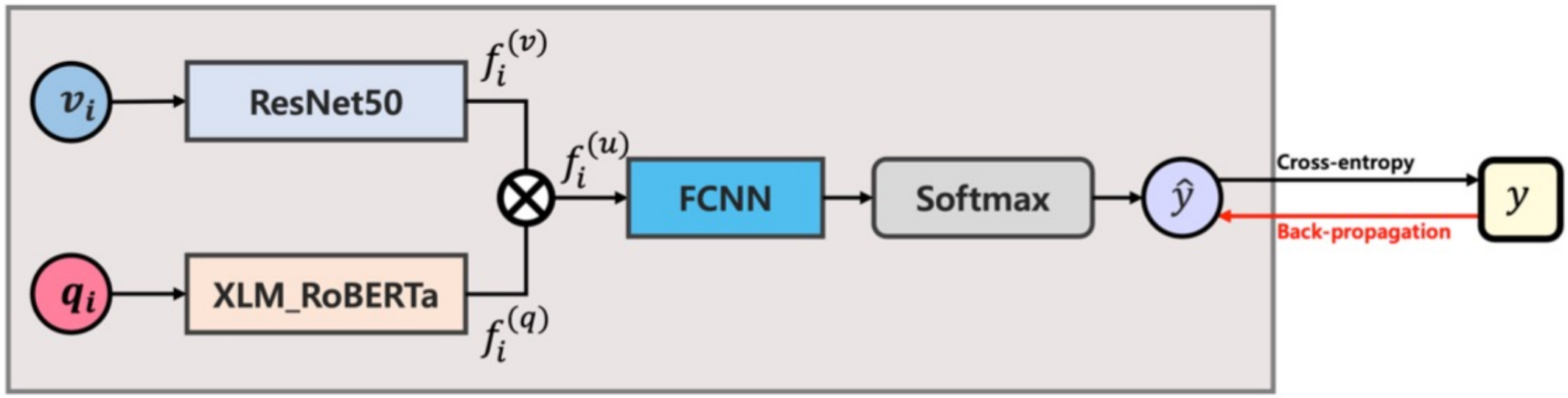
BOK-VQA dataset statistics

of images : 17,836
# of questions : 17,836
# of relations : 43
# of entities : 3,972

03. Method

VQA Baseline Model

v_i : $i\text{-}th$ image data
q_i : $i\text{-}th$ question data
$\otimes$: element-wise operation



$$f_i^{(v)} = \text{ResNet}(v_i)$$

$$f_i^{(q)} = \text{XLMRoBERTa}(q_i)$$

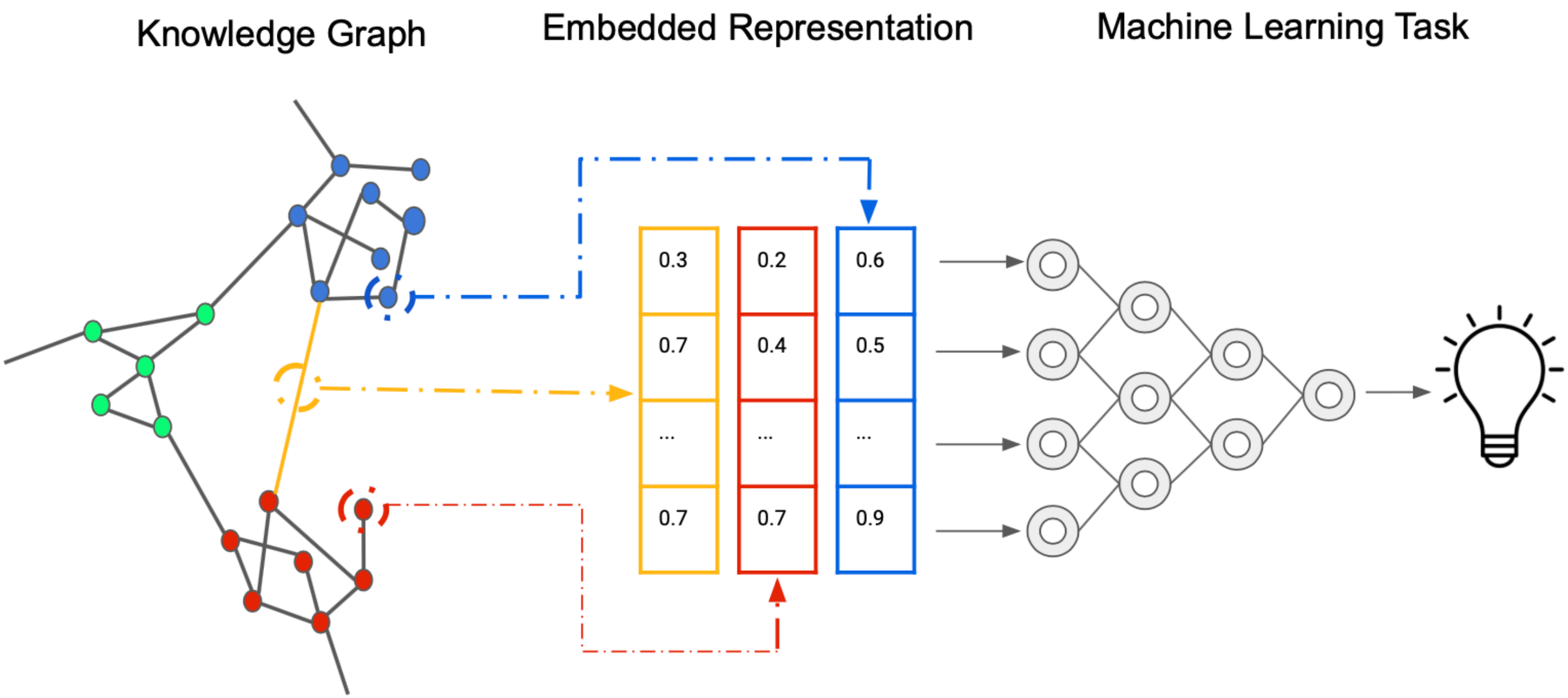
$$f_i^{(u)} = f_i^{(v)} \otimes f_i^{(q)}$$

$$f_i^{(u')} = W^{(\alpha)}\text{ReLU}(W^{(\beta)} f_i^{(u)} + b^{(\beta)}) + b^{(\alpha)}$$

$$\hat{y}_i = \arg\max_c (Softmax(f_{i,c}^{(u')}))$$

03. Method

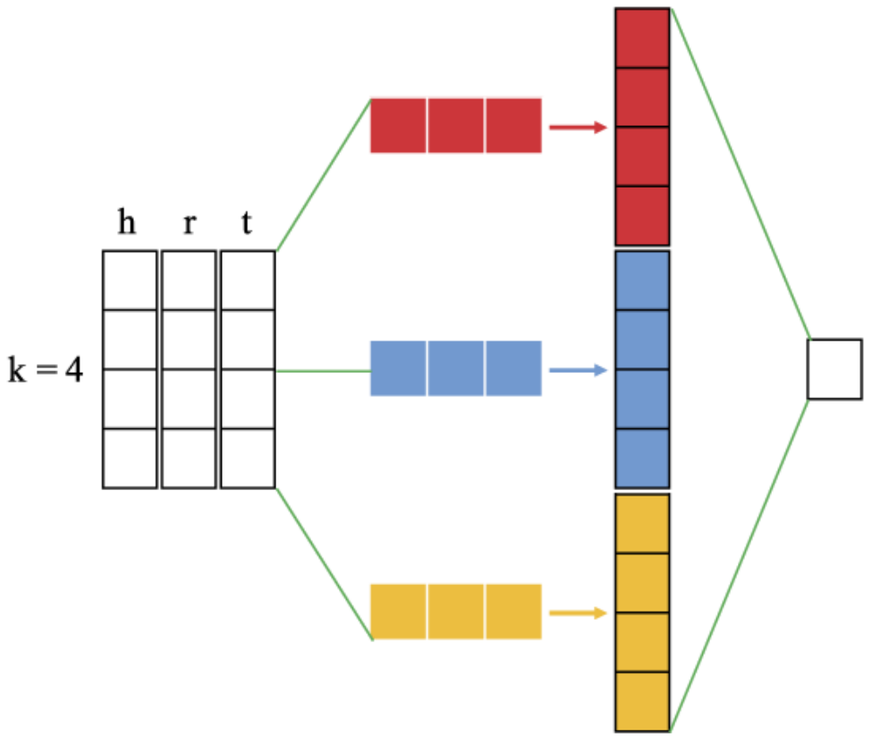
Knowledge Graph Embedding



이미지 출처 : https://en.wikipedia.org/wiki/Knowledge_graph_embedding

03. Method

Knowledge Graph Embedding : ConvKB



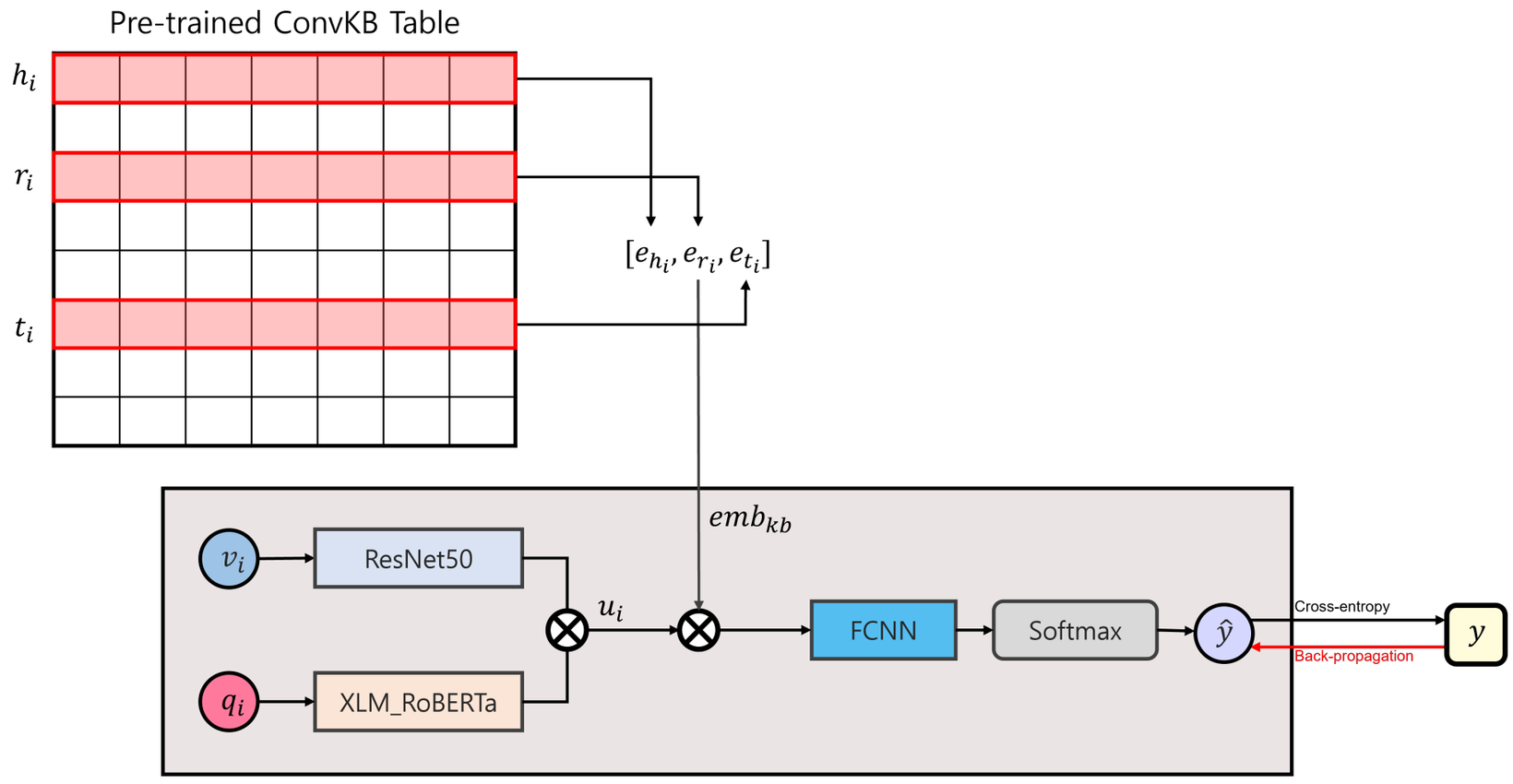
$$\mathcal{L}_c = \sum_{(h,r,t)\in\{\mathcal{K}\cup\mathcal{K}'\}} \log(1 + e^{g(h,r,t)} + \frac{\lambda}{2}||w||_2^2)$$

$$g(h,r,t) = \begin{cases} f(h,r,t), & \text{for } (h,r,t) \in \mathcal{K} \\ -f(h,r,t), & \text{for } (h,r,t) \in \mathcal{K}' \end{cases}$$

$$f(h,r,t) = W \cdot (\mathbf{ReLU}([e_h; e_r; e_t] * \Omega)) + b$$

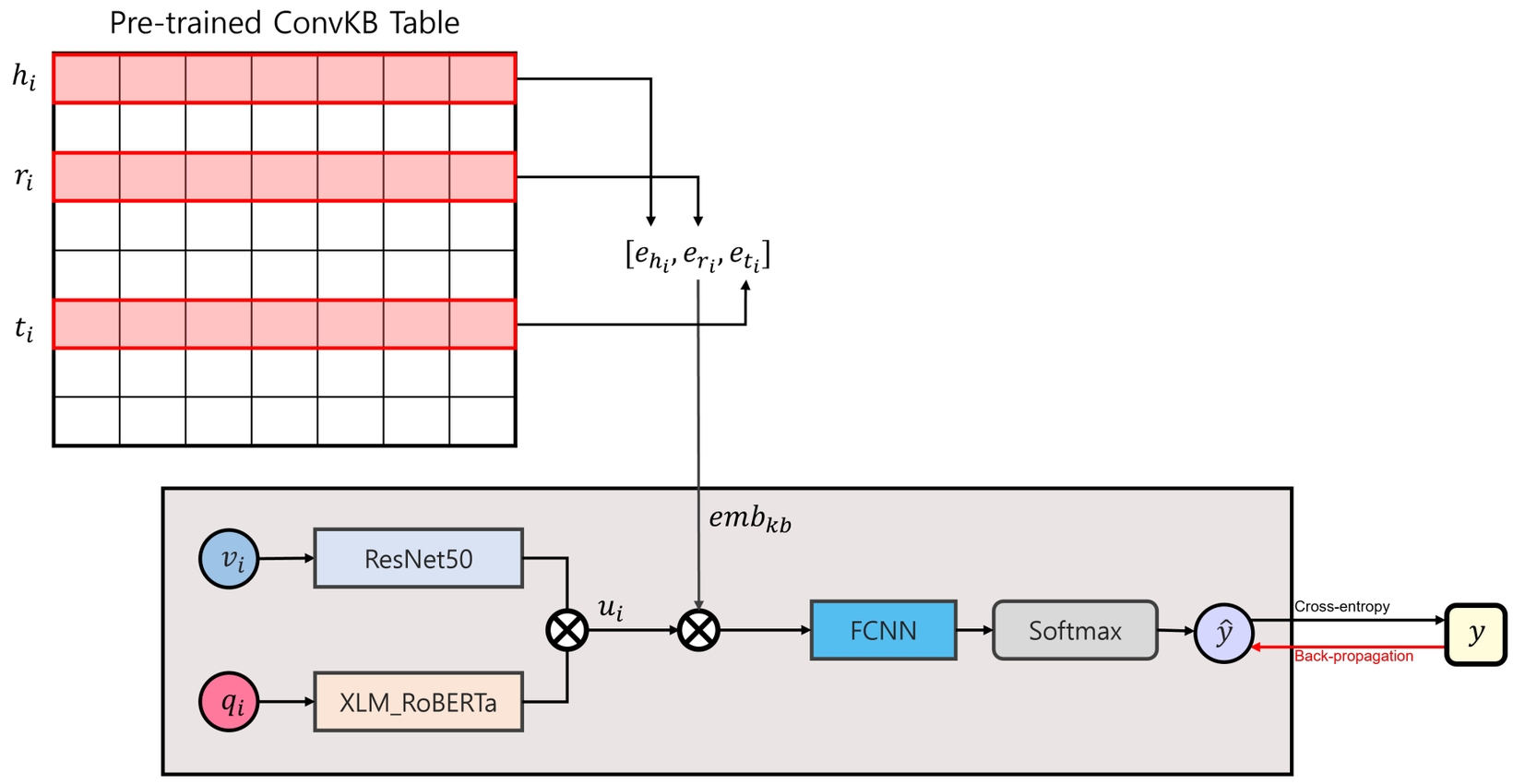
03. Method

VQA with KGE



03. Method

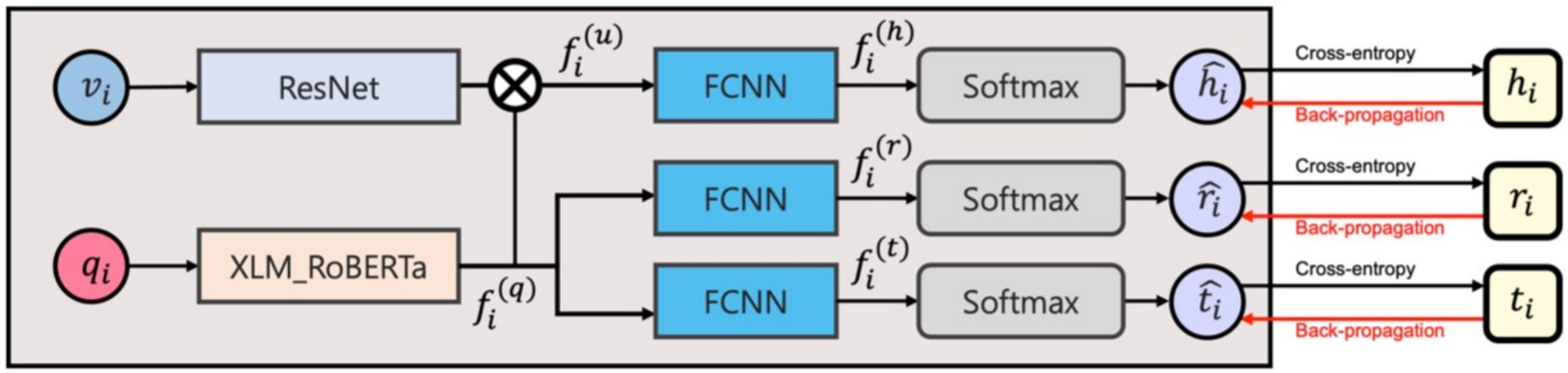
VQA with KGE



- 주어지는 이미지-질의와 관련된 외부 지식이 무엇인지 알고 있다는 가정
- 현실 세계에서 적용 불가능

03. Method

Triple prediction



- Head는 이미지(객체)와 질의에, Relation, Tail은 질의에 많은 정보가 있다는 점 활용

Proposed Model

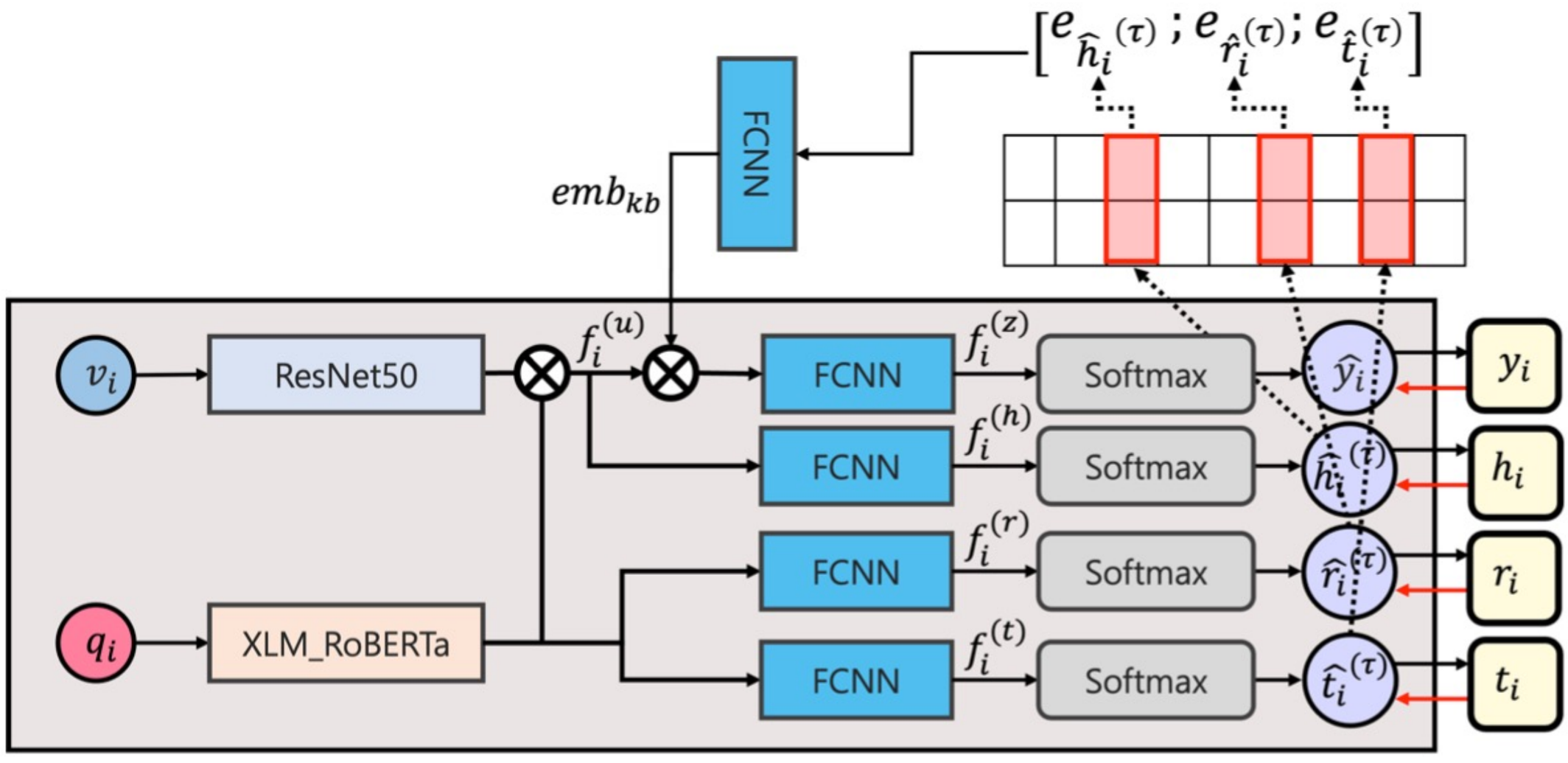


Figure 4: Illustration of GEL-VQA architecture

04. Results

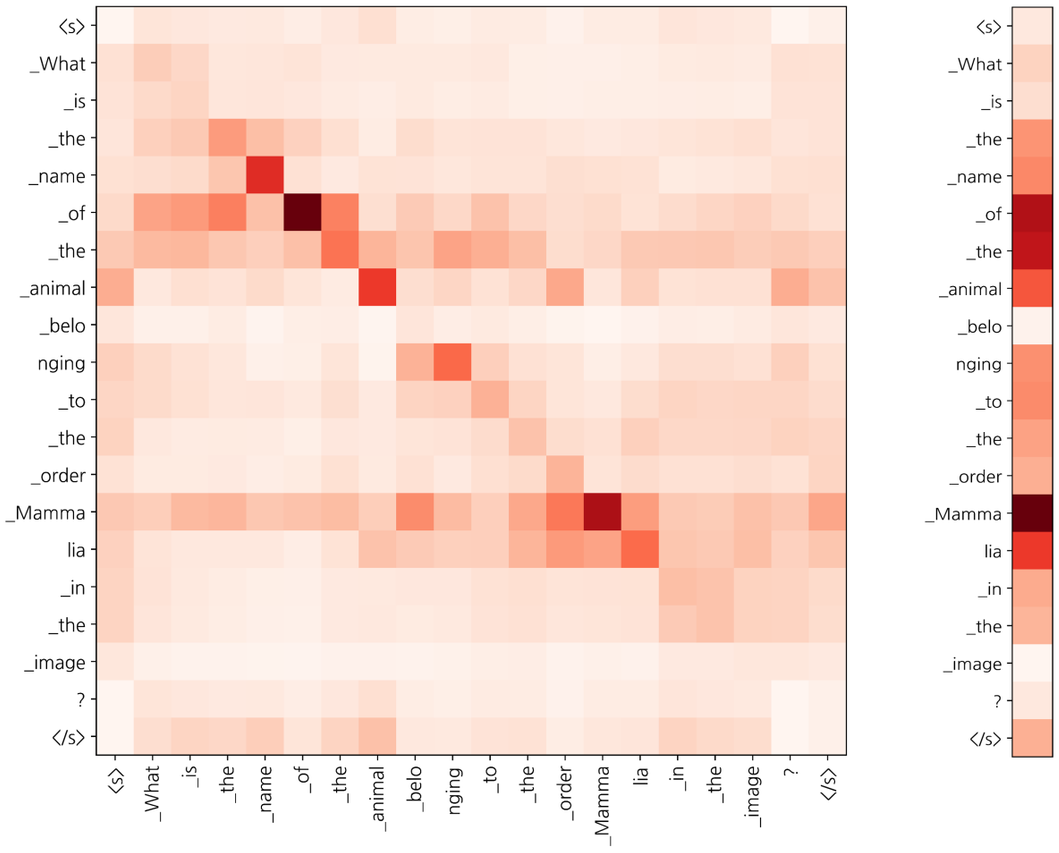
Accuracy

Language	BASELINE	GEL-VQA(IDEAL)	GEL-VQA	GEL-VQA + TF	GEL-VQA + TF + ATTN
Bilingual	21.51±1.81	66.01±1.83	45.08±0.94	48.07±1.33	48.11±1.50
English	21.90±2.34	66.68±1.15	47.83±0.56	51.64±0.88	50.74±0.90
Korean	21.16±1.50	72.25±1.29	50.30±2.24	53.40±2.73	55.48±1.89

Table 1: Table summarizing the performance of five different models across three languages: Bilingual, English, and Korean.

04. Results

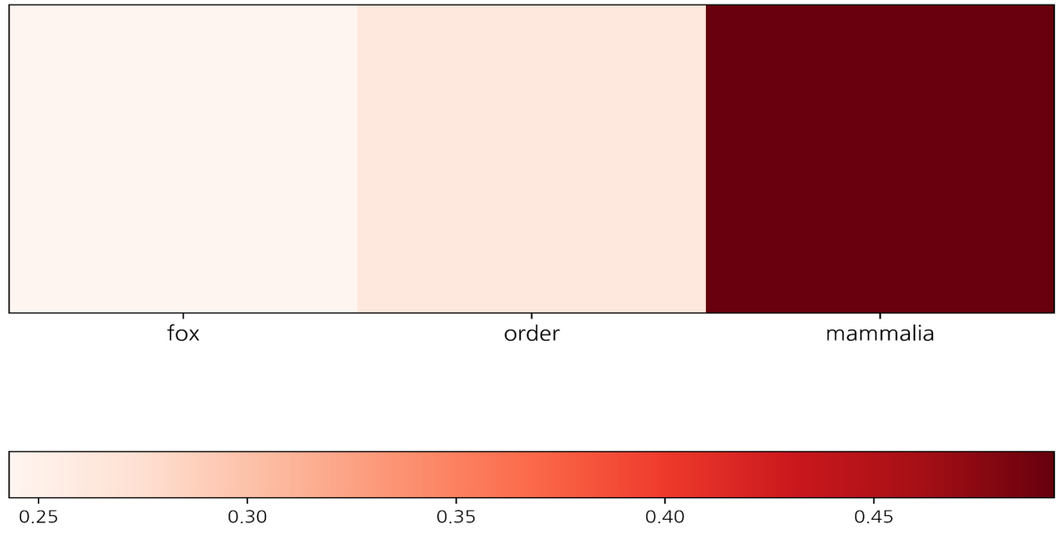
Analysis



Q : What is the name of the animal belonging to the order Mammalia in the image?
(이미지에서 포유강에 해당하는 동물이 뭐야?)

KB : [fox, order, mammalia]

Answer : fox



Q&A