SERMON:
Aspect-Enhanced Explainable Recommendation with Multi-modal Contrastive Learning

APL Paper Review Semina

UIHYUN KIM

1. Background
2. Main Concept
3. Model
4. Experiment

Background

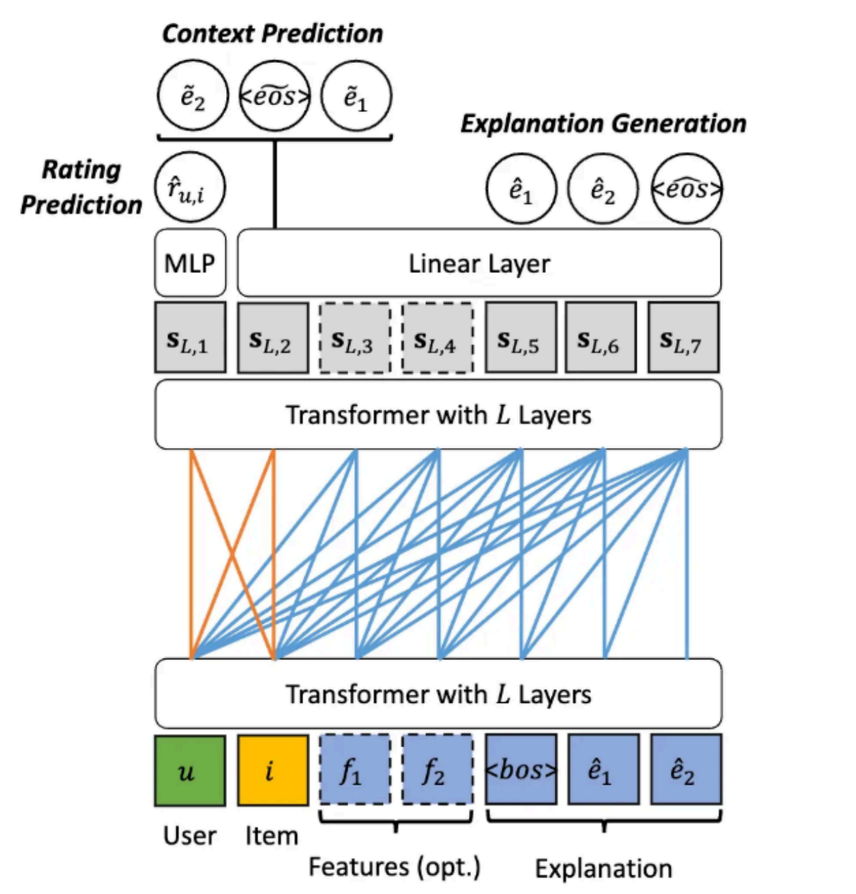
Case 1–Truth	The price of this hotel is very cheap and it also has Internet access and the sound insulation is very good. Will come again! Aspects: (price, cheap) (hotel, insulation)
NRT	The price of this hotel is best!
PETER	The hotel location is well and looks very nice!
SERMON	The room is pretty **insulation** and **traffic** here is very **convenient** and it's **cheap**.
Case 2–Truth	This phone charges very fast and takes beautiful pictures, the only downside is that it is too heavy. Aspects: (charges, fast) (phone, heavy)
NRT	The phone is bad.
PETER	The phone is poor, and it is very unreasonable to buy it.
SERMON	The phone **charges fast**, the **screen** is **well**. The **weight of phone** is a bit **unreasonable**.
Case 3–Truth	The restaurant is very expensive and there is no signal in the house, the worst thing is the food is very bad, I can't stand it. Aspects: (restaurant, expensive) (foods, bad)
NRT	The foods is nice.
PETER	It tastes terrible! The waiting time here is too long!
SERMON	**Terrible**! The **restaurant** here is **expensive** and the service here is very **poor**. Food here is **very not delicious**!

- ERS (Explainable Recommender Systems)
  - 사용자 리뷰를 활용하여 사용자-아이템 선호도를 예측하고, 추천 근거 문장을 생성
  - NLP 기술이 빠르게 발전하면서, 생성되는 설명의 품질도 향상되어 빠르게 발전
  - Ex) 타이베이 호텔 추천 + 근거 설명 "중심지와 가깝고 방이 깨끗함"

Background

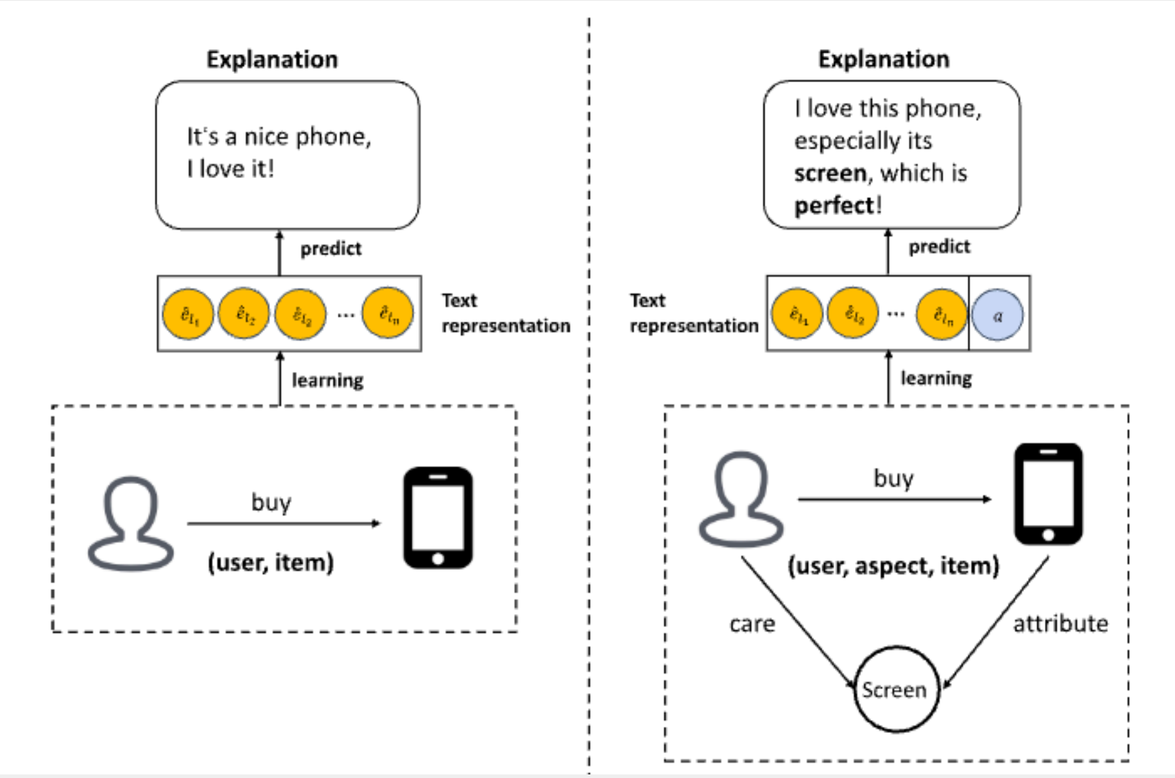
- 기존 ERS의 한계
  - User ID, Item ID 와 리뷰 텍스트의 Embedding Space 불일치
    → ID와 텍스트를 동시에 모델에 쓸 수 없어 추천 근거 문장을 생성할 때 Personalization 이 어려움
    → 추천을 위한 User-Item 모델과 설명 생성을 위한 모델 분리 (GRU, Knowledge Graph)

Background



- **PETER: PErsonalized Transformer for Explainable Recommendation (2020)**
  - 당시 ERS에 Transformer를 처음으로 적용한 논문
  - Rating Prediction, **Context Prediction**, Explanation Generation 을 묶어 Multi-task learning 으로 학습
  - **Context Prediction Task**: User ID, Item ID, Explanation으로 주요 단어를 추출 → ID 벡터가 언어적 특성을 가지게 학습

Main Concept

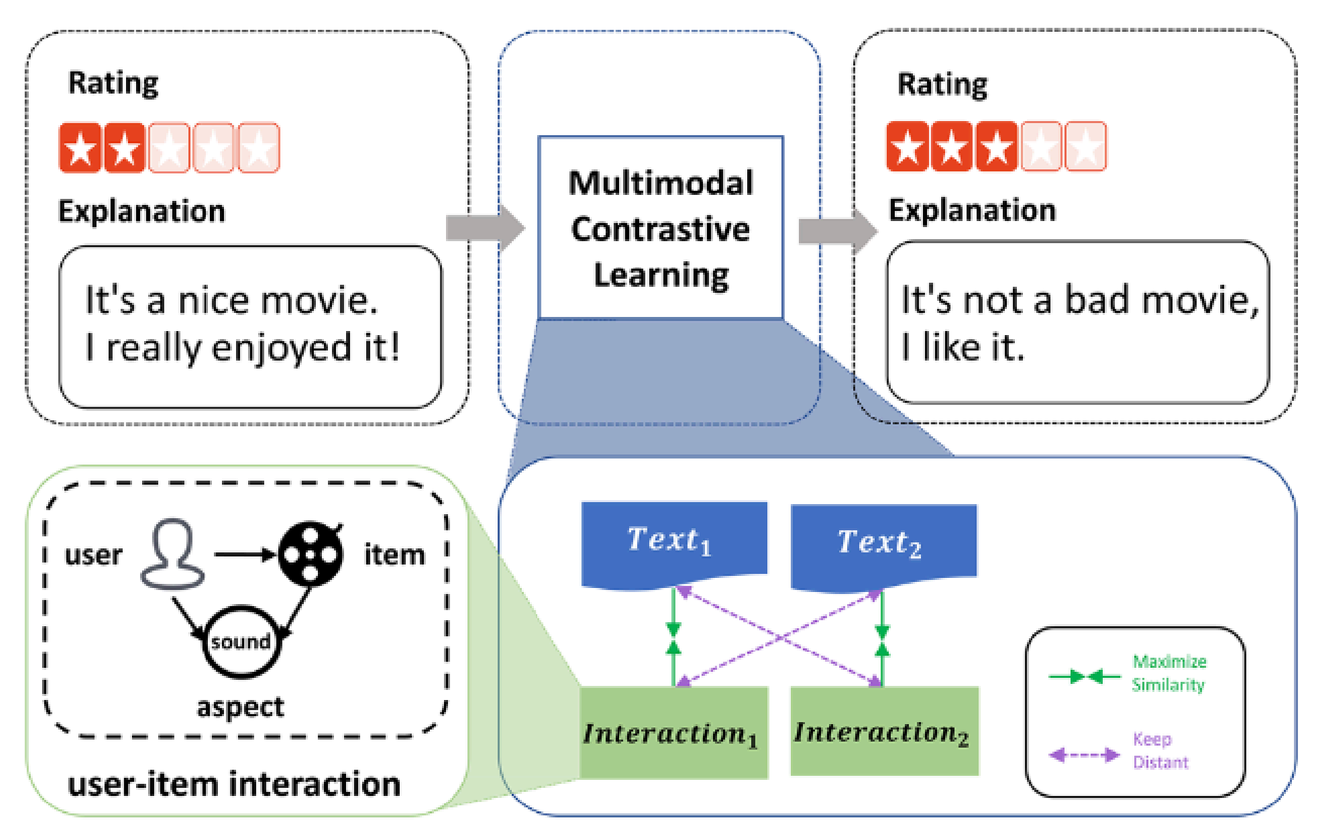


- **SERMON: Aspect-Enhanced Explainable Recommendation with Multi-modal Contrastive Learning (2025)**
  - 유저-아이템 간의 상호작용 정보로서 ***Aspect** 를 ERS에 함께 적용
  - **Aspect**: 리뷰로부터 추출한 (명사, 형용사) 쌍, 리뷰 텍스트로 부터 얻을 수 있는 상호작용 정보

  → 평점 예측(추천) 정확도 향상

  → 문장 생성에 User-Item의 상호 작용 반영하여 개인화 생성 능력 향상

Main Concept

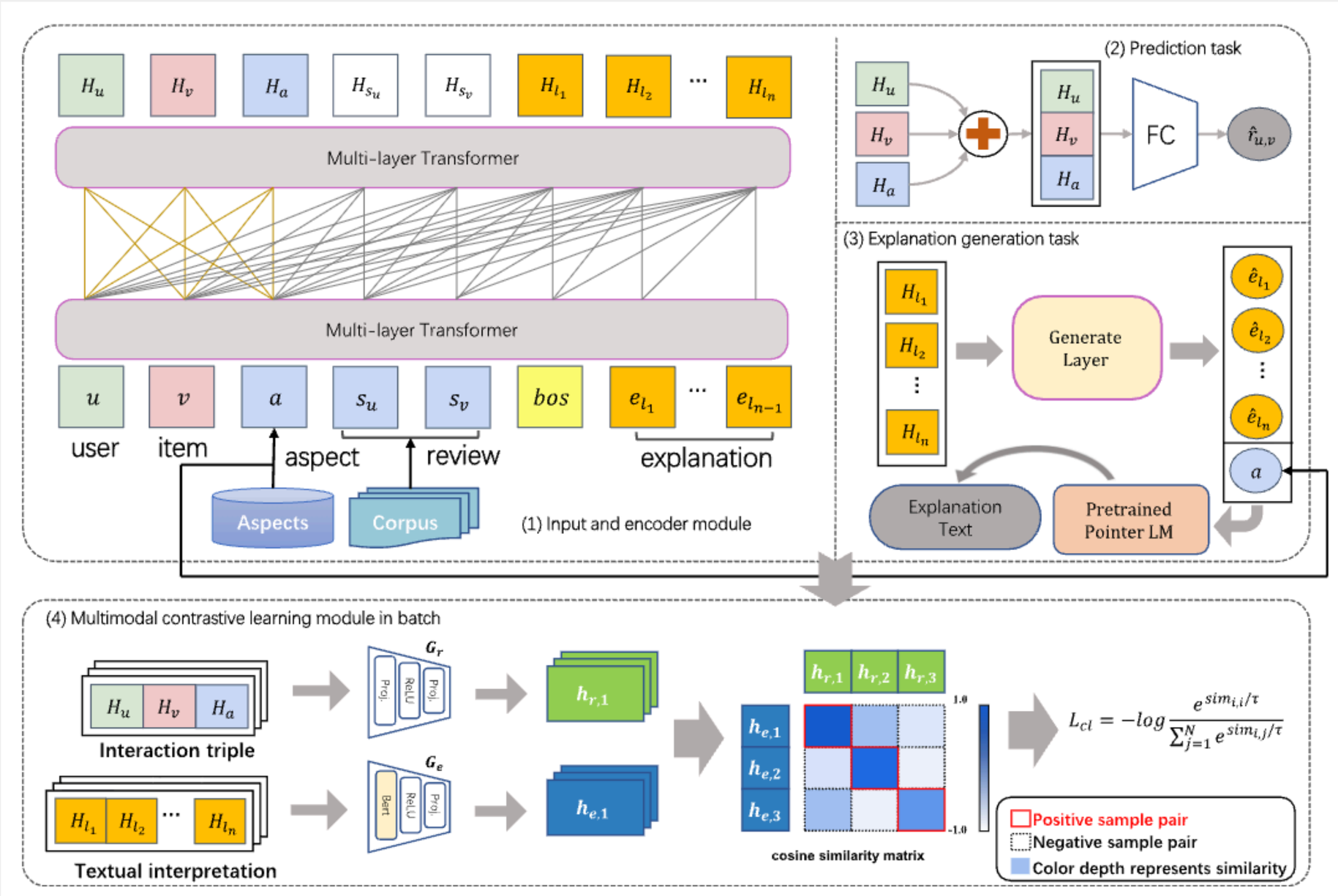


◦ 특정 User-Item 쌍의 Interaction과 Explanation 의 내용은 양의 상관 관계가 있을거라 전제

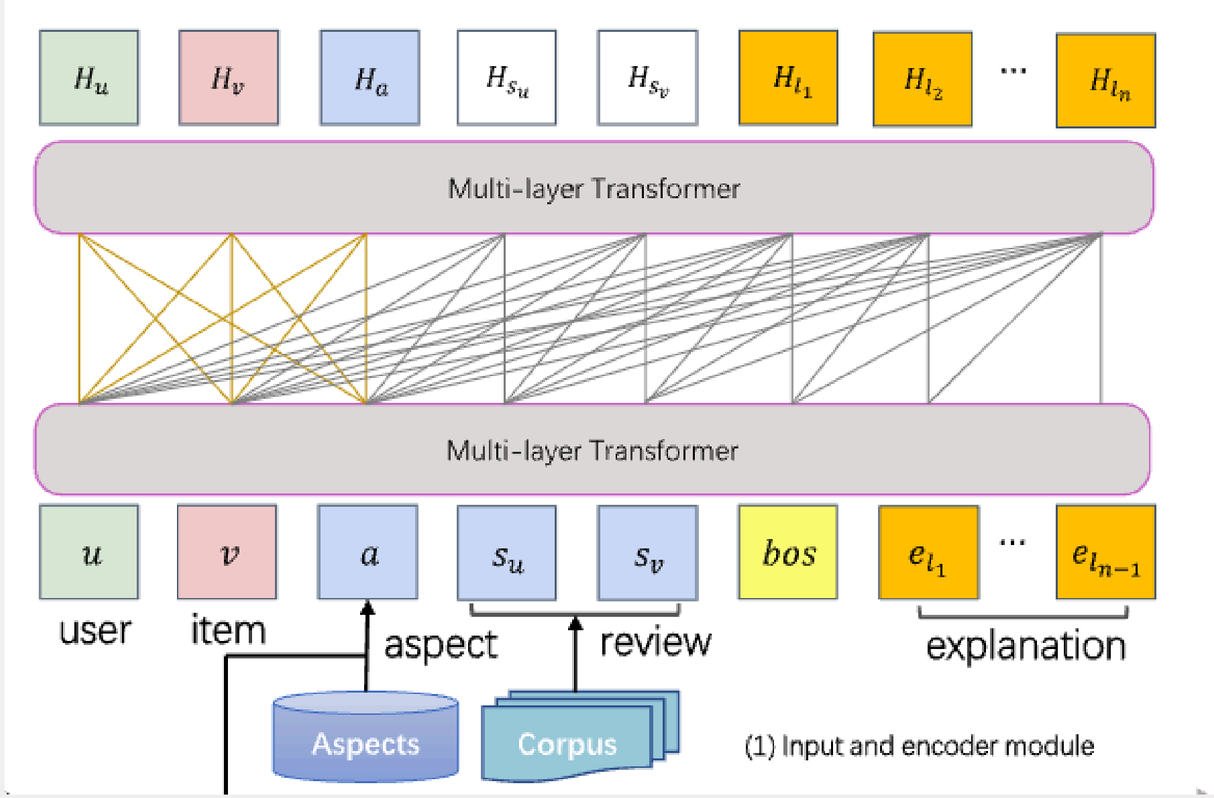
◦ Embedding space가 다른 둘 사이의 상관 관계를 학습하기 위해 Contrastive Learning 적용

→ 평점과, 설명이 서로를 참조하도록 학습

Model


Multi-layer Transformer
Multi-layer Transformer
user
item
aspect
review
explanation
Aspects
Corpus
(1) Input and encoder module
(2) Prediction task
FC
(3) Explanation generation task
Generate Layer
Explanation Text
Pretrained Pointer LM
(4) Multimodal contrastive learning module in batch
Interaction triple
Textual interpretation
Proj.
ReLU
Bart
cosine similarity matrix
Positive sample pair
Negative sample pair
Color depth represents similarity

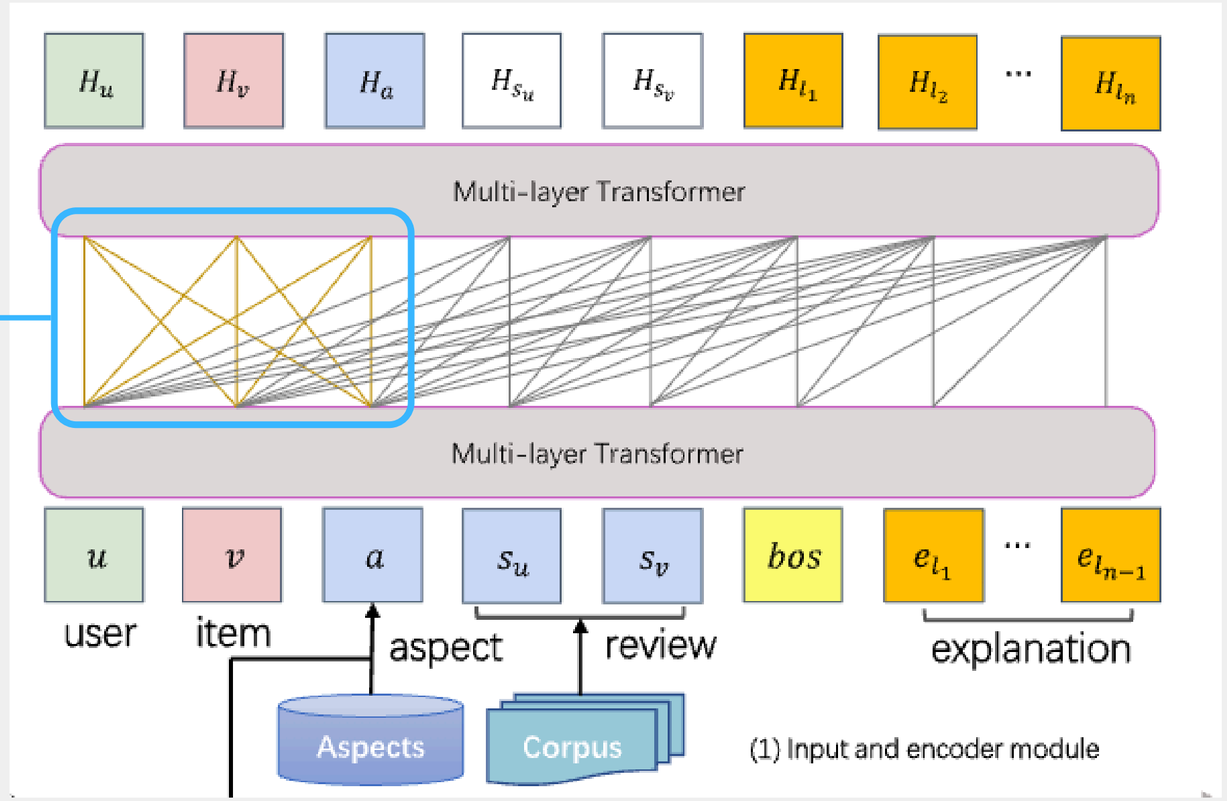
Model



- **Input Representation**
  - User ID, Item ID 임베딩 → u, v
  - 특정 User-Item 쌍에서 Aspect 추출, SBERT 임베딩 → a
  - User, Item Feature 리뷰 문장, SBERT 임베딩 → s
  - Explanation Ground Truth, 토큰 단위 임베딩 → e
  - 4 종류의 토큰들에 Self-Attention 을 수행하여 히든 벡터 H 추출

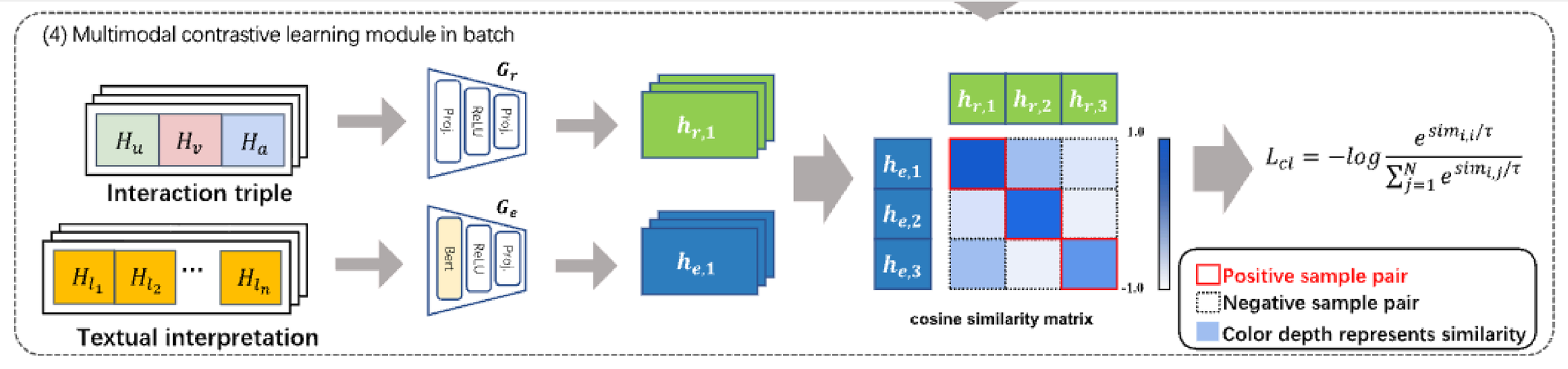
Model

Self-Attention Mask를 u, v, a 엔 주지 않아
u, v, a 가 서로를 참조하여 상호 작용 처럼 학습



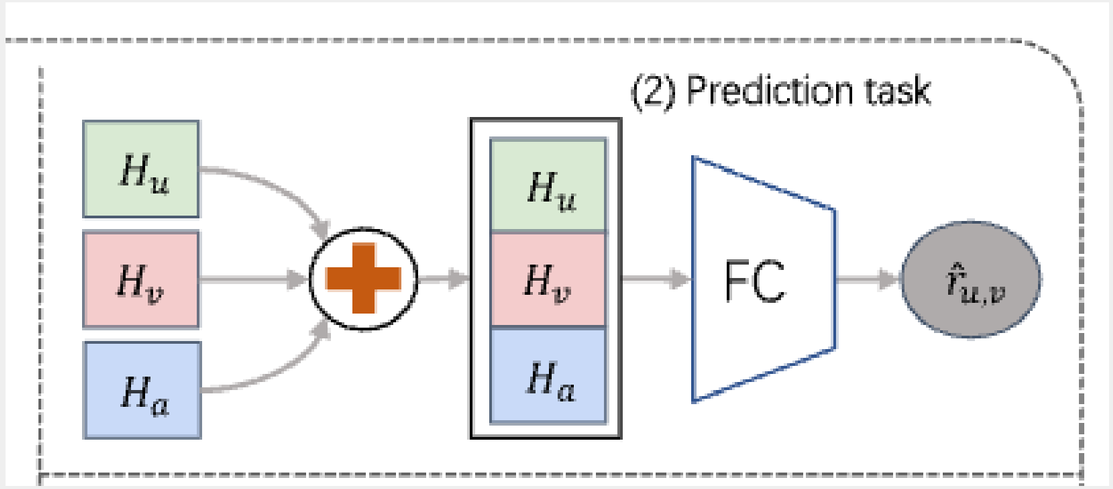
- **Input Representation**
  - User ID, Item ID 임베딩 → u, v
  - 특정 User-Item 쌍에서 Aspect 추출, SBERT 임베딩 → a
  - User, Item Feature 리뷰 문장, SBERT 임베딩 → s
  - Explanation Ground Truth, 토큰 단위 임베딩 → e
  - 4 종류의 토큰들에 Self-Attention 을 수행하여 히든 벡터 H 추출

Model

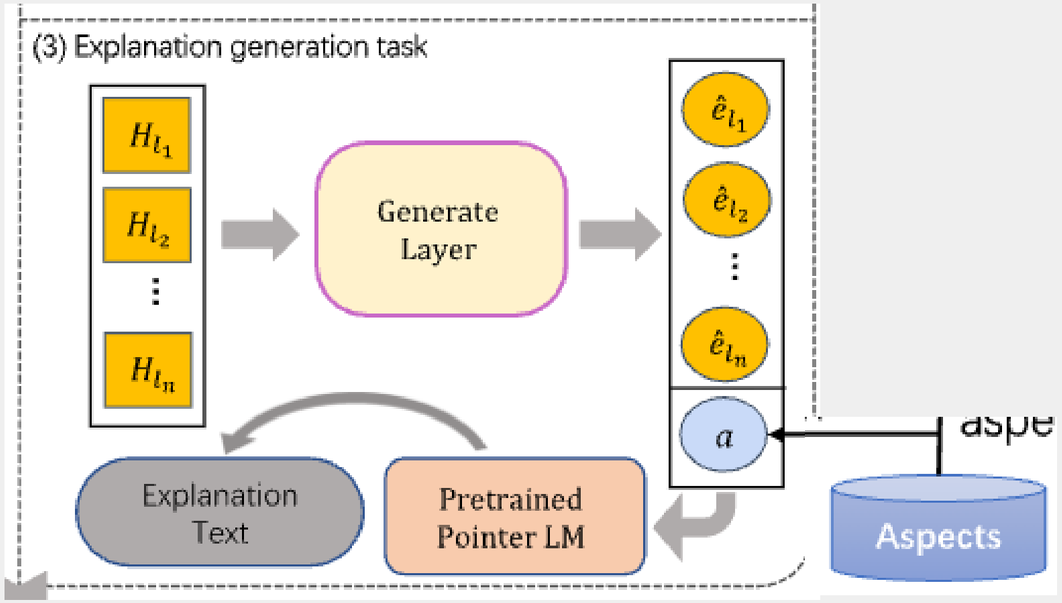


- **Multi-modal Contrastive Learning**
  - User ID, Item ID, Aspect 벡터 concatenate → Interaction triple vector
  - Explanation 벡터
  - 변환용 레이어 Gr, Ge 를 통해 두 벡터를 같은 차원으로 매핑
  - 같은 차원으로 매핑된 두 벡터로 contrastive learning 수행

Model

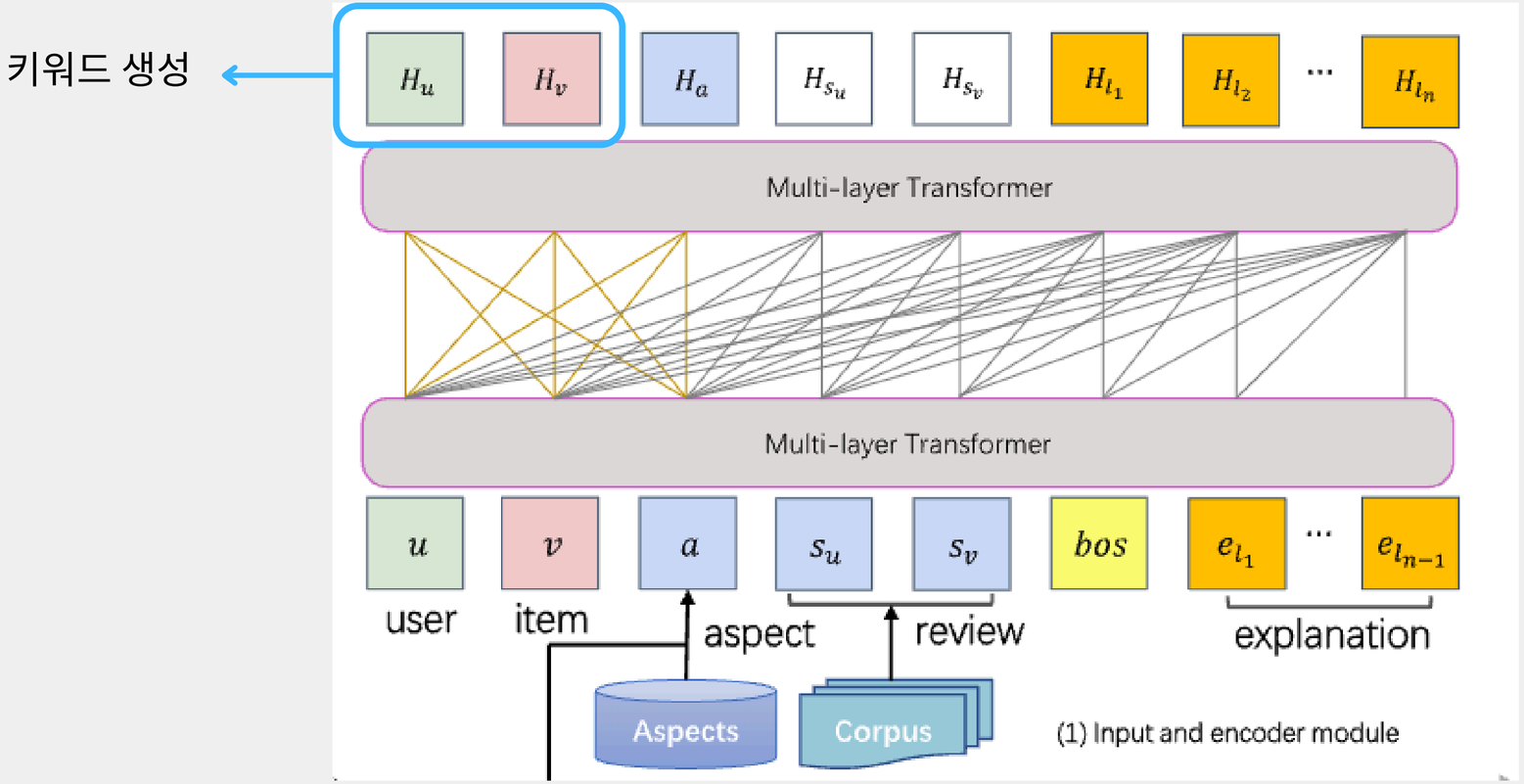


- **Rating Prediction**
  - Interaction triple vector를 MLP에 통과 시켜 평점 예측
  - 손실 함수로 MAE / RMSE 사용



- **Explanation Generation**
  - Explanation 의 last hidden 벡터로 문장 생성 → e
  - 생성된 문장과 input 에 넣었던 Aspect 를 합쳐 *POINTER에 전달하여 최종 Explanation 문장 생성
  - POINTER: 입력으로 넣은 특정 단어를 반드시 포함하도록 문장을 생성
  - 손실 함수로 Cross-Entropy Loss 사용

Model



- **Context Prediction**
  - Transformer를 통과한 결과 ID들의 임베딩 벡터들이 단어 임베딩 벡터들과 유사해져 고유성을 잃고 반복적인 문장을 생성하는 문제 발생
  - ID 벡터들로만 문장을 생성하여 Ground-truth와 비교하여 이를 방지
  - 손실 함수로 Cross-Entropy Loss 사용

Model

$$\mathcal{L} = rl\mathcal{L}_r + \lambda_c\mathcal{L}_c + el\mathcal{L}_e + cl\mathcal{L}_{cl} + \lambda_l\|\Theta\|_2^2$$

- **Multi-Task Learning**
  - Lr: rating prediction loss
  - Lc: context prediction loss
  - Le: text generation loss
  - Lcl: multi-modal contrastive loss
  - Θ, λl: L2정규화

Experiment

- **Data**
  - Yelp: 식당 리뷰
  - Amazon: 핸드폰 리뷰
  - TripAdvisor: 호텔 리뷰
  - 학습(80%) / 검증(10%) / 테스트(10%)

Datasets	Yelp	Amazon	TripAdvisor
Number of users	27,147	157,212	9,765
Number of items	20,266	48,186	6,280
Number of reviews	1,293,247	1,128,437	320,023
Records per user	47.64	7.18	32.77
Records per item	63.81	23.41	50.96

- **Setting**
  - 임베딩 벡터 차원 수: 384
  - 생성 문장 최대 길이: 15~20 단어
  - Loss Parameters
    - λc = 1.0 (context explanation)
    - rl = 1.0 (rating prediction)
    - cl = 0.2 (multimodal contrastive learning)
    - el = 1.0 (explanation), Xavier 초기화 사용
  - Optimization
    - Adam
    - Learning rate: 0.1
    - L2 정규화: 0.0001
    - Early stopping: 3 epochs

- **POINTER 모델 학습**
  - Wiki corpus로 POINTER 사전 학습 (10일 소요)
  - 이후 Amazon, Yelp, TripAdvisor에 대해 각각 4~5일 동안 파인튜닝

Experiment

- **Baseline Methods**
  - Rating Prediction
    - PMF: 행렬 분해 협업 필터링
    - NARRE: 텍스트 피처에 특화된 신경망 협업 필터링
    - DAML: user-item 문서를 사용하여 연관 관계 표현을 개선
    - RGCL: 리뷰 텍스트 이용한 그래프 신경망
  - ERS
    - NRT: GRU based
    - CAML: 유저의 과거 리뷰들로 유저 표현, 그 중 가장 연관 있는 리뷰들을 추려 문장 생성
    - ReXPlug: 문장 생성에 GPT-2 사용
    - PETER: ID와 Explanation 학습에 Transformer 적용

Experiment

	Yelp		Amazon		TripAdvisor	
	R ↓	M ↓	R ↓	M ↓	R ↓	M ↓
PMF	1.097	0.883	1.235	0.913	0.870	0.704
NARRE	1.028	0.791	1.176	0.865	0.796	0.612
DAML	1.014	0.784	1.173	0.858	0.793	0.617
RGCL	1.008	0.784	1.160	0.872	0.791	0.611
NRT	1.016	0.796	1.188	0.853	0.797	0.611
CAML	1.036	0.798	1.191	0.888	0.818	0.622
PETER	1.017	0.793	1.181	0.863	0.814	0.635
SERMON	**1.002**	**0.784**	**1.159**	**0.837**	**0.791**	**0.599**

Rating Prediction

Datasets	Metrics	Baselines				Ours				Improvement
		NRT	CAML	ReXPlug	PETER	SERMON-M	SERMON-A	SERMON-R	SERMON	
Yelp	BLEU1	10.5	9.91	8.59	10.29	10.42	10.54	10.58	**10.66**	3.4%
	BLEU4	0.67	0.56	0.57	0.69	0.70	0.72	0.73	**0.76**	10.9%
	R2-P	1.95	1.78	1.49	1.91	2.02	2.01	2.13	**2.19**	11.4%
	R2-R	1.29	1.05	1.07	1.31	1.33	1.38	1.40	**1.48**	10.6%
	R2-F	1.35	1.25	1.11	1.43	1.46	1.52	1.55	**1.60**	3.88%
	RL-P	15.88	14.25	13.32	16.07	16.60	16.68	16.69	**16.72**	9.86%
	RL-R	10.72	14.26	9.56	10.14	11.23	10.98	11.01	**11.25**	9.7%
	RL-F	9.53	9.16	8.70	10.26	10.61	10.77	10.79	**10.87**	5.6%
	BERT-S	83.6	83.2	82.2	83.3	84.7	84.9	85.0	**85.1**	2.1%
TripAdvisor	BLEU1	15.78	14.43	12.64	15.33	16.52	16.61	16.62	**16.69**	5.5%
	BLEU4	0.85	0.86	0.71	0.89	1.06	1.10	1.12	**1.18**	24.5%
	R2-P	1.98	1.49	1.61	1.92	2.40	2.48	2.53	**2.66**	25.5%
	R2-R	1.92	1.91	1.49	2.01	2.32	2.46	2.54	**2.68**	25.0%
	R2-F	1.90	1.92	1.61	1.94	2.22	2.36	2.41	**2.49**	22.0%
	RL-P	14.85	13.36	11.38	13.54	15.08	15.23	15.42	**15.67**	5.2%
	RL-R	14.03	12.38	10.22	14.75	15.34	15.59	15.60	**15.86**	6.9%
	RL-F	12.25	12.39	9.97	12.61	13.17	13.30	13.44	**13.63**	7.4%
	BERT-S	82.7	84.8	83.2	86.4	88.1	88.1	88.2	**88.3**	2.15%
Amazon	BLEU1	13.37	11.19	10.8	13.78	14.13	14.27	14.31	**14.46**	4.7%
	BLEU4	1.44	1.12	1.29	1.68	1.70	1.82	1.84	**1.92**	12.5%
	R2-P	2.06	1.48	2.17	2.21	2.56	2.74	2.80	**2.93**	24.5%
	R2-R	2.08	1.23	1.12	2.02	2.59	2.78	2.91	**3.00**	26.3%
	R2-F	1.97	1.24	1.22	1.97	2.11	2.47	2.49	**2.74**	28.1%
	RL-P	12.52	9.32	9.20	12.62	15.88	16.03	16.11	**16.25**	22.3%
	RL-R	12.20	10.11	10.58	12.06	13.97	14.04	14.36	**14.47**	16.6%
	RL-F	10.77	8.11	8.73	11.07	13.28	13.47	13.74	**13.81**	19.8%
	BERT-S	75.4	74.9	75.3	76.2	79.1	79.3	79.3	**79.4**	4.0%

Explanation Generation

BLEU, ROGUE, BERT-S

Experiment

Datasets	Number of users	Number of items	Number of reviews	Review retrieval time (seconds)	Aspect retrieval time (seconds)
TripAdvisor	9,765	6,280	320,023	7.02	46.23
Amazon Cell	157,212	48,186	1,128,437	31.96	75.22
Yelp	27,147	20,266	1,293,247	12.37	62.44

Retrieval Time

Aspect-Enhanced Explainable Recommendation with Multi-mod…

Explainable recommender systems (ERS) aim to enhance users' trust in the systems by offering personalized…

dl.acm.org