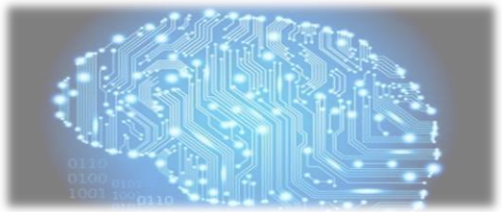
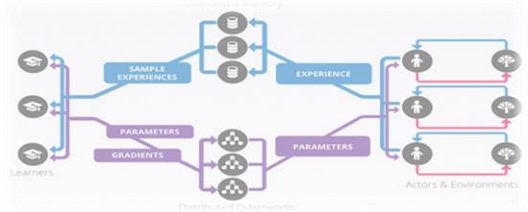



Multimodal Text and Image Classification

AutoML 연구실
석사과정 문아성

22.01.25

TO DO List

- Last week
  - [x] Multimodal Text and Image Classification: 멀티모달 데이터를 학습하는 모델(paper) 탐색
  - [x] 담낭용종 연구 실험 진행
    - 서울대 Test Dataset에 대해 Activation map, Confusion matrix, ROC curve 추가

- This week
  - [x] Multimodal Text and Image Classification: 선정한 모델 구현 및 실험
  - [x] 담낭용종 연구 실험 진행
    - Test Dataset에 대해 Activation map과 Non-neoplastic, Neoplastic 최종 확률 표기

Multimodal Text and Image Classification

- Image and Text fusion for UPMC Food-101 using BERT and CNNs

Gallo, Ignazio, et al. "Image and Text fusion for UPMC Food-101 using BERT and CNNs."

2020 35th International Conference on Image and Vision Computing New Zealand (IVCNZ). IEEE, 2020.

- *Dataset Specification*
  - Number of images: 90,704 *(train: 67,988 test: 22,716)*
  - Number of text: 93,409
  - annotation: 101

Image

Examples of images in the UPMC Food-101 Dataset

Text

Text(recipe) data sample: Apple Pie

Multimodal Text and Image Classification

Proposed Methods

- *Image Classification Model (Inception v3)*
- *Text Classification Model (BERT + LSTM)*
- *Image + Text Classification Model: late fusion method*
- *Image + Text Classification Model:* ***early fusion method***

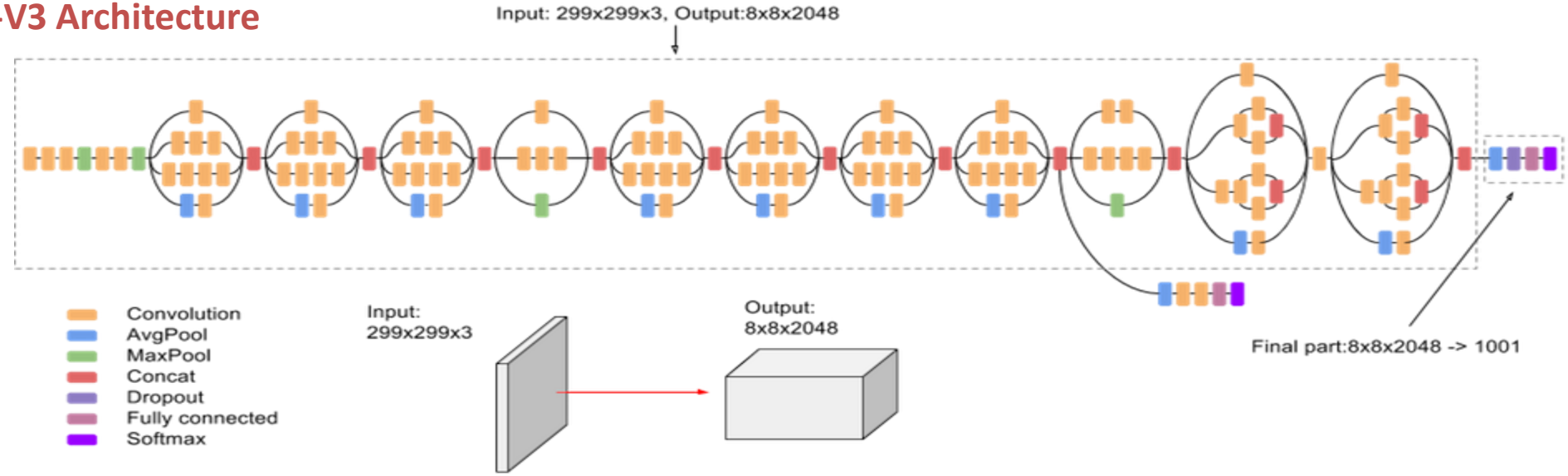
Multimodal Text and Image Classification

Image Classification Model

TABLE I: Custom model defined for the classification of UPMC Food-101 [4] images

Model for UPMC Food-101 Images Classification
Input Image - shape (299 x 299 x 3)
Inception V3 (without the last two layers)
Average Pooling 2D (8 x 8)
Dropout - probability 40%
Flatten
Dense *Softmax* - 101 units

Inception-V3 Architecture



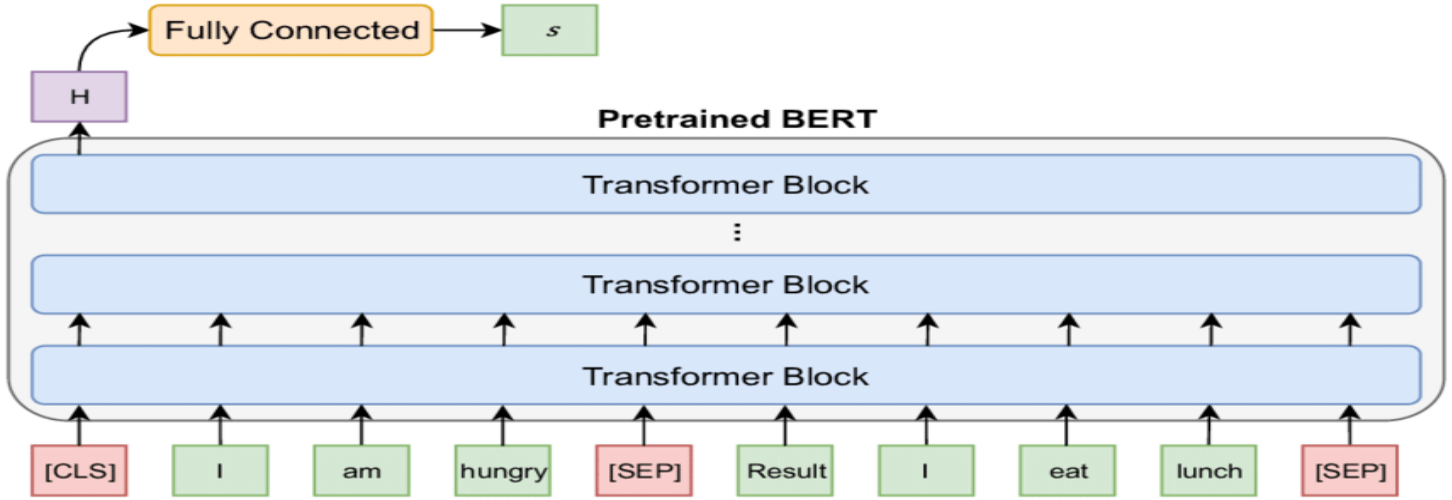
Multimodal Text and Image Classification

Text Classification Model

TABLE II: Custom model defined for the classification of UPMC Food-101 [4] texts

Model for for UPMC Food-101 Texts Classification
Inputs (word ids, masks, segment ids)
BERT Model
LSTM - 128 units
Dropout - probability 50%
Dense - 256 units
Dropout - probability 50%
Dense *Softmax* - 101 units

BERT Architecture



Multimodal Text and Image Classification

Image + Text Classification Model: late fusion method

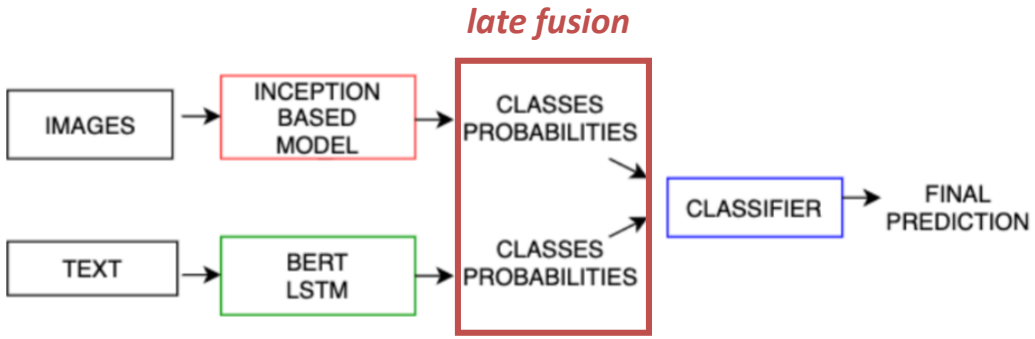


Fig. 1: Scheme of the implemented **late fusion** model with with a stacking approach.

TABLE III: Late fusion model architecture.

Input Image	Input Text
Inception V3	BERT Model
Average Pooling 2D (8 x 8)	LSTM - 128 units
Dropout - probability 40%	Dropout - probability 50%
Flatten	Dense - 256 units
	Dropout - probability 50%
Dense *Softmax* - 101 units	Dense *Softmax* - 101 units
Concatenate	
Dense - 256 units	
Dropout - probability 20%	
Dense - 128 units	
Dropout - probability 20%	
Dense *Softmax* - 101 units	

Multimodal Text and Image Classification

Image + Text Classification Model: early fusion method

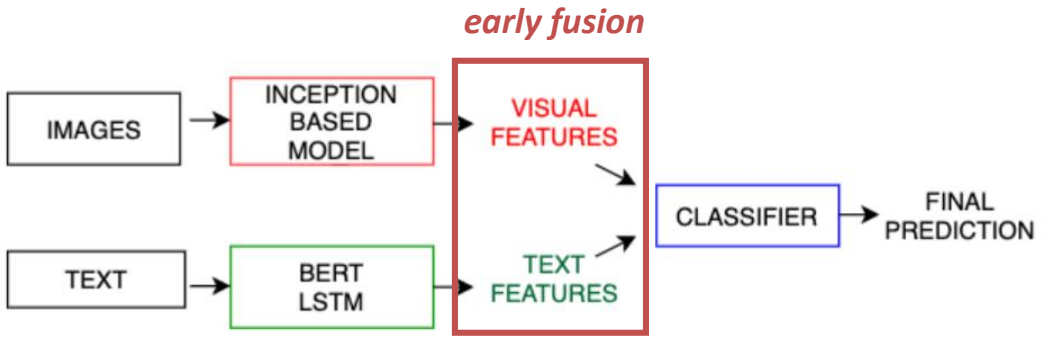


Fig. 2: Scheme of the implemented early fusion model with a stacking approach.

TABLE IV: Proposed model architecture.

Input Image	Input Text
Inception V3	
Average Pooling 2D (8 x 8)	
Dropout - probability 40%	
Flatten	BERT Model
Dense - 128 units	LSTM - 128 units
Concatenate	
Dense - 256 units	
Dense *Softmax* - 101 units	

Multimodal Text and Image Classification

Experimental Results

Table Ⅴ. Comparison of classification results available in literature on UPMC Food-101

- Metric: Accuracy

	Image	Text	Fusion
Wang et al. [1]	40.2	82.0	85.1
Kiela et al. [2]	56.7	**88.0**	90.8
Narayana et al. [3]	**73.8**	87.3	92.3
Proposed	71.7	84.6	**92.5**

Table Ⅵ. Comparison of the models for the classification of UPMC Food-101.

Test Accuracy on UPMC Food-101				
Images Model	Text Model	Late Fusion Model	Early Fusion Model (Paper)	Early Fusion Model (Our)
71.67	84.41	84.59	**92.50**	74.91

References

[1] X. Wang, D. Kumar, N. Thome, M. Cord, and F. Precioso, “Recipe recognition with large multimodal food dataset,” in 2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), pp. 1–6, IEEE, 2015.

[2] D. Kiela, E. Grave, A. Joulin, and T. Mikolov, “Efficient large-scale multi-modal classification,” arXiv preprint arXiv:1802.02892, 2018.

[3] P. Narayana, A. Pednekar, A. Krishnamoorthy, K. Sone, and S. Basu, “Huse: Hierarchical universal semantic embeddings,” arXiv preprint arXiv:1911.05978, 2019.

기존 서울대 데이터 실험 세팅

➢ **데이터 구성**

[기존 서울대 데이터]

- 177명의 환자 데이터 (전체 데이터)
- Neoplastic: 100명
- Non-neoplastic : 77명

➢ **Train & Test Set**

[기존 서울대 데이터]

- 2,888장의 데이터셋
- Train Set : 141명
- Test Set : 36명

➢ **Input Image**

- Outline, Inline, Polyp, 담낭벽+Polyp 이미지

➢ **모델 구성**

- 갑상선에서 사용한 모델을 담낭 용종 연구에 맞게 조정한 모델
- 데이터 불균형 문제를 해소하기 위한 Focal loss function 적용

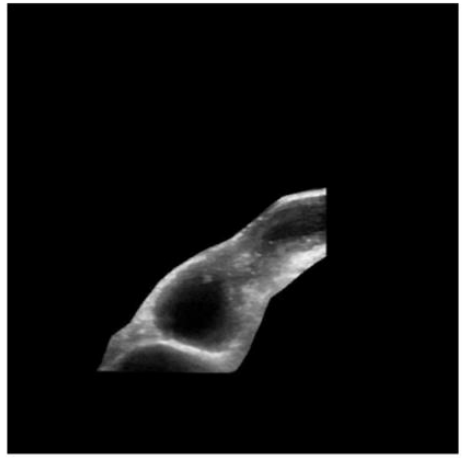
➢ **반복 실험 30회**

Activation Map Sample

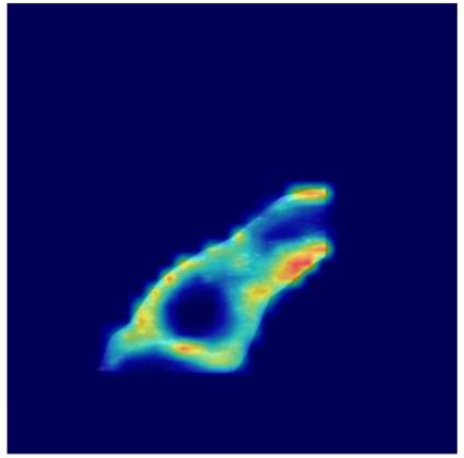
Activation Map: *Non-neoplastic*

Example: snuh_0097 (8)

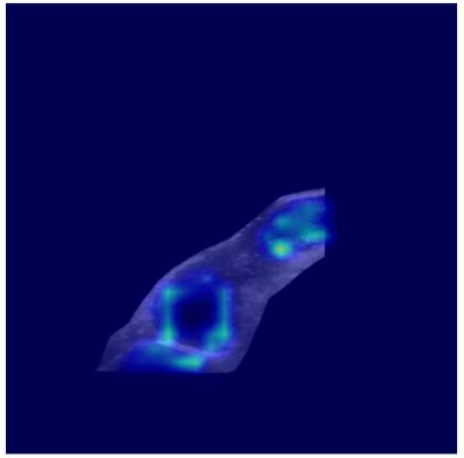
** All: outline x 0.25, inline x 0.25, polyp x 0.25, 담낭벽 + polyp: 0.25*



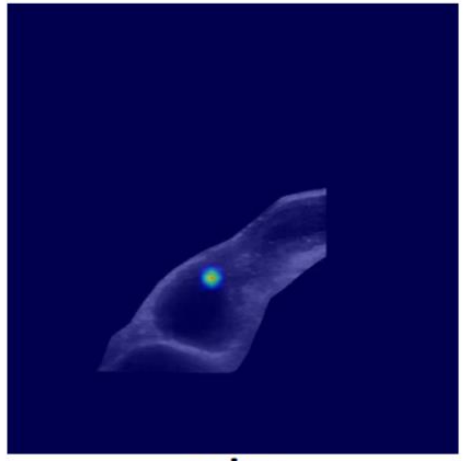
original



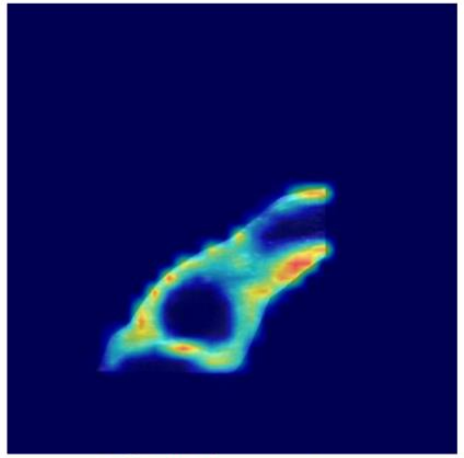
outline



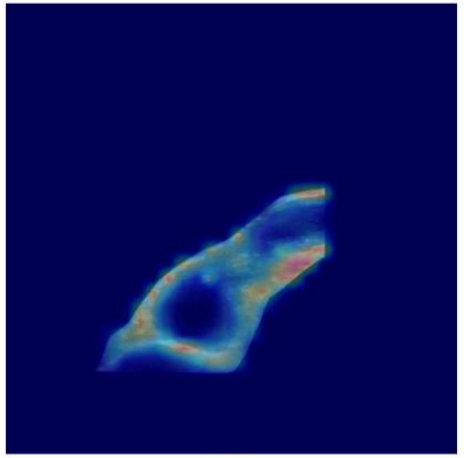
inline



polyp



담낭벽 +polyp



All

Activation Map Sample *[All]*