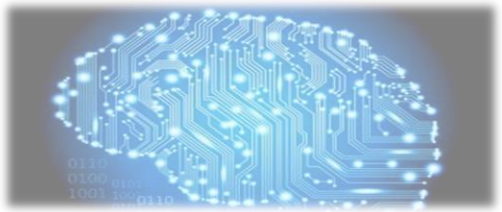
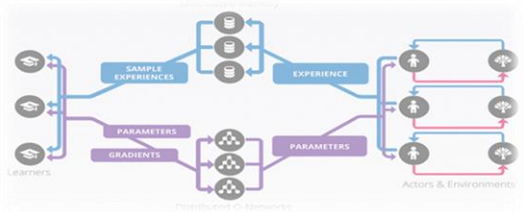



# Multimodal Text and Image Classification

AutoML 연구실
석사과정 문아성
22.01.11

# TO DO List

- Last week
  - [x] Bird Dataset: *"Fine-grained Image Classification via Combining Vision and Language"*
  - [x] 담낭용종 연구 실험 진행 (기존 서울대 데이터)
    - 기존 서울대 데이터로 adenoma, cancer, non-neoplastic 클래스 분류 실험 완료
  - [x] 졸업논문 – "An Efficient Neural Network based on Early Compression of Sparse CT Slice Images"

- This week
  - [x] Multimodal Text and Image Classification
  - [x] 담낭용종 연구 실험 진행 (기존 서울대 데이터)
    - Confusion matrix, ROC Curve 추가
  - [ ] 졸업논문 – "An Efficient Neural Network based on Early Compression of Sparse CT Slice Images"
    - Introduction 내용 추가 (medical field efficiency의 중요성)
    - Related work 의학적, CNN 내용 추가 (CNN Algorithms)

# Multimodal Text and Image Classification

- Deep Modular Co-Attention Networks for Visual Question Answering

*Yu, Zhou, et al. "Deep modular co-attention networks for visual question answering."*

*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.*

- **Self-Attention & Guided-Attention Units**
- **Modular Co-Attention (MCA)**
- **Modular Co-Attention Network**

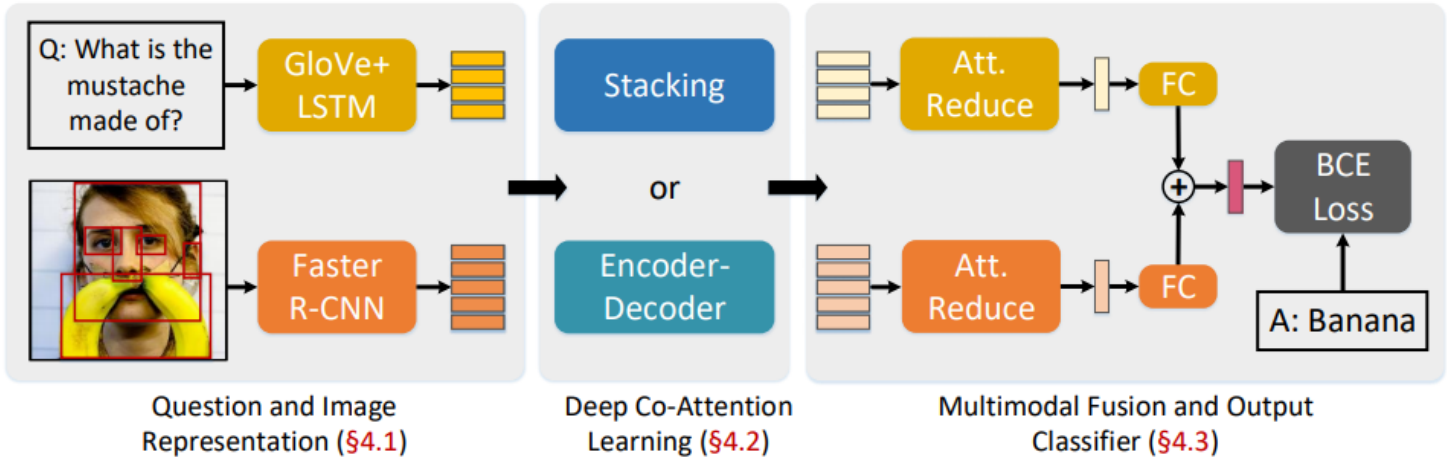



Figure 4: Overall flowchart of the deep Modular Co-Attention Networks (MCAN). In the Deep Co-attention Learning stage, we have two alternative strategies for deep co-attention learning, namely *stacking* and *encoder-decoder*.

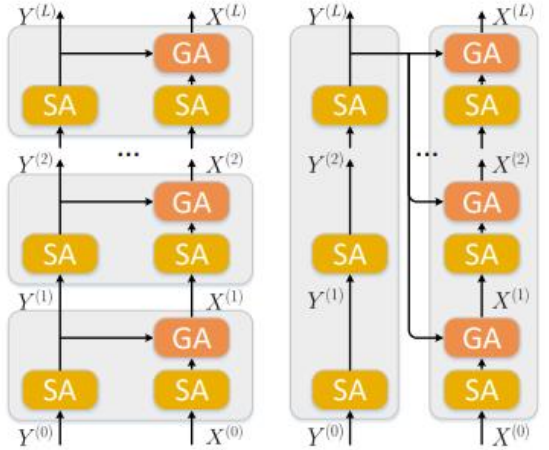



Figure 5: Two deep co-attention models based on a cascade of MCA layers (*e.g.*, SA(Y)-SGA(X,Y)).

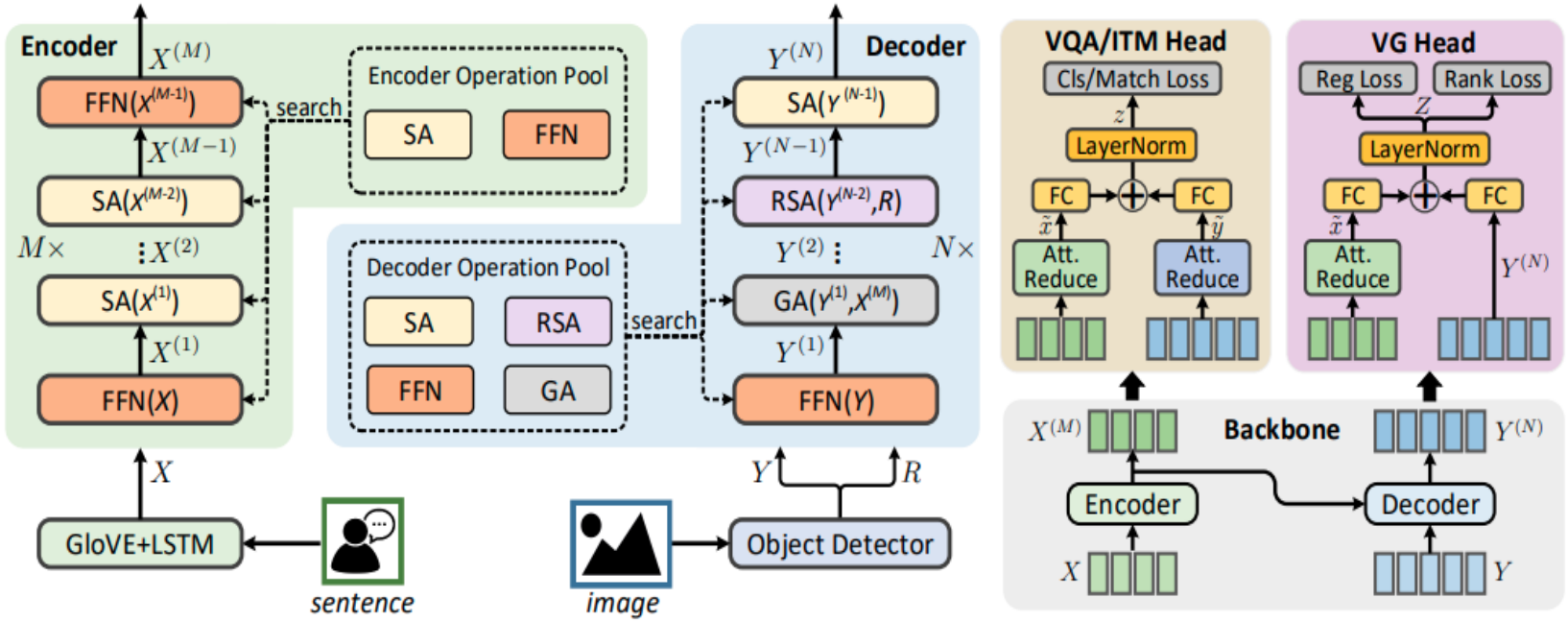
# Multimodal Text and Image Classification

- Deep Multimodal Neural Architecture Search

*Yu, Zhou, et al. "Deep Multimodal Neural Architecture Search."*

*Proceedings of the 28th ACM International Conference on Multimedia. 2020.*

- Unified Encoder-Decoder Backbone
  - *Self-attention operation*
  - *Guided-attention operation*
  - *Feed-forward network operation*
  - *Relation self-attention operation*




(a) Unified Encoder-Decoder Backbone (b) Task-specific Heads

**Figure 2: The flowchart of the MMnas framework, which consists of (a) unified encoder-decoder backbone and (b) task-specific heads on top the backbone for visual question answer (VQA), image-text matching (ITM), and visual grounding (VG).**

# Multimodal Text and Image Classification

- Are These Birds Similar: Learning Branched Networks for Fine-grained Representations

*Nawaz, Shah, et al. "Are these birds similar: Learning branched networks for fine-grained representations."*

*2019 International Conference on Image and Vision Computing New Zealand (IVCNZ). IEEE, 2019.*

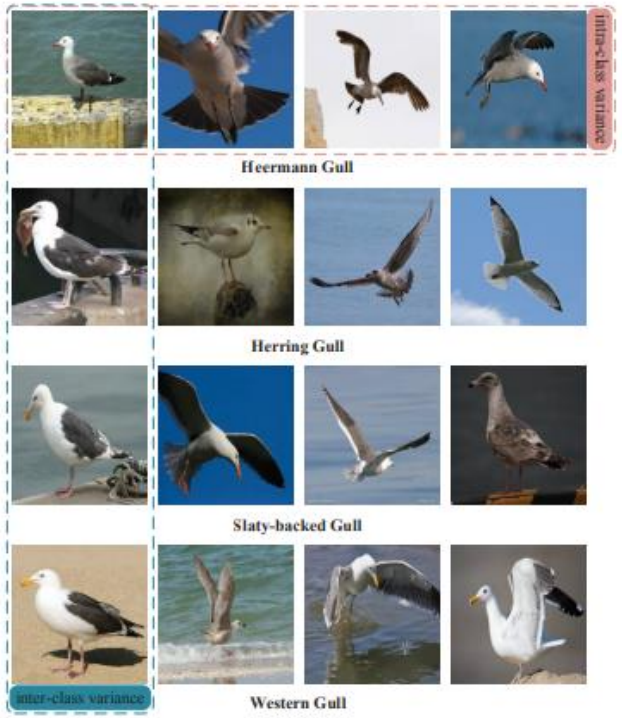



Fig. 1. Images extracted from CUB-200-2011. These examples indicate large intra-class variance and small inter-class variance, which makes the image classification task highly challenging.

| Class | Vision | Natural Language Description |
|---|---|---|
| Green Jay |  | – This bird has a black pointed bill, with a black and green breast.<br>– A colorful bird with shades of green, blue, and black. the beak is thick and pointed.<br>– Bird has a bluish cheek patch and a black eyebrow with a bluish spot.<br>... |
| Common Yellowthroat |  | – This bird has a brown crown as well as a yellow throat.<br>– This yellow breasted bird has a contrasting brown back and white flanks and thighs.<br>– A small multi colored bird with a white breast and yellow back feathers.<br>... |
| Red Eyed Vireo |  | – This beautiful gold and gray colored bird had a sharp pointed beak and black tail.<br>– This has a white underbelly with yellow and gray short feathers and a striped face.<br>– A bird with a white belly and a dark brown eyebrow.<br>... |

Fig. 3. Some random image-text pairs taken from the CUB-200-2011 dataset.

# Multimodal Text and Image Classification

## Model Architecture

- Vision Stream

*we exploited the visual information with the **NTS-NET** to extract features and perform the classification task.*

- Textual Stream

*we used only text descriptions and classified it with the **BERT model.***

- Two Branch Network

*we fused text and visual information with MLPs to perform multimodal classification.*

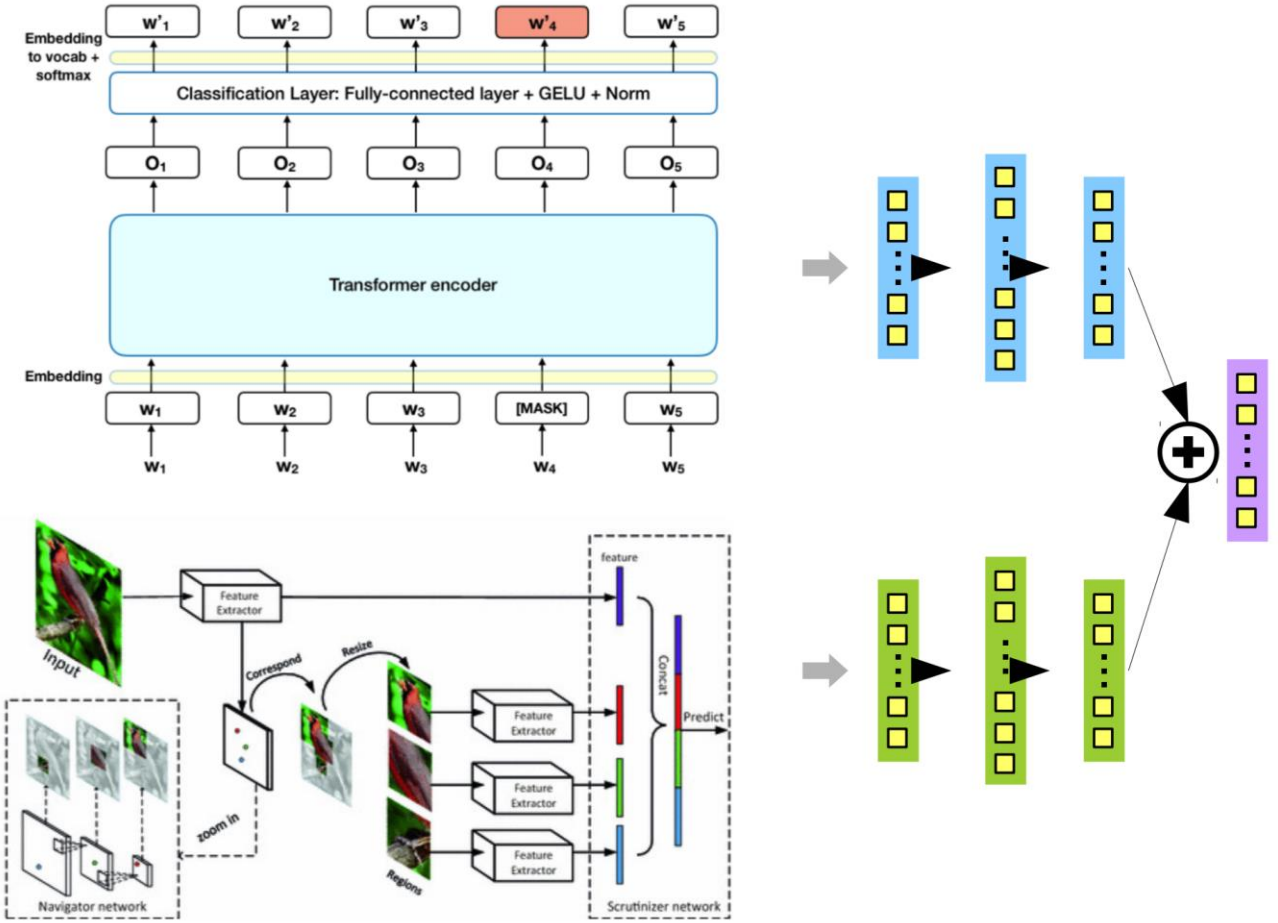



Fig. 2. The proposed approach extracts image and text representation from NTS-NET [9] and BERT [10] models. Finally, representations is fused after passing through a MLP.

# 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.*


Examples of images in the UPMC Food-101 Dataset


Text(recipe) data sample: Apple Pie

# Multimodal Text and Image Classification

## Model Architecture

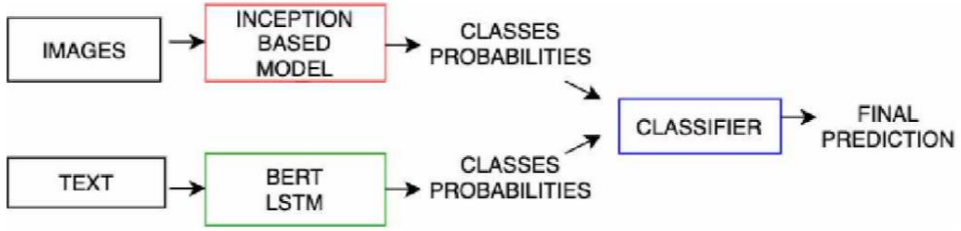



Fig. 1: Scheme of the implemented **late fusion** model with with a stacking approach.

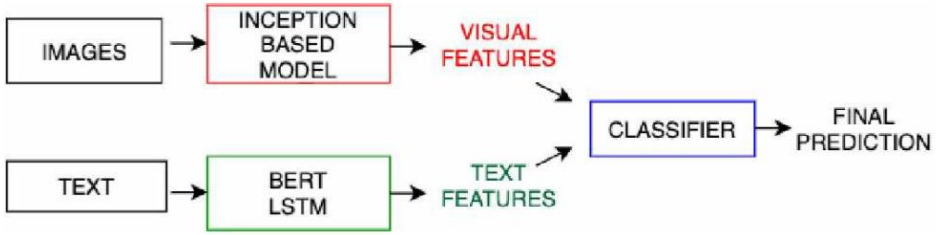



Fig. 2: Scheme of the implemented early fusion model with a stacking approach.

| 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 | |

TABLE IV: Proposed model architecture.

# 기존 서울대 데이터 실험 세팅

➢ **데이터 구성**

[기존 서울대 데이터]

- 177명의 환자 데이터 (전체 데이터)
- Cancer: 31명
- Adenoma: 69명
- Non-neoplastic : 77명

➢ **Train & Test Set**

[기존 서울대 데이터]

- 2,888장의 데이터셋
- Train Set : 141명(2,256장)
- Test Set : 36명(632장)

➢ **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**

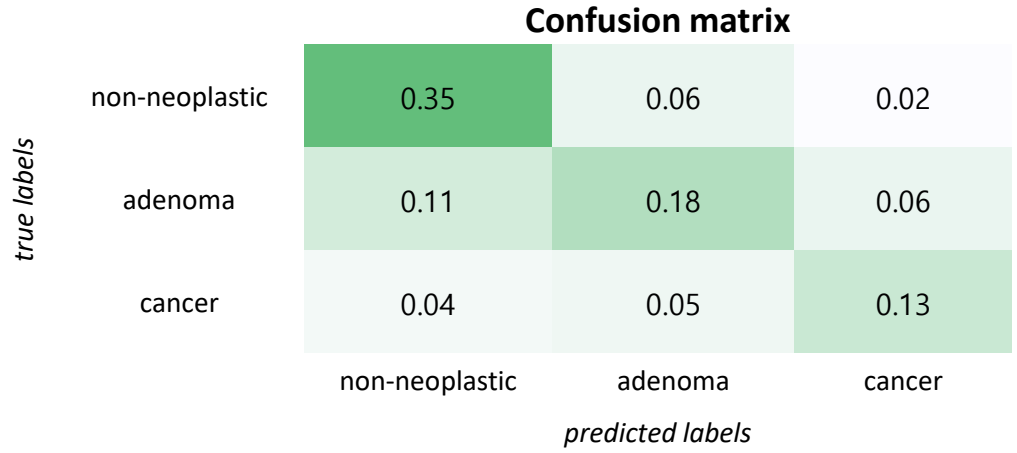
# 서울대 데이터 실험 결과

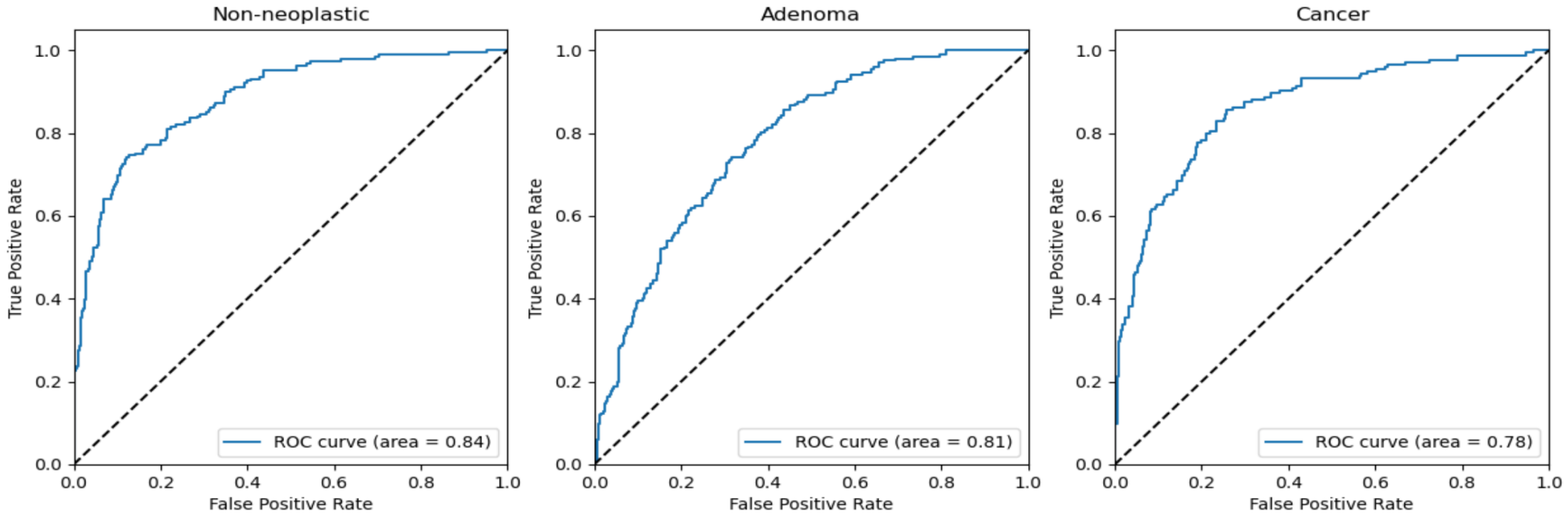
Table. The main classification metrics accuracy, AUC, sensitivity and specificity on a per-class basis.

| | Accuracy | AUC | Sensitivity | Specificity |
|---|---|---|---|---|
| **non-neoplastic** | **0.770** | 0.836 | 0.825 | 0.720 |
| **adenoma** | **0.715** | 0.805 | 0.519 | 0.833 |
| **cancer** | **0.821** | 0.784 | 0.586 | 0.897 |

AUC, area under the curve

# 서울대 데이터 실험 결과


Confusion matrix
non-neoplastic
adenoma
cancer
true labels
0.35
0.06
0.02
0.11
0.18
0.06
0.04
0.05
0.13
non-neoplastic
adenoma
cancer
predicted labels


Non-neoplastic
Adenoma
Cancer
True Positive Rate
False Positive Rate
ROC curve (area = 0.84)
ROC curve (area = 0.81)
ROC curve (area = 0.78)