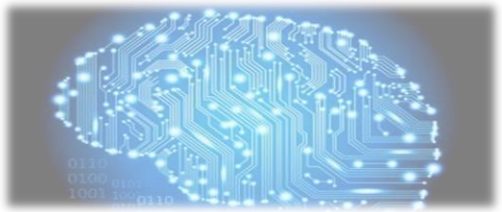
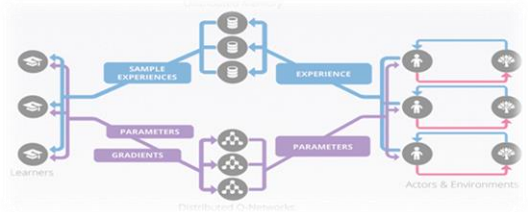



Fine-grained Image Classification via Combining Vision and Language

AutoML 연구실
석사과정 문아성

21.12.28

TO DO List

- Last week
  - ☑ ~~Multimodal Datasets Paper Search~~
  - ☑ 담낭용종 연구 실험 진행 (기존 서울대 데이터, 신규 서울대 데이터, 가톨릭 센터 데이터)
    - 각 데이터셋으로 학습 후 다른 데이터셋으로 테스트 완료
    - 각 데이터셋으로 학습 후 다른 데이터셋으로 전이 학습 실험 완료
  - ☑ 졸업논문 – "An Efficient Neural Network based on Early Compression of Sparse CT Slice Images"
- This week
  - ☑ Bird Dataset: *"Fine-grained Image Classification via Combining Vision and Language"*
  - ☑ 담낭용종 연구 실험 진행 (기존 서울대 데이터)
    - 기존 서울대 데이터로 adenoma, cancer, non-neoplastic 클래스 분류 실험 완료
  - ☐ 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)

Fine-grained Image Classification via Combining Vision and Language

- Fine-grained Image Classification via Combining Vision and Language

He, Xiangteng, and Yuxin Peng. "Fine-grained image classification via combining vision and language."

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.

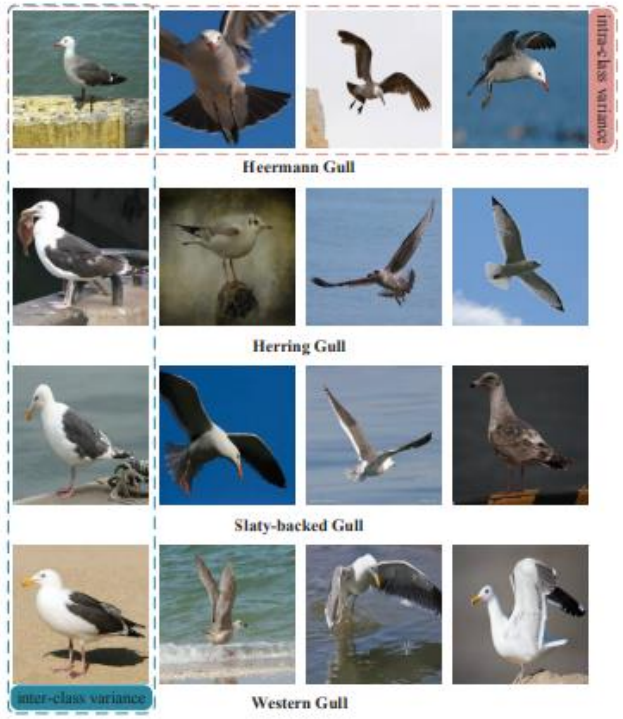


Figure 1. Examples from CUB-200-2011. Note that it is a technically challenging task even for humans to categorize them due to large intra-class variance and small inter-class variance.

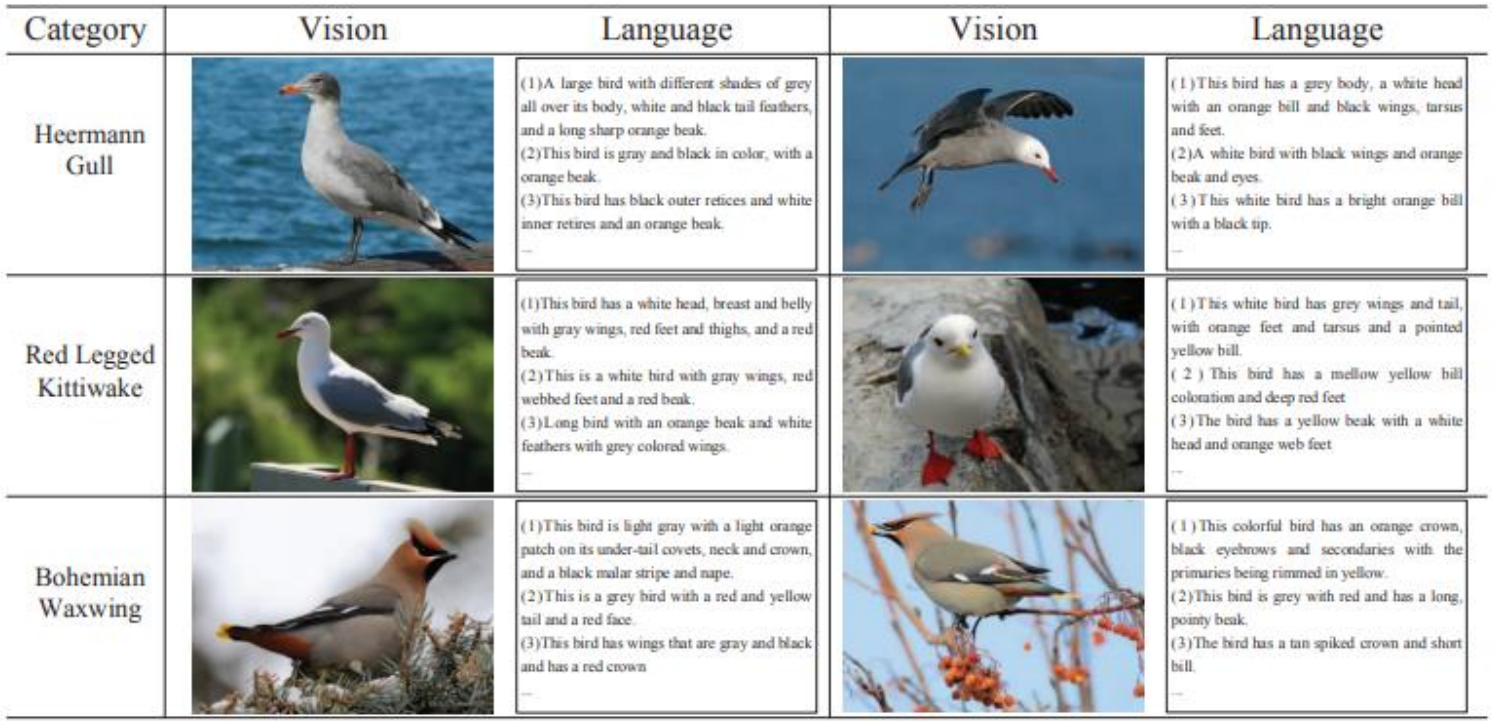


Figure 4. Sample natural language descriptions of CUB-200-2011.

Fine-grained Image Classification via Combining Vision and Language

Model Architecture

- Object localization
- Vision Steam
- Language Steam

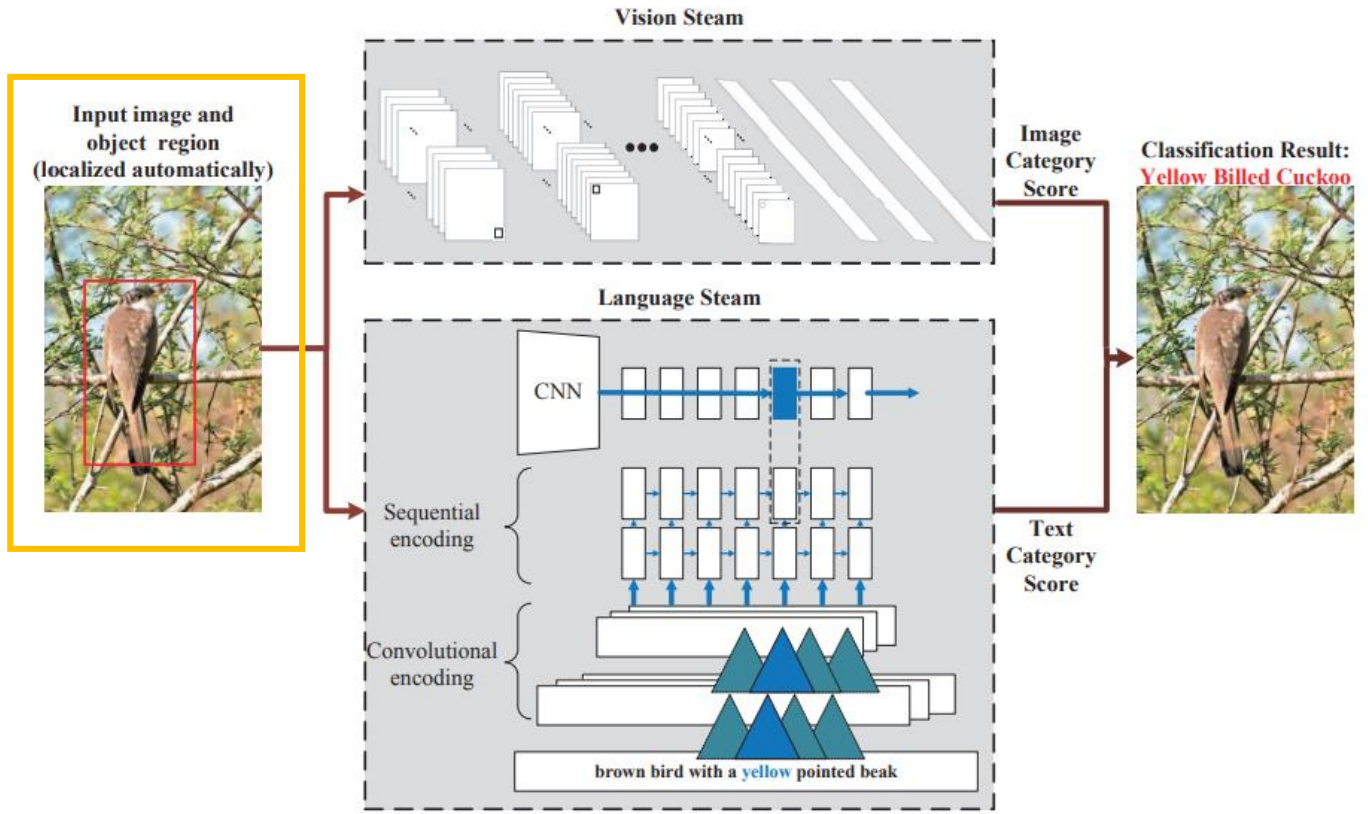
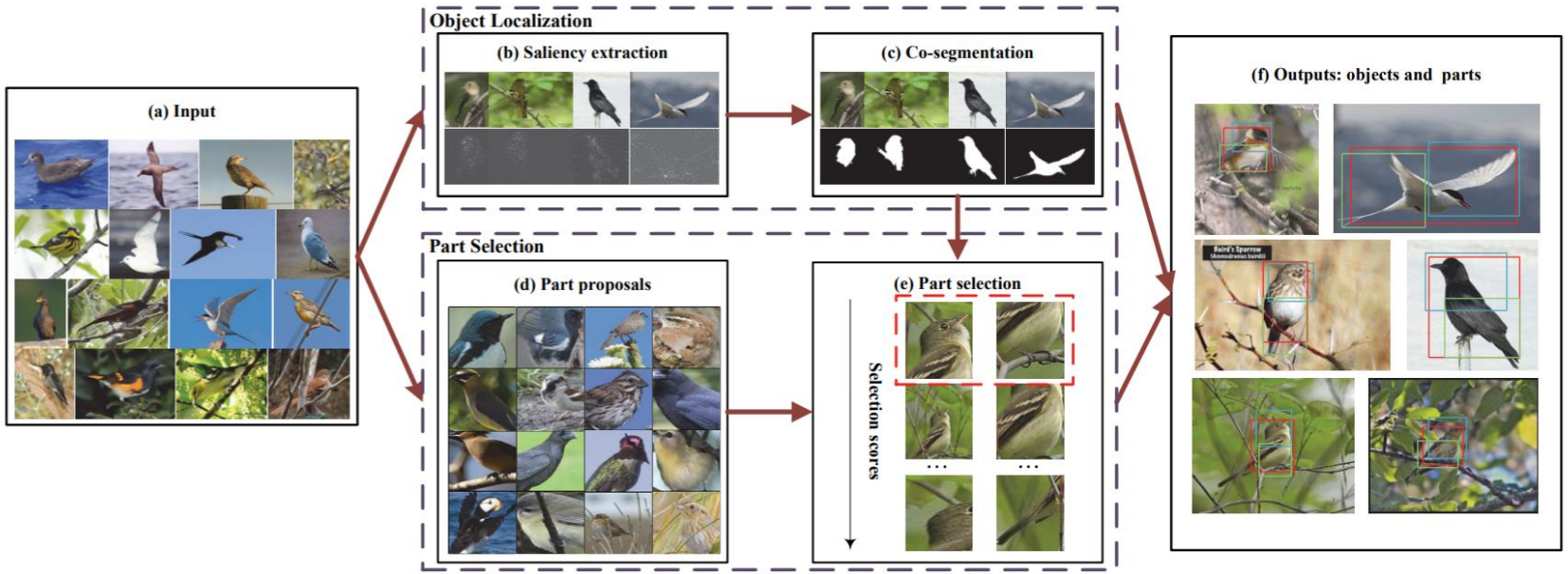


Figure 2. Overview of our CVL approach. The two-stream model conducts on the original images and their object localizations. One learns the deep representations directly from the vision information. The other learns the salient visual aspects for distinguishing sub-categories via jointly modeling vision and language. The classification results of the two streams are merged in later phase to combine the advantages of vision and language.

Fine-grained Image Classification via Combining Vision and Language

- Object localization
  - *Saliency extraction*
  - *Co-segmentation*
- Part Selection
  - *Part proposals*
  - *Part selection*



Method	Accuracy (%)
Language+ft+box	**81.81**
Language+ft	77.80
Language	50.54

Table 3. Effect of fine-tuning and object localization for fine-grained image classification. "ft" indicates fine-tuning is applied, and "box" indicates object localization is applied.

Fine-grained Image Classification via Combining Vision and Language

- Jointly Model Vision and Language: Vision Stream

  - *Pre-trained dataset: ImageNet*
  - *Fine-tuned dataset: CUB-200-2011*

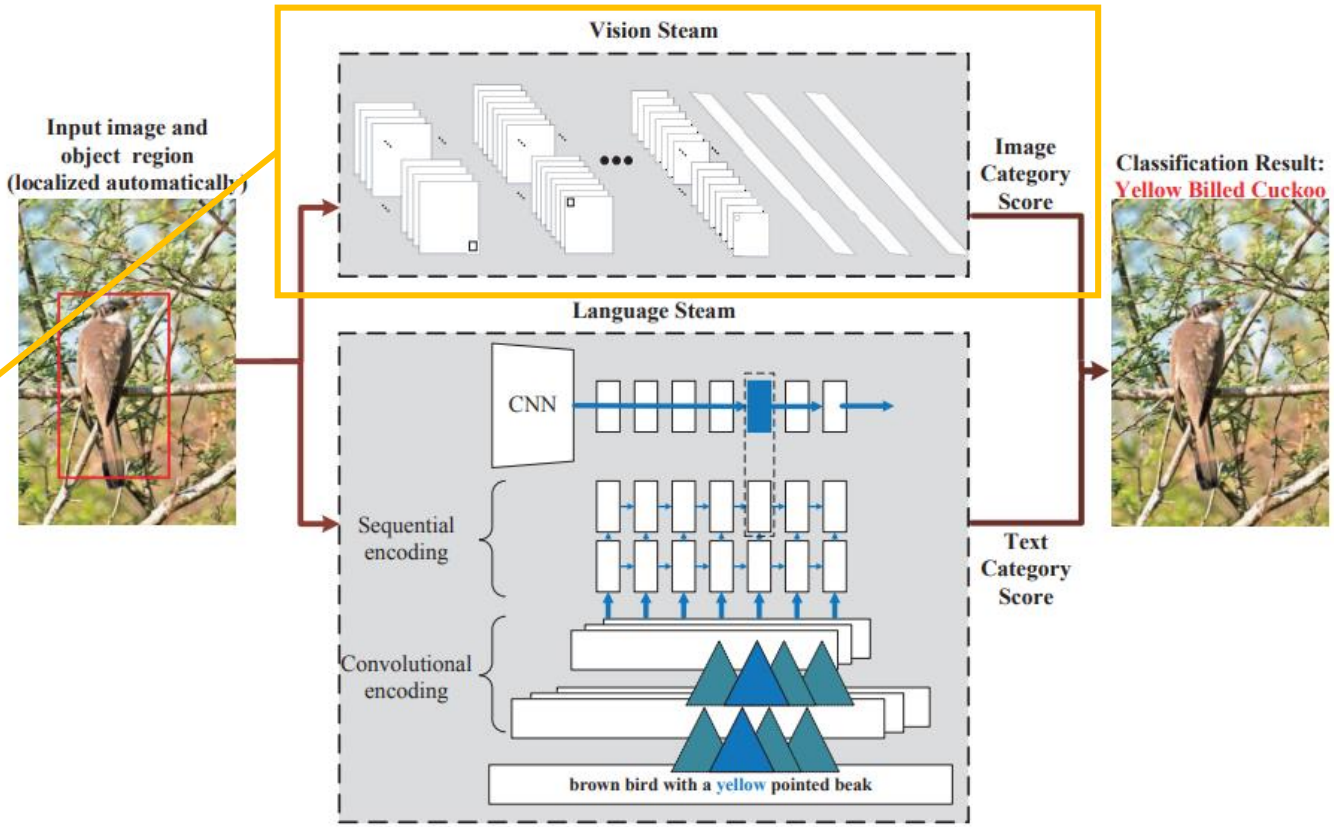


Figure 2. Overview of our CVL approach. The two-stream model conducts on the original images and their object localizations. One learns the deep representations directly from the vision information. The other learns the salient visual aspects for distinguishing sub-categories via jointly modeling vision and language. The classification results of the two streams are merged in later phase to combine the advantages of vision and language.

Fine-grained Image Classification via Combining Vision and Language

- Jointly Model Vision and Language: Language Stream

$$\frac{1}{N}\sum_{n=1}^{N}\Delta(y_n, f_v(v_n)) + \Delta(y_n, f_t(t_n)) \quad (1)$$

$$f_v(v) = arg\max_{y\in Y}\mathbb{E}_{t\sim T(y)}[F(v,t)] \quad (2)$$

$$f_t(t) = arg\max_{y\in Y}\mathbb{E}_{v\sim V(y)}[F(v,t)] \quad (3)$$

$$F(v,t) = \theta(v)^T\phi(t) \quad (4)$$

$$\phi(t) = \frac{1}{L}\sum_{i=1}^{L} h_i \quad (5)$$

$$f(I) = f_v(v) + \beta * f_t(t) \quad (6)$$

$D = (v_n, t_n, y_n)$

$v \in V, \quad t \in T, \quad y \in Y$

$f_v: V \to Y, \quad f_t: T \to Y$

θ: encoder function

h_i: hidden embedding vector

L: sequence length

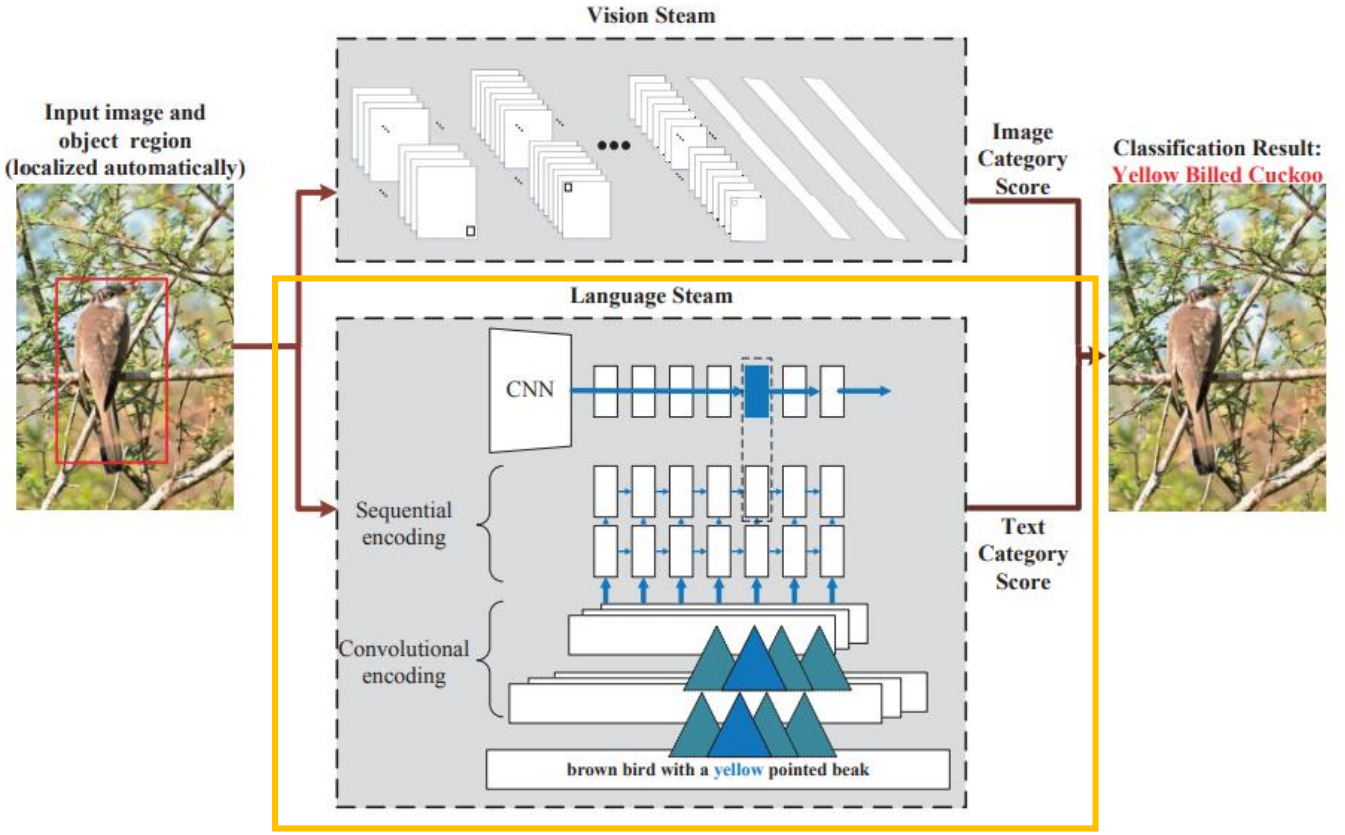


Figure 2. Overview of our CVL approach. The two-stream model conducts on the original images and their object localizations. One learns the deep representations directly from the vision information. The other learns the salient visual aspects for distinguishing sub-categories via jointly modeling vision and language. The classification results of the two streams are merged in later phase to combine the advantages of vision and language.

Fine-grained Image Classification via Combining Vision and Language

Experimental Results

Method	Train Annotation		Test Annotation		Accuracy (%)
	Bbox	Parts	Bbox	Parts	
Our CVL approach					**85.55**
PD [5]					84.54
Spatial Transformer [37]					84.10
Bilinear-CNN [38]					84.10
NAC [19]					81.01
TL Atten [2]					77.90
VGG-BGLm [39]					75.90
PG Alignment [18]	√		√		82.80
Triplet-A (64) [40]	√		√		80.70
VGG-BGLm [39]	√		√		80.40
Part-based R-CNN [3]	√	√			73.50
SPDA-CNN [41]	√	√	√		85.14
Part-based R-CNN [3]	√	√	√	√	76.37
POOF [14]	√	√	√	√	73.30
GPP [15]	√	√	√	√	66.35

Table 1. Comparisons with state-of-the-art methods on CUB-200-2011, sorted by amount of annotation used. "Our CVL" indicates our full method combining vision and language. "Bbox" indicates the object annotation (i.e. bounding box of object) provided by the dataset, and "Parts" indicates the parts annotations (i.e. parts locations). "√" indicates that one of bounding box and part locations is used in training or testing stage. Since the exact amount of annotation used varies from method to method, we defer to the original sources for details.

Method	Accuracy (%)
our CVL approach (Language-stream+Vision-stream)	**85.55**
Language-stream	81.81
Vision-stream	82.98
Original	76.17

Table 2. Effects of different variants of our method on CUB-200-2011. "Language-stream" refers to the classification result of the language stream, "Vision-stream" refers to the classification result of the vision stream, "CVL" refers to our CVL approach combining vision and language, and "Original" refers to the classification result that only use the original image for prediction.

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 . Metrics with 95% confidence interval of five Images.

Model	Metrics	AUC Mean	95% CI	
			Lower	Upper
기존 서울대 데이터 [neoplastic non-neoplastic]	AUC	0.968	0.932	1.003
	Accuracy	0.917	0.850	0.984
	Sensitivity	0.889	0.754	1.023
	Specificity	0.954	0.893	1.016
기존 서울대 데이터 [cancer adenoma non-neoplastic]	AUC	0.901	0.844	0.995
	Accuracy	0.752	0.636	0.943
	Sensitivity	0.718	0.441	0.903
	Specificity	0.715	0.428	0.915

AUC, area under the curve; CI, confidence interval;