Multimodal Segmentation

조 유 빈

Vision & Display Systems Lab.

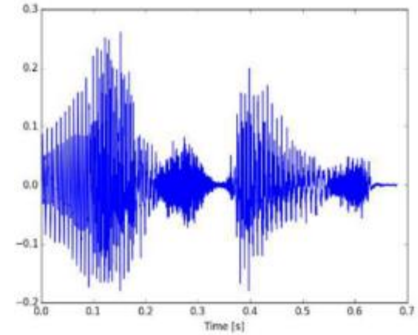
Dept. of Artificial Intelligence, Sogang University

Outline

- Introduction
  - Multi-modal learning
- Referring Expression Segmentation
  - Vision-Language Transformer and Query Generation for Referring Segmentation (ICCV 2021)
    - Attention Is All You Need (NIPS 2017)
- Zero-shot Segmentation
  - Language-Driven Semantic Segmentation (ICLR 2022)
- Conclusion

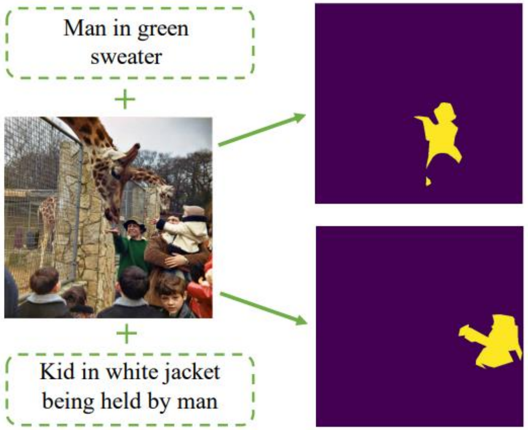
Introduction

- Multi-modal
  - 두가지 이상의 modality의 결합
  - Text – Image / Image – Audio / Audio – Image – Text
- Multi-modal learning
  - 인간의 인지적 학습법을 모방하여 다양한 형태(modality)의 데이터로 학습하는 방법
  - 다양한 모달 조합을 통해 task 확장 가능
  - Uni-modal보다 풍부한 정보 획득



Referring Expression Segmentation

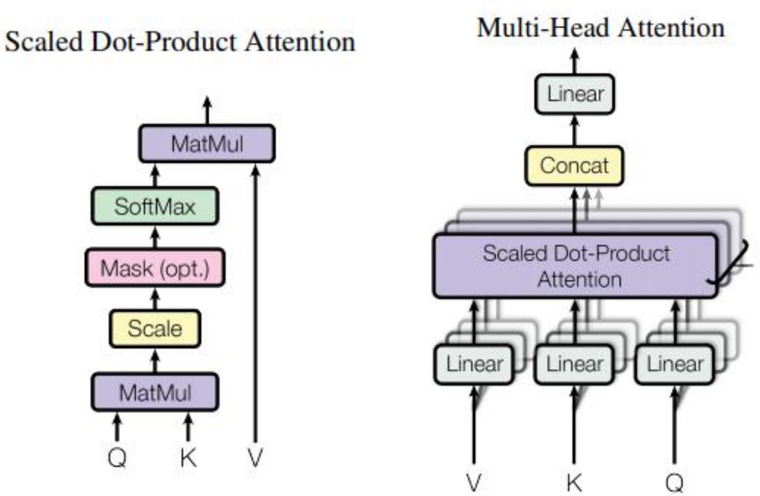
- What is referring expression segmentation?
  - Target object의 특성에 대한 language expression이 주어지면 이미지 내에서 해당 object만을 segmentation
  - Challenging points
    - Language expression은 target 객체와 다른 객체들과의 relationship에 대한 describing을 포함
      - 이미지의 holistic understanding을 위한 global context 정보 추출 필요
    - 이미지 내의 다양한 객체와 제약 없는 언어 표현으로 인한 randomness
      - Network의 robustness 필요



[1] Google Research, "Attention Is All You Need" (NIPS 2017)

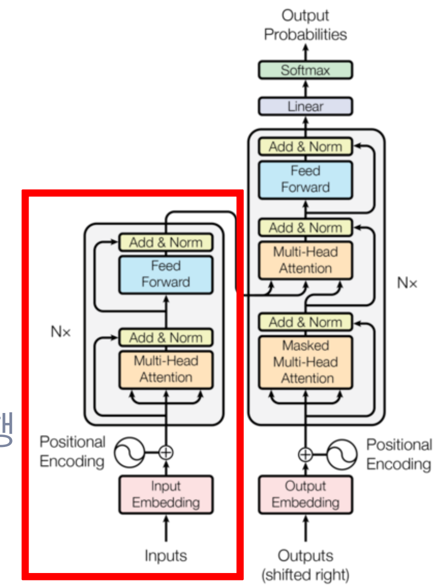
Referring Expression Segmentation

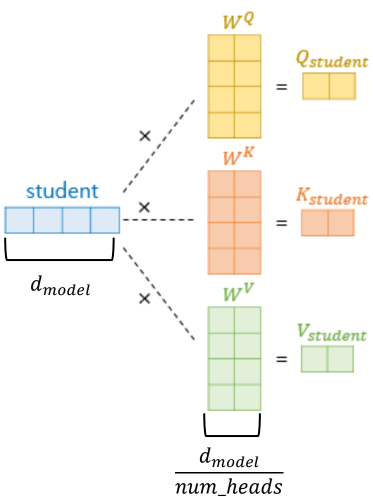
- Transformer [1]
  - Self-attention
    - 단일 sequence 내의 서로 다른 요소들을 관련시켜 한 position의 representation을 계산
  - Why self-attention
    - 병렬적으로 동시에 연산 가능
    - 멀리 떨어진 원소들 간의 path length 감소
      - Long-term dependency problem 해결
      - Global dependency 학습

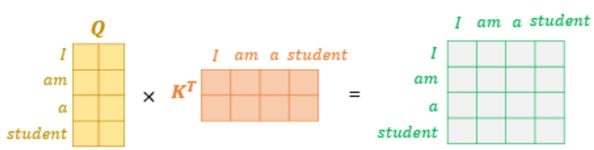


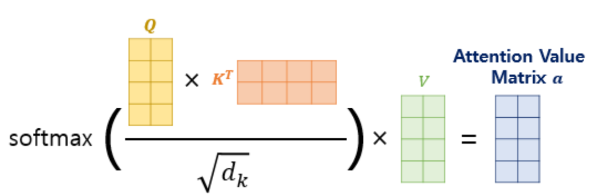
Referring Expression Segmentation

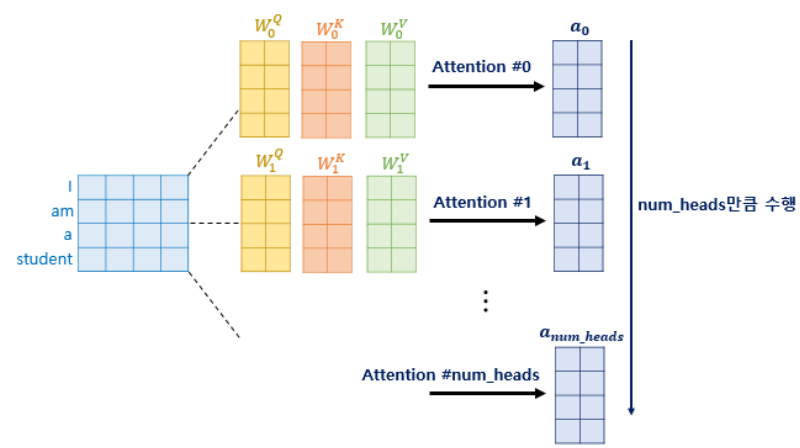
- Transformer [1]
  - Encoder
    - Multi-head self-attention
      - Self-attention : Q, K, V의 출처가 같음 (encoder vector)
      - Multi-head : 벡터의 차원을 축소시키고 attention을 병렬적으로 수행
        - ✓다른 관점에서 정보들을 수집
        - ✓W^Q, W^K, W^V는 각 attention head마다 값이 다름







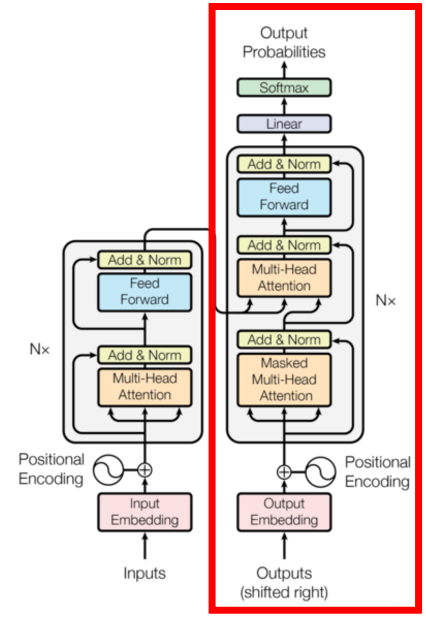


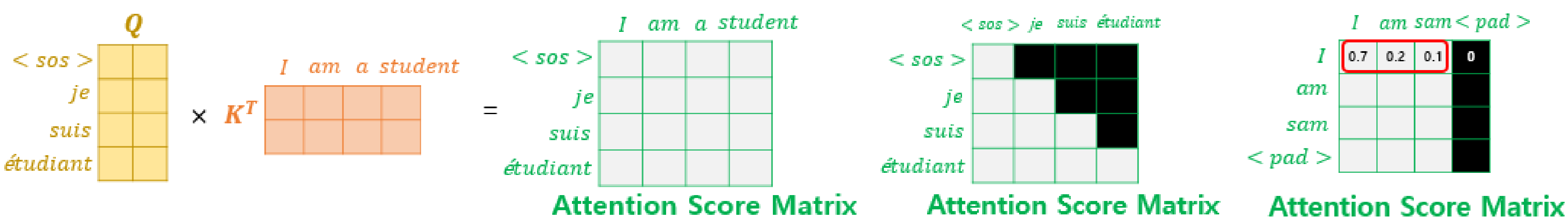


[1] Google Research, “Attention Is All You Need” (NIPS 2017)

Referring Expression Segmentation

- Transformer [1]
  - Decoder
    - Masked multi-head self-attention
      - Self-attention : Q, K, V의 출처가 같음 (decoder vector)
      - 일부 원소는 매우 작은 음수 값을 곱해 masking
        - ✓실질적인 의미를 가진 단어가 아닌 <pad>인 경우
        - ✓현재 시점보다 미래에 있는 단어인 경우
    - Multi-head attention (non self-attention)
      - Non self-attention : Q (decoder vector) / K, V (encoder vector)
      - Decoder 출력을 위해 encoder의 어떤 정보를 참고하면 좋을지 attention 수행



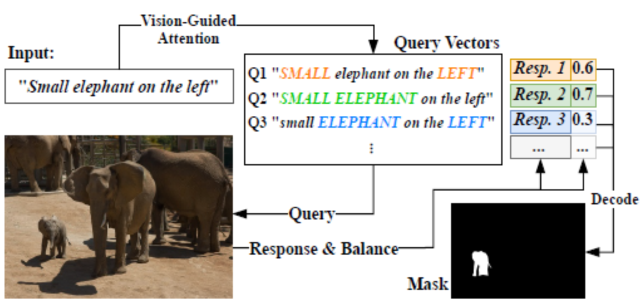


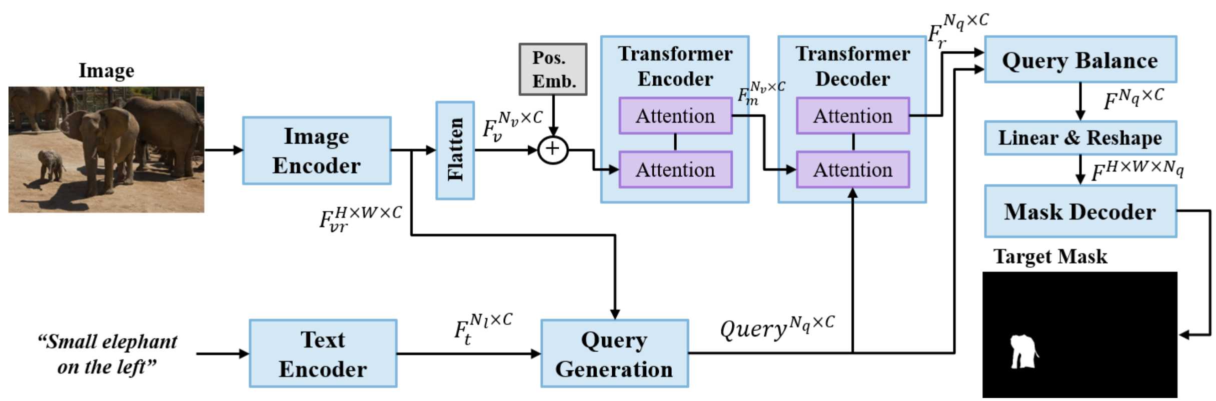
< Look-ahead mask >

< Padding mask >

Referring Expression Segmentation

- VLT [1]
  - NLP transformer [2] 의 encoder, decoder 구조를 적용
    - Build deep interactions among multi-modal information (vision-language)
    - Long-range dependencies modeling
      - Bring efficiencies to information interactions among pixels/words in a distance
      - Understand the global context of the image

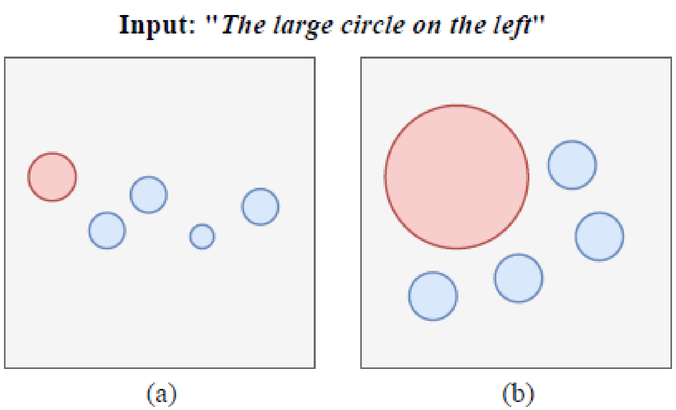




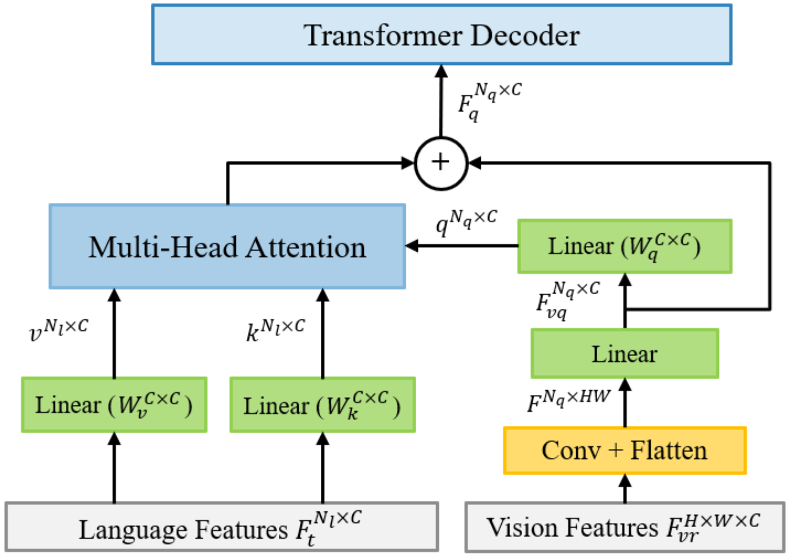
< Overall architecture >

Referring Expression Segmentation

- VLT [1]
  - Query Generation Module (QGM)
    - Comprehend the language expression from multiple aspects incorporating the image
      - A same sentence may have different understanding perspectives and emphasis
    - Learn different aspects of information & enhance the robustness of the queries
      - Extract the key information & address high randomness
      - Different queries emphasize different words

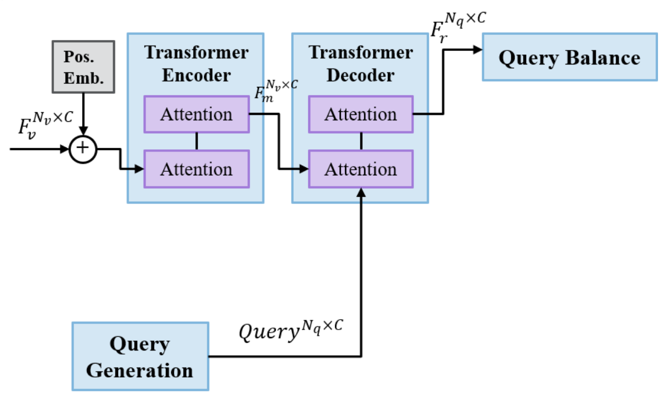


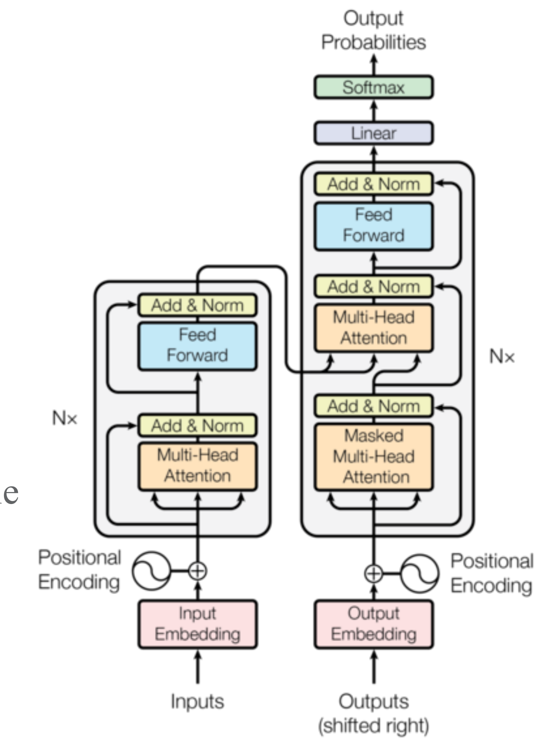
< Example of one sentence having different emphasis >



Referring Expression Segmentation

- VLT [1]
  - Transformer Encoder
    - Deriving the memory features about vision information
      - Extract the global context of vision information
  - Transformer Decoder
    - Outputs the response features corresponding to each query vector
      - Query the image with language vector
    - Q : N_q language query vectors produced by Query Generation module
    - K, V : vision memory features F_m of the transformer encoder

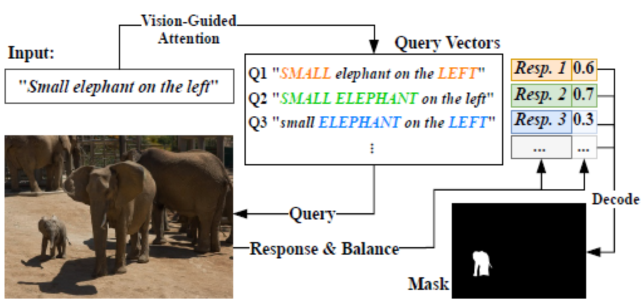


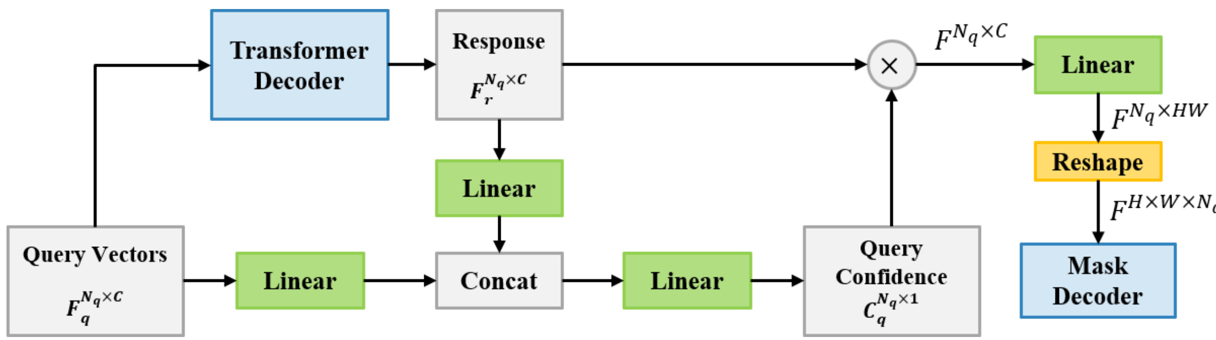


[1] Henghui Ding, Chang Liu, et al. "Vision-Language Transformer and Query Generation for Referring Segmentation" (ICCV 2021)

Referring Expression Segmentation

- VLT [1]
  - Query Balance Module (QBM)
    - Find the better comprehension ways to the image and language
      - Queries that provide better comprehensions are spotlighted
    - Balance the influence of different queries to the final output
    - Confidence shows how much the query fits the context of its prediction and controls the influence of its response to the mask decoding





< Query Balance Module >

Referring Expression Segmentation

- VLT [1]
  - Ablation study
    - Effectiveness of the transformer module (Table 1)
      - Compare the performance and parameter size of transformer with regular conv-nets
    - Effectiveness of the Query Generation Module (Table 2)
      - QGM effectively understands the sentence and generates valid attended language features guided by vision information
      - F_t : directly send the language features into the transformer decoder as the query
      - Learnt : 16 query vectors are learned during training and fixed during inference

Table 1

Type	#params	IoU	Pr@0.5	Pr@0.6	Pr@0.7	Pr@0.8	Pr@0.9
7 Conv Layers	~ 16.6M	44.28	49.54	42.16	35.24	25.98	10.47
Transformer	~ 17.5M	49.36	55.84	50.79	41.68	29.96	10.76

Table 2

No.	Method	IoU	Pr@0.5	Pr@0.6	Pr@0.7	Pr@0.8	Pr@0.9
1	F_t	45.05	52.69	46.08	36.20	20.97	3.42
2	Learnt	42.99	49.85	42.38	31.52	17.14	2.41
3	Ours	49.36	55.84	50.79	41.68	29.96	10.76

Referring Expression Segmentation

- VLT [1]
  - Ablation study
    - Performance gain by increasing the query number N_q (Figure 1 & Table 3)
      - Multiple queries are desired for the transformer network
      - Multiple queries generated by QGM represent different aspects of information
    - Effectiveness of the Query Balance Module (Figure 1 & Table 3)

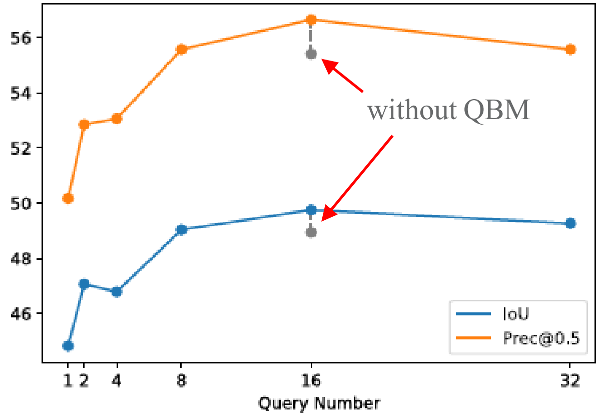


Figure 1

Table 3. Influence of Query Numbers. *: without Query Balance Module

N_q	IoU	Pr@0.5	Pr@0.6	Pr@0.7	Pr@0.8	Pr@0.9
1	44.83	50.17	43.94	34.75	21.64	4.66
2	47.07	52.85	47.31	39.66	28.90	8.30
4	46.79	53.06	47.54	40.38	28.23	8.92
8	49.04	55.57	50.58	44.24	32.99	12.62
16	**49.36**	**55.84**	50.79	41.68	29.96	10.76
32	49.27	55.57	50.48	44.43	33.87	12.50
16*	48.94	55.41	50.32	43.84	32.56	12.99

Table 3

Referring Expression Segmentation

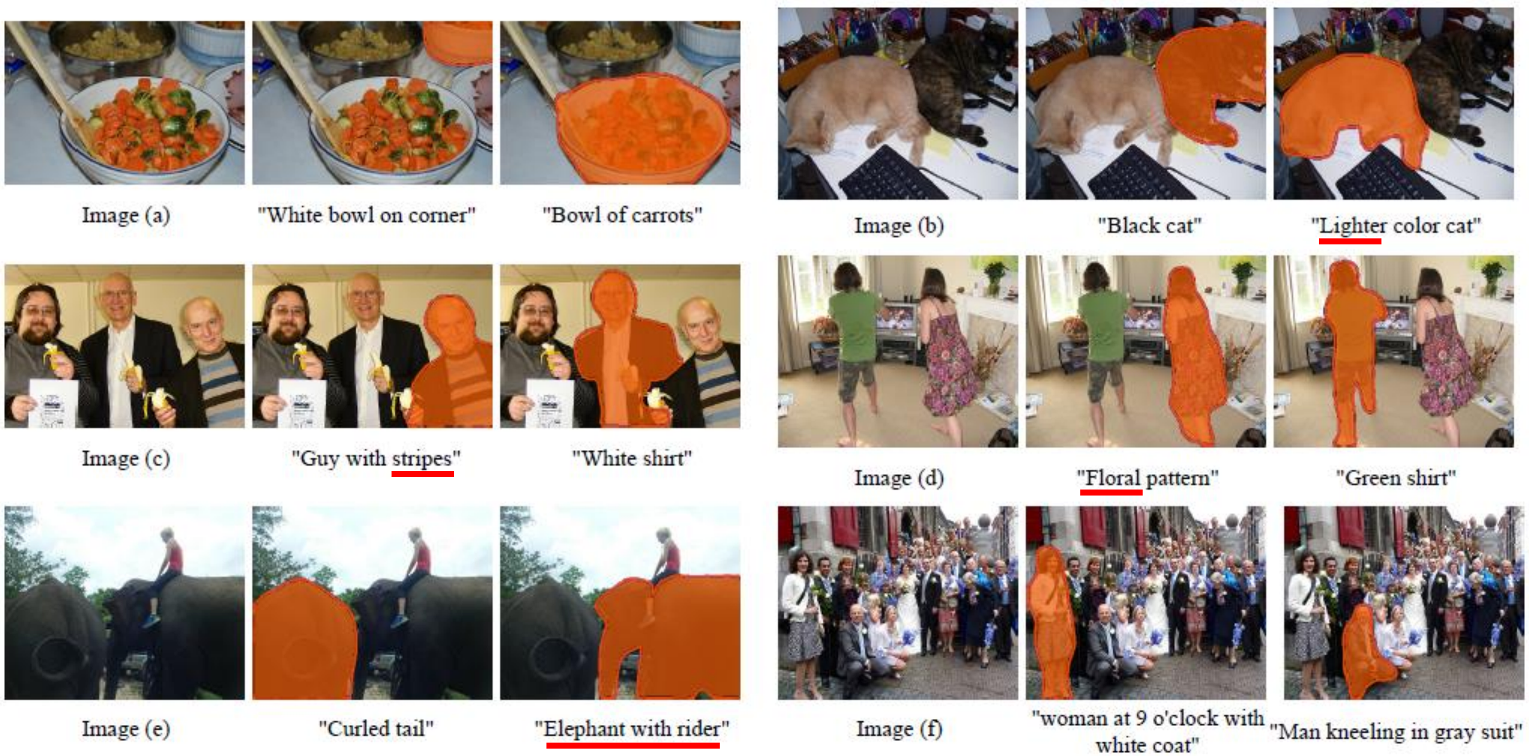
- VLT [1]
  - Quantitative results
    - Proposed method has good abilities on hard cases and long expressions
      - Long and complex sentences usually contain more information and more emphasis
        - ✓QGM and QBM can detect multiple emphasis and find the more informative ones

	RefCOCO			RefCOCO+			G-Ref		
	val	test A	test B	val	test A	test B	val (U)	test (U)	val(G)
DMN [22]	49.78	54.83	45.13	38.88	44.22	32.29	-	-	36.76
RRN [15]	55.33	57.26	53.93	39.75	42.15	36.11	-	-	36.45
MAttNet [29]	56.51	62.37	51.70	46.67	52.39	40.08	47.64	48.61	-
CMSA [28]	58.32	60.61	55.09	43.76	47.60	37.89	-	-	39.98
BRINet [11]	60.98	62.99	59.21	48.17	52.32	42.11	-	-	48.04
CMPC [12]	61.36	64.53	59.64	49.56	53.44	43.23	-	-	39.98
LSCM [13]	61.47	64.99	59.55	49.34	53.12	43.50	-	-	48.05
MCN [19]	62.44	64.20	59.71	50.62	54.99	44.69	49.22	49.40	-
CGAN [18]	64.86	68.04	62.07	51.03	55.51	44.06	51.01	51.69	46.54
VLT (ours)	**65.65**	**68.29**	**62.73**	**55.50**	**59.20**	**49.36**	**52.99**	**56.65**	**49.76**
Prec@0.5	76.20	80.31	71.44	64.19	68.40	55.84	61.03	60.24	56.65

[1] Henghui Ding, Chang Liu, et al. "Vision-Language Transformer and Query Generation for Referring Segmentation" (ICCV 2021)

Referring Expression Segmentation

- VLT [1]
  - Qualitative results

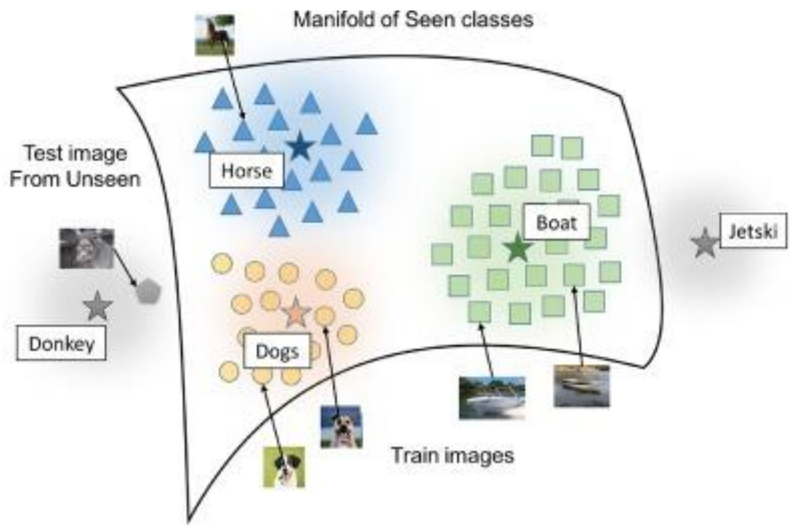


expression describing the interaction between objects

contain multiple aspects of information

Zero-shot Segmentation

- What is zero-shot method?
  - 추가적인 학습 과정 없이도 unseen class를 예측하는 것
    - 일반적인 딥러닝은 training 과정에서 학습한 제한된 class만으로 예측
  - Novel class 학습을 위한 추가적인 annotation 필요 없음

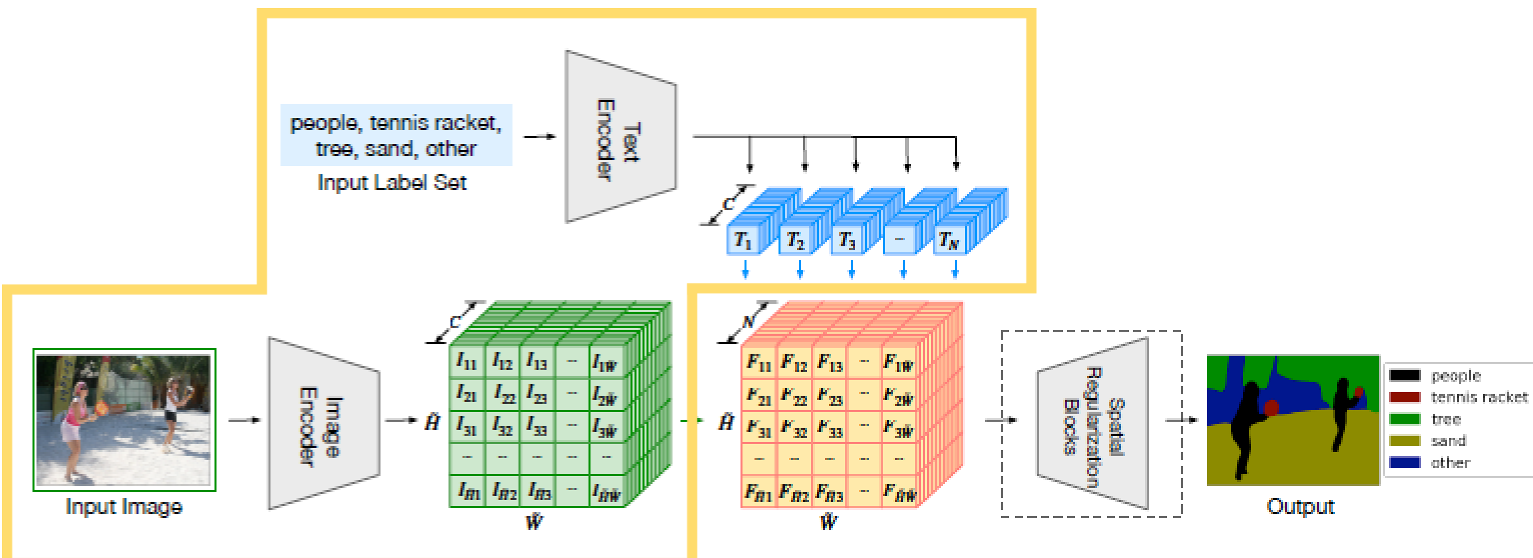


Zero-shot Segmentation

- LSeg [1]
  - Transfer the flexibility of the text encoder to the visual recognition module
    - Text encoder is trained to embed closely related concepts near one another
    - Increase the flexibility and generality of semantic segmentation model
  - Use CLIP [2] text encoder that has been co-trained on visual data
    - To embed labels from the training set into an embedding space

Zero-shot Segmentation

- LSeg [1]
  - Text encoder (using CLIP [2])
    - Embed the set of N potential labels into a vector space
  - Image encoder
    - Extract an embedding vector for each pixel

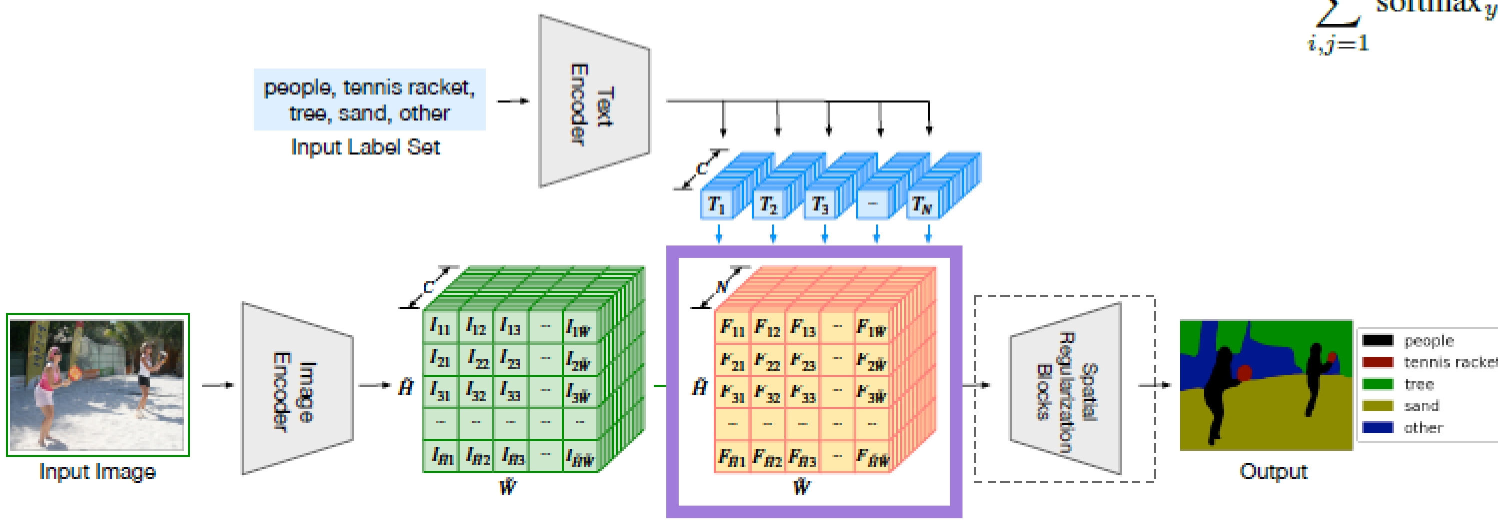


[1] Apple, Intel Labs, et al. “Language Driven Semantic Segmentation” (ICLR 2022)

Zero-shot Segmentation

- LSeg [1]
  - Word-pixel correlation tensor
    - Correlate the embedding of each pixel to all label embeddings by the inner product
    - Train image encoder to provide pixel embeddings close to text embedding of the corresponding GT class
      - Maximize the dot product of the F_{ijk} that corresponds to the GT label k

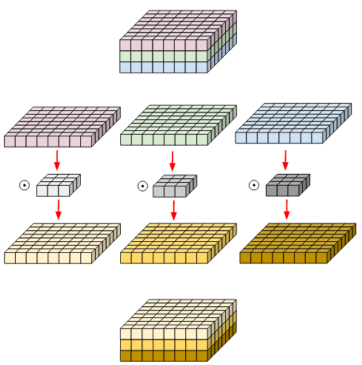
$$\sum_{i,j=1}^{H,W} \text{softmax}_{y_{ij}}\left(\frac{F_{ij}}{t}\right)$$

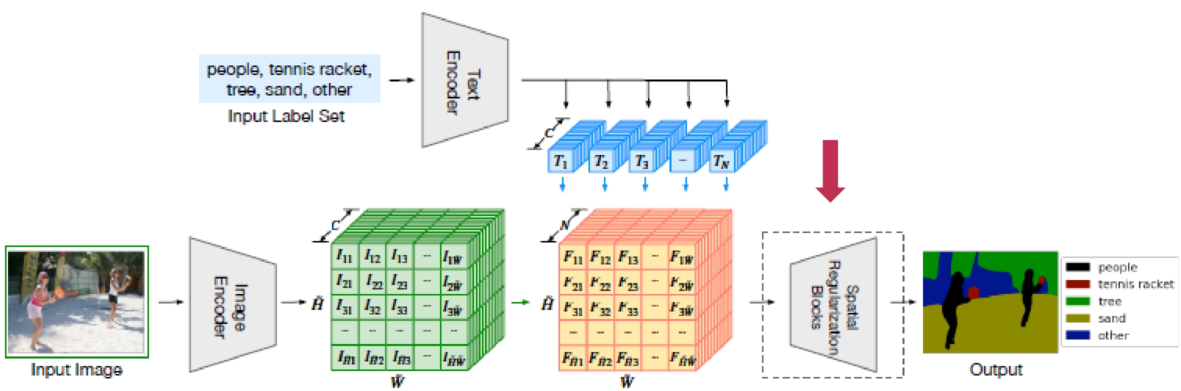


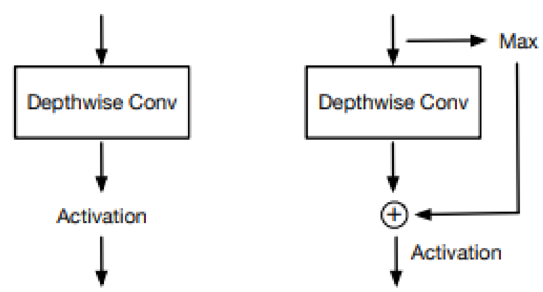
Zero-shot Segmentation

- LSeg [1]
  - Spatial regularization blocks
    - Spatially regularize & upsample the predictions to the original input resolution
    - All operations stay equivariant with respect to the labels
      - ⁘ No interactions between the input channels
        - ✓Use depthwise block or bottleneck block

< Depthwise Conv >







(a) DepthwiseBlock (b) BottleneckBlock

Zero-shot Segmentation

- LSeg [1]
  - Quantitative results
    - Comparison of mIoU on FSS-1000 (Table 1)
      - Train classes : 520 / validation classes : 240 / test classes : 240
    - Comparison on a fixed label set (Table 2)
      - No unseen class labels at test time

Model	Backbone	Method	mIoU
OSLSM	VGG16	1-shot	70.3
GNet	VGG16	1-shot	71.9
FSS	VGG16	1-shot	73.5
DoG-LSTM	VGG16	1-shot	80.8
DAN	ResNet101	1-shot	85.2
HSNet	ResNet101	1-shot	86.5
LSeg	ResNet101	zero-shot	84.7
LSeg	ViT-L/16	zero-shot	**87.8**

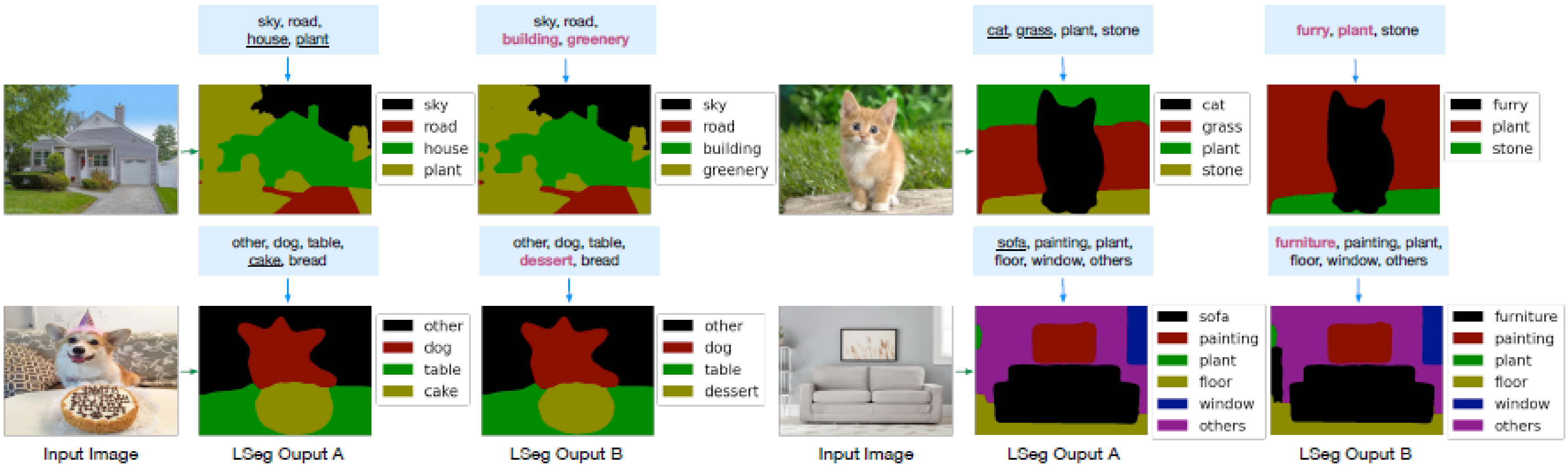
Table 1

Method	Backbone	Text Encoder	pixAcc [%]	mIoU [%]
OCNet	ResNet101	-	-	45.45
ACNet	ResNet101	-	81.96	45.90
DeeplabV3	ResNeSt101	-	82.07	46.91
DPT	ViT-L/16	-	**82.70**	**47.63**
LSeg	ViT-L/16	ViT-B/32	82.46	46.28
LSeg	ViT-L/16	RN50 × 16	**82.78**	**47.25**

Table 2

Zero-shot Segmentation

- LSeg [1]
  - Qualitative results



(a) Related unseen labels. (b) Hierarchical unseen labels.

Zero-shot Segmentation

- LSeg [1]
  - Qualitative results
    - Failure cases
      - Negative samples
      - Assing multiple labels

(a) (b)

Conclusion

- VLT [1]
  - Design vision-language transformer method
    - Holistic understanding of the whole image
    - Comprehend the language expression in different ways incorporating with image information
      - Robustness

- LSeg [2]
  - Increase the flexibility and generality of semantic segmentation model
    - Embed text labels and image pixels into a common space and assigns the closest label to each pixel