BrAIn Lab.
Brain and Artificial Intelligence Lab.

SEOULTECH
SEOUL NATIONAL UNIVERSITY OF SCIENCE & TECHNOLOGY

Segformer & VideoBERT

2026.01.20

BrainLAB Journal Club
신동헌

Overview

01 **Segformer**

02 **VideoBERT**

Enze Xie

팔로우

NVIDIA Research, MMLab@HKU
connect.hku.hk의 이메일 확인됨 - 홈페이지
computer vision generative AI

제목	인용	연도
SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers E Xie, W Wang, Z Yu, A Anandkumar, JM Alvarez, P Luo Conference on Neural Information Processing Systems (NeurIPS), 2021	8844	2021
Pyramid vision transformer: A versatile backbone for dense prediction without convolutions W Wang, E Xie, X Li, DP Fan, K Song, D Liang, T Lu, P Luo, L Shao IEEE International Conference on Computer Vision (ICCV)	6272	2021
PVT v2: Improved baselines with Pyramid Vision Transformer W Wang, E Xie, X Li, DP Fan, K Song, D Liang, T Lu, P Luo, L Shao Computational Visual Media 8 (3), 415-424	2489	2022
Bevformer: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers Z Li, W Wang, H Li, E Xie, C Sima, T Lu, Q Yu, J Dai IEEE Transactions on Pattern Analysis and Machine Intelligence	2187	2024

01 Segformer

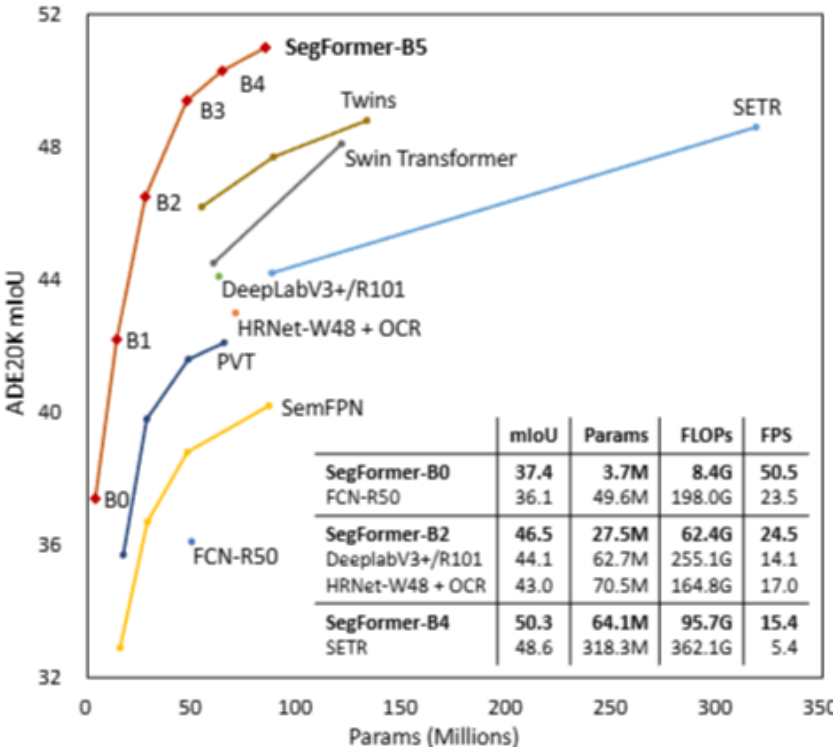
BrAIn Lab.
Brain and Artificial Intelligence Lab.

이전 연구의 한계

- ViT는 다중 스케일 특징이 아닌 단일 스케일의 특징만을 출력
- 큰 이미지에서 계산 비용이 높음
- Encoder 설계에 집중, Decoder를 통한 개선을 고려하지 않음

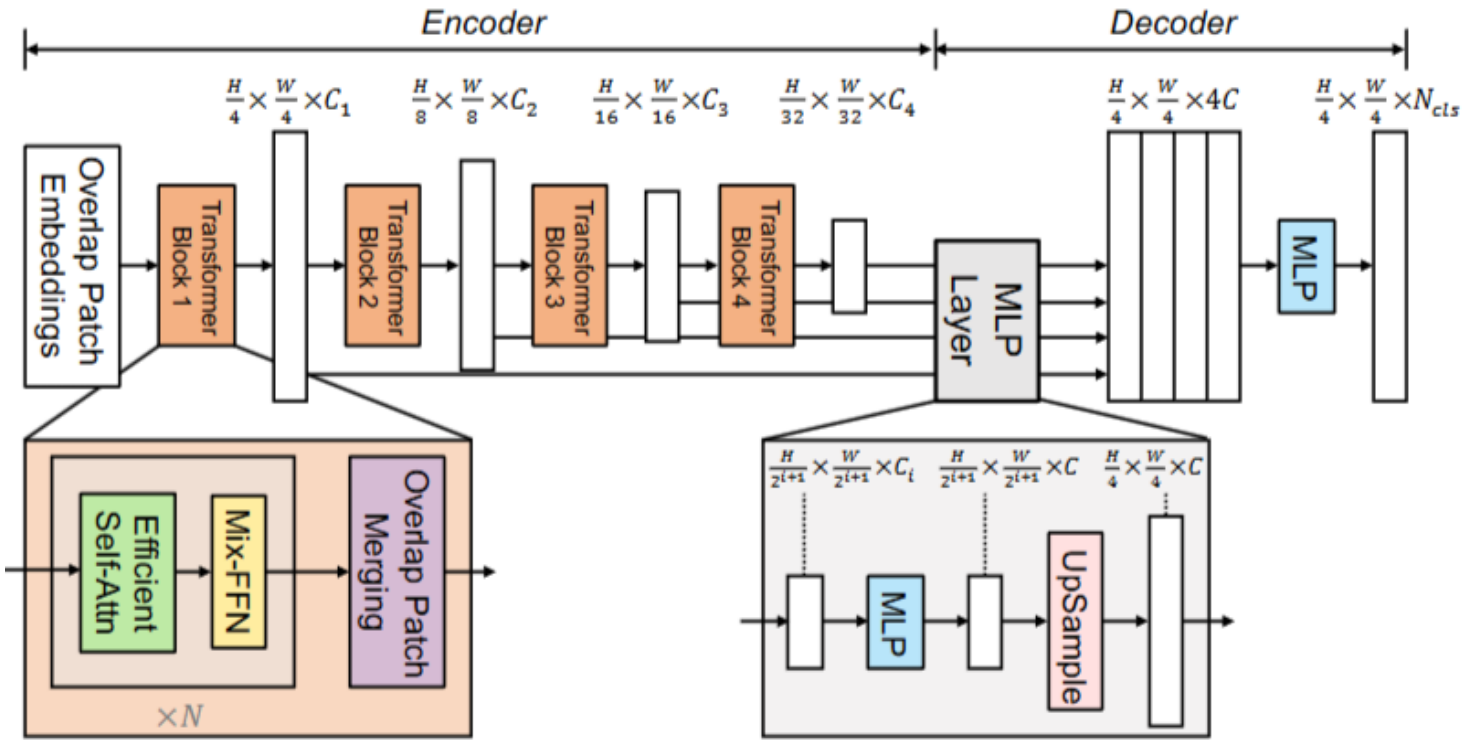
Segformer의 핵심 기여

- 다중 스케일 특징을 출력하는 Transformer Encoder 제안
- MLP Decoder를 활용해 local attention과 global attention을 통합



	mIoU	Params	FLOPs	FPS
SegFormer-B0	**37.4**	**3.7M**	**8.4G**	**50.5**
FCN-R50	36.1	49.6M	198.0G	23.5
SegFormer-B2	**46.5**	**27.5M**	**62.4G**	**24.5**
DeeplabV3+/R101	44.1	62.7M	255.1G	14.1
HRNet-W48 + OCR	43.0	70.5M	164.8G	17.0
SegFormer-B4	**50.3**	**64.1M**	**95.7G**	**15.4**
SETR	48.6	318.3M	362.1G	5.4

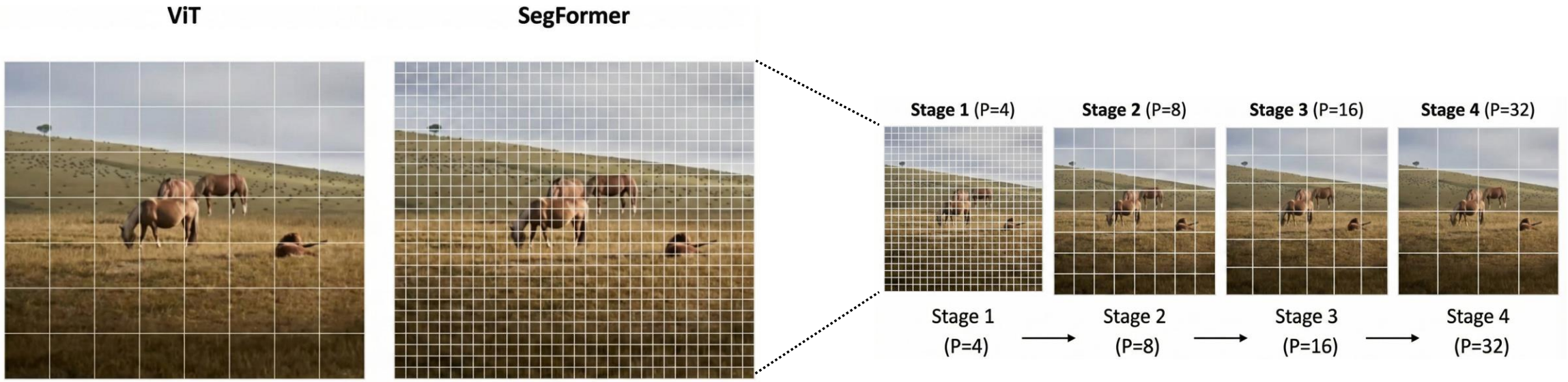
01 Segformer



Segformer Encoder

1. **Overlap patch Merging**
2. **Efficient Self-Attention**
3. **Mix-FFN**

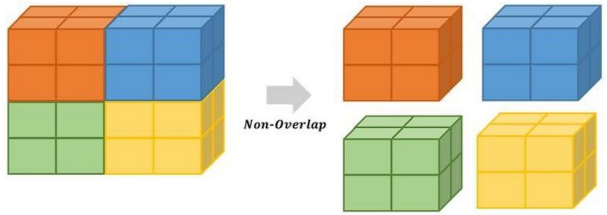
01 Segformer



- ViT는 고정된 patch 사이즈를 사용
- Segformer는 stage별로 다른 가변 patch 사이즈를 사용
- Stage 4는 non-local 정보, Stage 1은 local 정보를 캡처

Overlap patch Merging

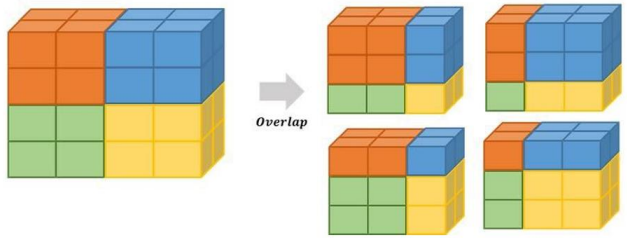
ViT



- ViT의 Merging 방법은 local continuity를 보존 하지 못함

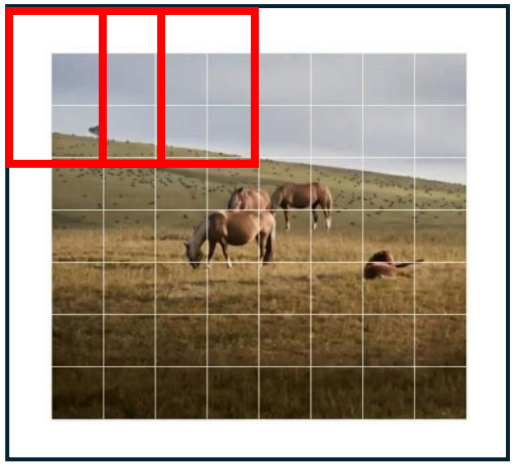
local continuity : 인접한 패치 간의 연결성이 유지되는 성질

Segformer



- Overlap을 통해 기존 ViT의 문제 해결
- Feature map의 해상도를 단계적으로 낮춤

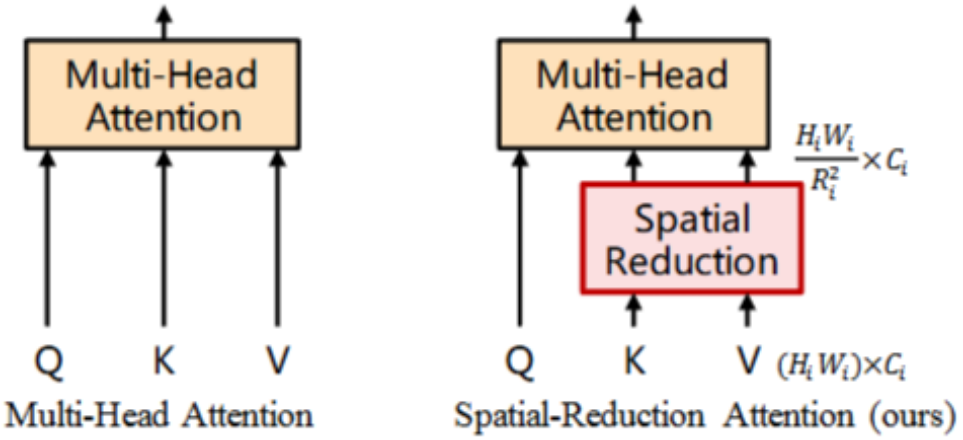
(K=3, S=2, P=1)



Ex) $output = \frac{input + 2P - K}{S} + 1$

input : 8

Efficient Self-Attention



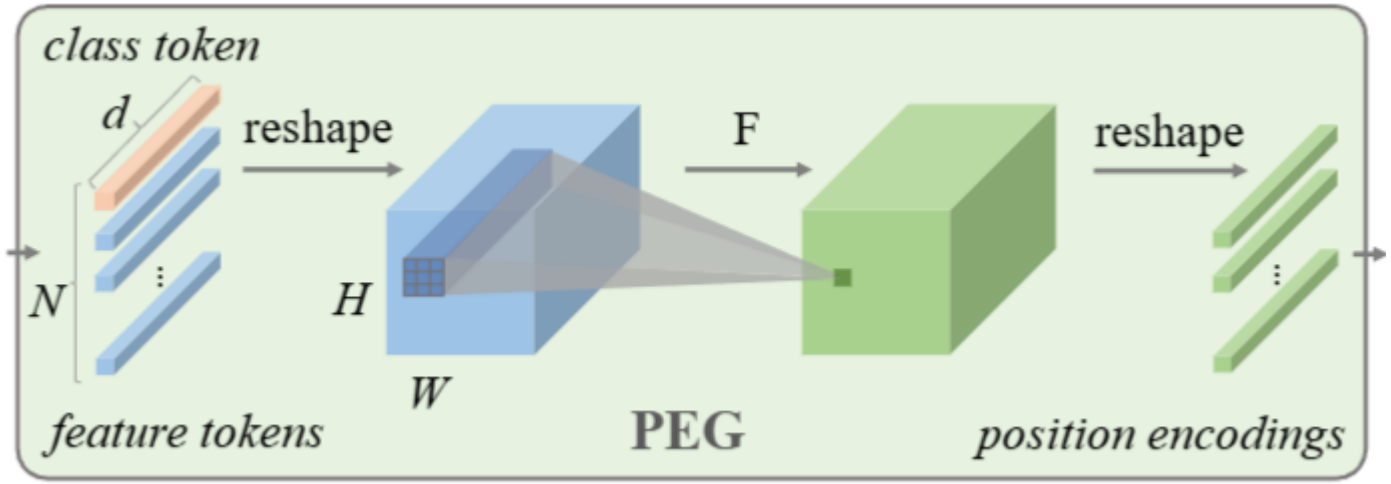
$$\hat{K} = \text{Reshape}(\frac{N}{R}, C \cdot R)(K)$$

$$K = \text{Linear}(C \cdot R, C)(\hat{K}),$$

- Segformer는 ViT 보다 patch를 여러 개 쪼개 연산량이 증가
- Reduction ration(R)를 사용해 cost 문제 해결
- 각 Stage의 R : [64, 16, 4, 1]
- $O(n^2)\ to\ O(\frac{N^2}{R})$

Wang, Wenhai, et al. "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions." *Proceedings of the IEEE/CVF international conference on computer vision*. 2021.

Mix-FFN



- 기존의 Positional Embedding은 고정된 해상도를 가짐

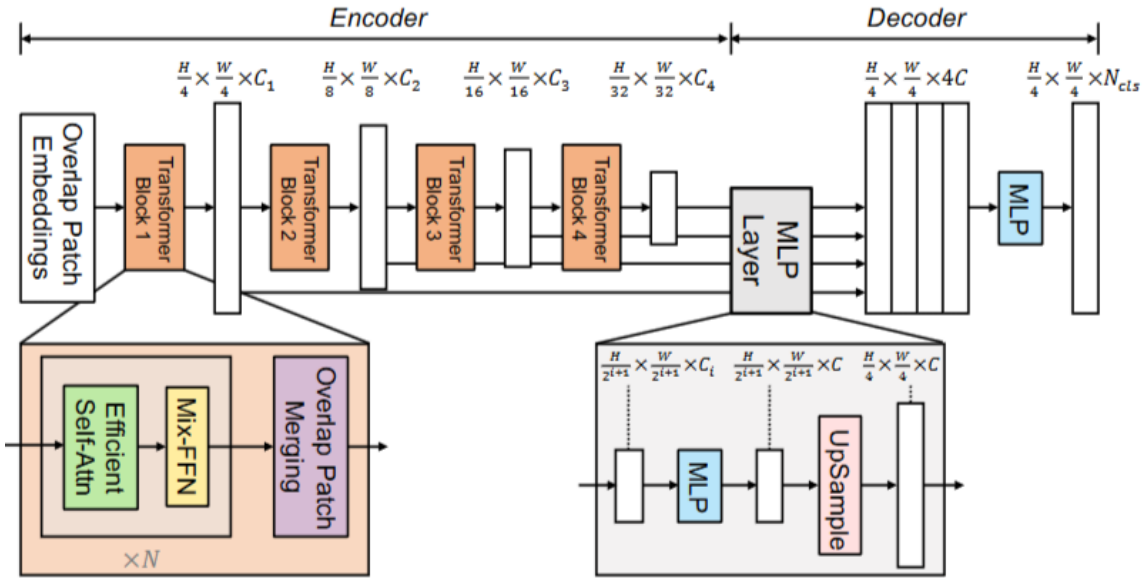
  문제점 : 입력 크기가 변하면 성능 저하 발생

$$\mathbf{x}_{out} = \mathrm{MLP}(\mathrm{GELU}(\mathrm{Conv}_{3\times3}(\mathrm{MLP}(\mathbf{x}_{in})))) + \mathbf{x}_{in}$$

Chu, Xiangxiang, et al. "Conditional positional encodings for vision transformers." *arXiv preprint arXiv:2102.10882* (2021).

BrAIn Lab.
Brain and Artificial Intelligence Lab.

All-MLP Decoder

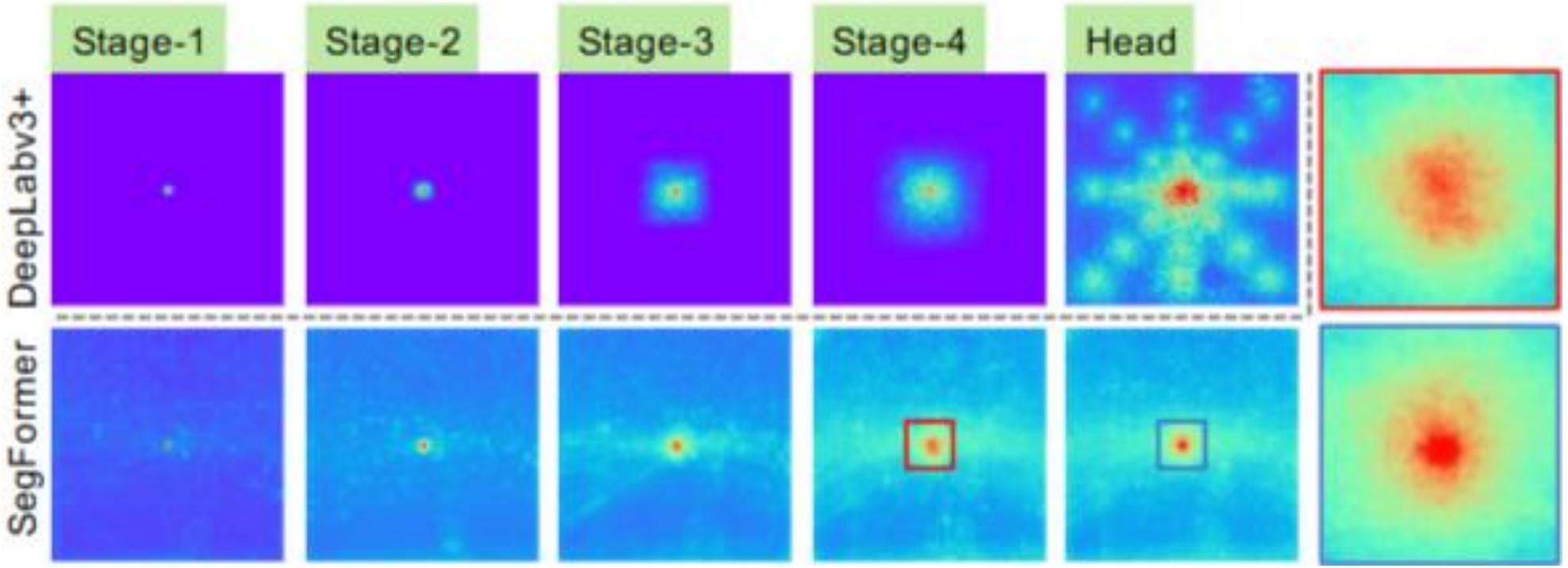


$$\hat{F}_i = \text{Linear}(C_i, C)(F_i), \forall i$$
$$\hat{F}_i = \text{Upsample}(\frac{W}{4} \times \frac{W}{4})(\hat{F}_i), \forall i$$
$$F = \text{Linear}(4C, C)(\text{Concat}(\hat{F}_i)), \forall i$$
$$M = \text{Linear}(C, N_{cls})(F),$$

- Transformer encoder가 CNN보다 큰 수용 영역을 갖기 때문에 간단한 decoder 사용이 가능

01 Segformer

BrAIn Lab.
Brain and Artificial Intelligence Lab.



BrAIn Lab.
Brain and Artificial Intelligence Lab.

Encoder	Params		ADE20K		Cityscapes		COCO-Stuff	
Model Size	Encoder	Decoder	Flops ↓	mIoU(SS/MS) ↑	Flops ↓	mIoU(SS/MS) ↑	Flops ↓	mIoU(SS) ↑
MiT-B0	3.4	0.4	8.4	37.4 / 38.0	125.5	76.2 / 78.1	8.4	35.6
MiT-B1	13.1	0.6	15.9	42.2 / 43.1	243.7	78.5 / 80.0	15.9	40.2
MiT-B2	24.2	3.3	62.4	46.5 / 47.5	717.1	81.0 / 82.2	62.4	44.6
MiT-B3	44.0	3.3	79.0	49.4 / 50.0	962.9	81.7 / 83.3	79.0	45.5
MiT-B4	60.8	3.3	95.7	50.3 / 51.1	1240.6	82.3 / 83.9	95.7	46.5
MiT-B5	81.4	3.3	183.3	51.0 / 51.8	1460.4	82.4 / 84.0	111.6	46.7

B0-1 (C : 256)

B2-5 (C : 768)

	Method	Encoder	Params ↓	ADE20K			Cityscapes		
				Flops ↓	FPS ↑	mIoU ↑	Flops ↓	FPS ↑	mIoU ↑
Real-Time	FCN [1]	MobileNetV2	9.8	39.6	64.4	19.7	317.1	14.2	61.5
	ICNet [11]	-	-	-	-	-	-	30.3	67.7
	PSPNet [17]	MobileNetV2	13.7	52.9	57.7	29.6	423.4	11.2	70.2
	DeepLabV3+ [20]	MobileNetV2	15.4	69.4	43.1	34.0	555.4	8.4	75.2
	SegFormer (Ours)	MiT-B0	**3.8**	**8.4**	**50.5**	**37.4**	125.5	15.2	**76.2**
				-	-	-	51.7	26.3	75.3
				-	-	-	31.5	37.1	73.7
				-	-	-	**17.7**	**47.6**	71.9
Non Real-Time	FCN [1]	ResNet-101	68.6	275.7	14.8	41.4	2203.3	1.2	76.6
	EncNet [24]	ResNet-101	**55.1**	218.8	14.9	44.7	1748.0	1.3	76.9
	PSPNet [17]	ResNet-101	68.1	256.4	15.3	44.4	2048.9	1.2	78.5
	CCNet [41]	ResNet-101	68.9	278.4	14.1	45.2	2224.8	1.0	80.2
	DeeplabV3+ [20]	ResNet-101	62.7	255.1	14.1	44.1	2032.3	1.2	80.9
	OCRNet [23]	HRNet-W48	70.5	164.8	**17.0**	45.6	1296.8	**4.2**	81.1
	GSCNN [35]	WideResNet38	-	-	-	-	-	-	80.8
	Axial-DeepLab [74]	AxialResNet-XL	-	-	-	-	2446.8	-	81.1
	Dynamic Routing [75]	Dynamic-L33-PSP	-	-	-	-	**270.0**	-	80.7
	Auto-Deeplab [50]	NAS-F48-ASPP	-	-	-	44.0	695.0	-	80.3
	SETR [7]	ViT-Large	318.3	-	5.4	50.2	-	0.5	82.2
	SegFormer (Ours)	MiT-B4	64.1	**95.7**	15.4	51.1	1240.6	3.0	83.8
	SegFormer (Ours)	MiT-B5	84.7	183.3	9.8	**51.8**	1447.6	2.5	**84.0**

(b) Accuracy as a function of the MLP dimension C in the decoder on ADE20K.

C	Flops ↓	Params ↓	mIoU ↑
256	25.7	24.7	44.9
512	39.8	25.8	45.0
768	62.4	27.5	45.4
1024	93.6	29.6	45.2
2048	304.4	43.4	45.6

(c) Mix-FFN vs. positional encoding (PE) for different test resolution on Cityscapes.

Inf Res	Enc Type	mIoU ↑
768×768	PE	77.3
1024×2048	PE	74.0
768×768	Mix-FFN	80.5
1024×2048	Mix-FFN	79.8

Chen Sun

팔로우

Assistant Professor, Brown University
brown.edu의 이메일 확인됨 - 홈페이지

Computer Vision Machine Learning Artificial Intelligence

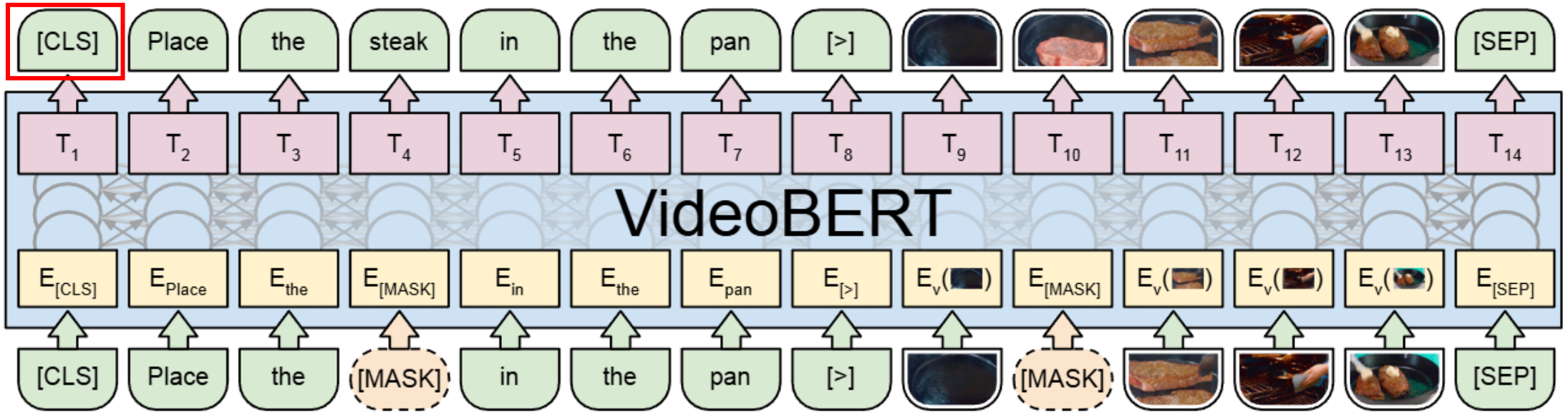
제목	인용	연도
OpenImages: A public dataset for large-scale multi-label and multi-class image classification I Krasin, T Duerig, N Alldrin, A Veit, S Abu-El-Haija, S Belongie, D Cai, ... github.com/openimages	4628 *	2016
Speed/accuracy Trade-offs for Modern Convolutional Object Detectors J Huang, V Rathod, C Sun, M Zhu, A Korattikara, A Fathi, I Fischer, ... CVPR 2017	3850	2017
Revisiting Unreasonable Effectiveness of Data in Deep Learning Era C Sun, A Shrivastava, S Singh, A Gupta ICCV 2017	3812	2017
ViViT: A Video Vision Transformer A Arnab, M Dehghani, G Heigold, C Sun, M Lučić, C Schmid ICCV 2021	3718	2021
The iNaturalist species classification and detection dataset G Van Horn, O Mac Aodha, Y Song, Y Cui, C Sun, A Shepard, H Adam, ... CVPR 2018	2420 *	2018
Rethinking Spatiotemporal Feature Learning For Video Understanding S Xie, C Sun, J Huang, Z Tu, K Murphy ECCV 2018	2202 *	2018
What makes for good views for contrastive learning Y Tian, C Sun, B Poole, D Krishnan, C Schmid, P Isola NeurIPS 2020	1845	2020
VideoBERT: A Joint Model for Video and Language Representation Learning C Sun, A Myers, C Vondrick, K Murphy, C Schmid ICCV 2019	1739	2019

- Visual Domain과 Language Domain 관계를 모델링하는 방법
- Text를 자동으로 시각화하는 text-to-video 예측을 수행

- BERT는 **Masked Language Model**
- 비디오(시각 정보)는 Mask로 가린 뒤 해당 위치에 적절한 영상을 찾는 작업이 어려움

  ➔ Vector Quantization을 진행

Text와 영상이 일치하는지 확인

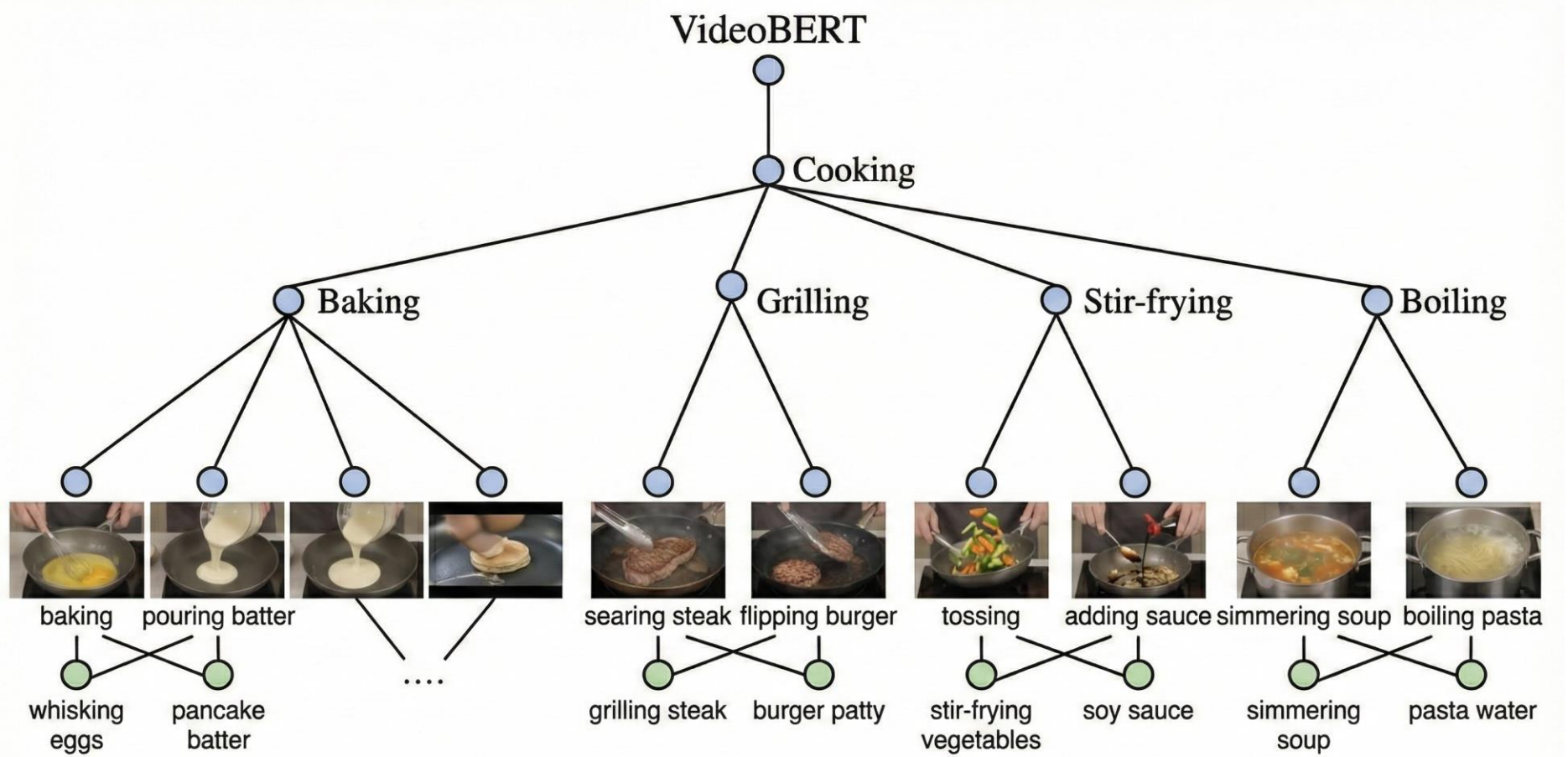


Vector Quantization

1. 비디오를 20fps 샘플링
2. 30 프레임 window를 사용해 클립 생성
3. 클립에 S3D를 사용해 특징 추출(1024 차원) S3D : Separable 3D CNN
4. 계층적 K-means clustering을 사용해 토큰화(d : 4, k : 12)

"but in the meantime, you're just kind of moving around your cake board and you can keep reusing make sure you're working on a clean service so you can just get these all out of your way but it's just a really fun thing to do especially for a birthday party."

"apply a little bit of butter on one side and place a portion of the stuffing and spread evenly cover with another slice of the bread and apply some more butter on top since we're gonna grill the sandwiches."



Hierarchical clustering of cooking videos in VideoBERT

영상과 Text의 불일치 문제

- 영상에서 화자가 시각적으로 나타나지 않는 것을 언급하는 경우
- 영상과 Text가 같은 시점에 정확히 나타나지 않는 경우

해결방법

- 인접한 문장을 하나의 긴 문장으로 연결

  ➔ 긴 문맥 파악 가능
- 영상마다 재생 속도가 다를 수 있기 때문에 시각적 토큰에 1-5 단계 sub sampling 비율을 무작위로 선택

  ➔ 시간적 변동성에 대한 내성 강화

영상을 보고 **동사**와 **목적어** 추출

now let me show you how to **[MASK]** the **[MASK]**

Top verbs: make, assemble, prepare
Top nouns: pizza, sauce, pasta

Top verbs: make, do, pour
Top nouns: cocktail, drink, glass

영상을 보고 영상의 **특징을 text**로 추출

now let's **[MASK]** the **[MASK]** to the **[MASK]**, and then **[MASK]** the **[MASK]**.

GT: add some chopped basil leaves into it

VideoBERT: chop the basil and add to the bowl

S3D: cut the tomatoes into thin slices

GT: cut the top off of a french loaf

VideoBERT: cut the bread into thin slices

S3D: place the bread on the pan

Thank You

2026.01.20

BrainLAB Journal Club
신동헌