Vision AI

2021 arXiv Trends

2021-06

hutom

no.	Paper Title	Correspondence	h-index
1	Deep Learning based Multi-modal Computing with Feature Disentanglement for MRI Image Synthesis	Yan Wang	10
2	Do You Even Need Attention? A Stack of Feed-Forward Layers Does Surprisingly Well on ImageNet	Luke Melas-Kyriazi	4
3	Vision Transformers for Dense Prediction	Vladlen Koltun	75
4	Is Space-Time Attention All You Need for Video Understanding?	Gedas Bertasius	13
5	LocalViT: Bringing Locality to Vision Transformers	Luc Van Gool	158

Deep Learning based Multi-modal Computing with Feature Disentanglement for MRI Image Synthesis

Yuchen Fei[1], Bo Zhan[1], Mei Hong[1], Xi Wu[2], Jiliu Zhou[1,2], Yan Wang[1,*]

https://arxiv.org/ftp/arxiv/papers/2105/2105.02835.pdf

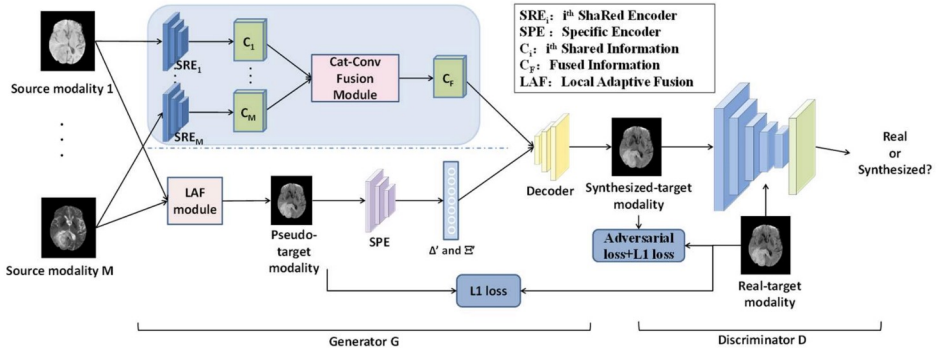


Figure 1: Framework of the proposed method.

- **Purpose**
  - full-sequence MRI images 를 획득하는 것은 어려움.
  - 높은 정확도의 **target MRI sequences prediction** 과 clinical 진단을 위한 정보를 제공하는 것.

- **Method**
  - Deep learning 기반의 multi-modal computing model 제안.
    - 다른 modalities 에서 충분한 정보를 얻기 위하여, multi-modal MRI sequences 가 input 으로 사용됨.
  - Input modality 의 feature를 효과적으로 구분하기 위하여 다음 2가지로 분리함.
    - Modality-invariant space : shared information
    - Modality-specific space : specific information
    - 각각의 input 은 Adaptive Instance Normalization (AdaIN) layer 를 통해 다시 결합됨.
  - Target modality 의 **test phase 에서 specific information 이 부족할 경우와 관련**하여
    - Local Adaptive Fusion (LAF) module 로써 GT 와 유사한 specific information 정보를 생성함.

- **Result**
  - Synthesis performance 를 평가하기 위해, BRATS2015 dataset 사용.
  - 실험 결과) 정성적, 정량적 측정 모두에서 벤치마크 방법과 SOTA medical image synthesis methods 를 능가함.

Do You Even Need Attention? A Stack of Feed-Forward Layers Does Surprisingly Well on ImageNet

Luke Melas-Kyriazi
Oxford University
lukemk@robots.ox.ac.uk

https://arxiv.org/pdf/2105.02723.pdf

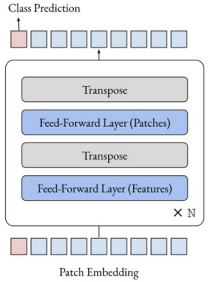


Figure 1: The architecture explored in this report is extremely simple, consisting of a patch embedding followed by a series of feed-forward layers. These feed-forward layers are alterately applied to the patch and feature dimensions of the image tokens. The architecture is identical to that of ViT [4] with the attention layer replaced by a feed-forward layer.

		Params	ImageNet Top-1
Tiny ($P = 16$)	ViT	-	-
	DeiT	5.7M	72.2
	FF Only	7.7M	61.4
Base ($P = 16$)	ViT	86M	77.9
	DeiT	86M	79.9
	FF Only	62M	74.9
Large ($P = 32$)	ViT	306M	71.2
	DeiT	-	-
	FF Only	206M	71.4

Table 1: A comparison of ImageNet top-1 accuracies for different model sizes. In the first column, P refers to the patch size in pixels. Overall, the models with only feed-forward layers (*FF Only*) perform worse than their counterparts with attention, but they perform surprisingly well regardless. Performance deteriorates for the largest models both with and without attention.

- Short report
  - transformer-style networks **without attention layers** make for surprisingly strong image classifiers.
  - pytorch 코드 제공.

- Image classification 과 다른 vision tasks 에서의 **Vision transformer** 는 강력한 성능을 보임.
- **However, 이러한 성과에 있어 Attention 이 어느 정도까지 영향을 미치는지는 여전히 불분명함.**
  - **Is the attention layer even necessary?**

- **본 실험에서, vision transformer 의 attention layer 를 patch dimension을 적용한 feed-forward layer 로 대체함.**
  - The resulting architecture is simply series of **feed-forward layers applied over the patch and feature dimensions.**

- Experiments
  - On ImageNet, this architecture performs surprisingly well.
- **Vision transformers 에서 attention 이외의 patching embedding 등 다른 요인이 중요하게 작동하고 있음을 예상할 수 있음.**

- Do You Even Need Feed-Forward Layers?
  - replaced the feed-forward layer over the feature dimension with an **attention layer over the feature dimension.**
  - but it performed spectacularly poorly.

- Github : https://github.com/lukemelas/do-you-even-needattention

Vision Transformers for Dense Prediction

René Ranftl Alexey Bochkovskiy Vladlen Koltun

Intel Labs

rene.ranftl@intel.com

https://arxiv.org/pdf/2103.13413.pdf

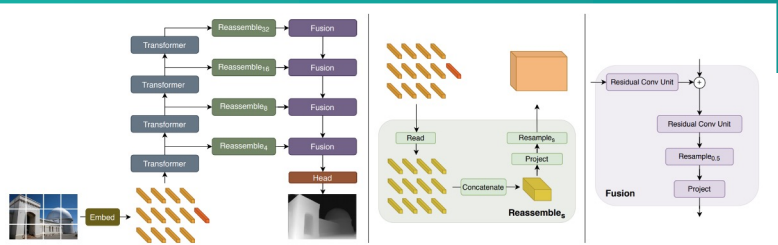


Figure 1. *Left*: Architecture overview. The input image is transformed into tokens (orange) either by extracting non-overlapping patches followed by a linear projection of their flattened representation (DPT-Base and DPT-Large) or by applying a ResNet-50 feature extractor (DPT-Hybrid). The image embedding is augmented with a positional embedding and a patch-independent readout token (red) is added. The tokens are passed through multiple transformer stages. We reassemble tokens from different stages into an image-like representation at multiple resolutions (green). Fusion modules (purple) progressively fuse and upsample the representations to generate a fine-grained prediction. *Center*: Overview of the Reassemble$_s$ operation. Tokens are assembled into feature maps with $\frac{1}{s}$ the spatial resolution of the input image. *Right*: Fusion blocks combine features using residual convolutional units [23] and upsample the feature maps.

	Training set	DIW WHDR	ETH3D AbsRel	Sintel AbsRel	KITTI δ>1.25	NYU δ>1.25	TUM δ>1.25
DPT - Large	MIX 6	**10.82** (-13.2%)	**0.089** (-31.2%)	**0.270** (-17.5%)	**8.46** (-64.6%)	**8.32** (-12.9%)	**9.97** (-30.3%)
DPT - Hybrid	MIX 6	11.06 (-11.2%)	0.093 (-27.6%)	0.274 (-16.2%)	11.56 (-51.6%)	8.69 (-9.0%)	10.89 (-23.2%)
MiDaS	MIX 6	12.95 (+3.9%)	0.116 (-10.5%)	0.329 (+0.5%)	16.08 (-32.7%)	8.71 (-8.8%)	12.51 (-12.5%)
MiDaS [30]	MIX 5	12.46	0.129	0.327	23.90	9.55	14.29
Li [22]	MD [22]	23.15	0.181	0.385	36.29	27.52	29.54
Li [21]	MC [21]	26.52	0.183	0.405	47.94	18.57	17.71
Wang [40]	WS [40]	19.09	0.205	0.390	31.92	29.57	20.18
Xian [45]	RW [45]	14.59	0.186	0.422	34.08	27.00	25.02
Casser [5]	CS [8]	32.80	0.235	0.422	21.15	39.58	37.18

Table 1. Comparison to the state of the art on monocular depth estimation. We evaluate zero-shot cross-dataset transfer according to the protocol defined in [30]. Relative performance is computed with respect to the original MiDaS model [30]. Lower is better for all metrics.

- Transformer가 Vision 에 많이 적용되므로 해당 논문 선정.
  - 3월 말부터 꾸준히 Top Hype
- **Dense Vision Transformers**
  - Dense prediction tasks (semantic image segmentation) 를 위함.
  - An architecture that CNN networks 대신에 vision transformers 를 백본으로 활용.
- **Vision Transformer 를 점진적으로 사용**
  - 각 단계별로 Token을 다양한 Resolution 으로 Assemble
    - image embedding 은 positional embedding 과 함께 augmented 됨. + patch-independent readout token 추가.
    - 각각의 token은 multiple transformer stages 통과.
  - Convolutional decoder 구조를 사용하여 Token을 Full-Resolutions Predictions 으로 점진적 Combine.
- 기존 FCN과 비교하여 Dense Prediction tasks 에서 상당한 개선을 가져옴.
  - 특히 train dataset 이 많을 때
  - Monocular Depth Estimation 으로 SOTA FCN 과 성능 비교시 최대 28%의 개선을 보임.
- Github : : https://github.com/intel-isl/DPT

Is Space-Time Attention All You Need for Video Understanding?

Gedas Bertasius [1] **Heng Wang** [1] **Lorenzo Torresani** [1 2]

https://arxiv.org/pdf/2102.05095.pdf

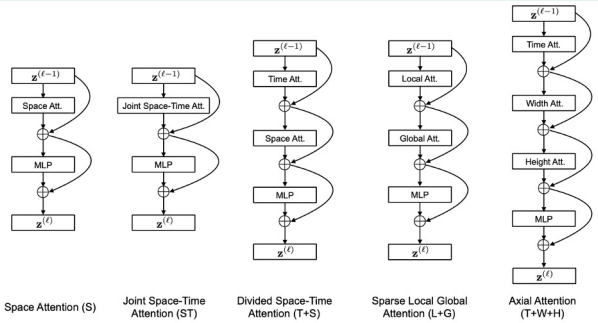
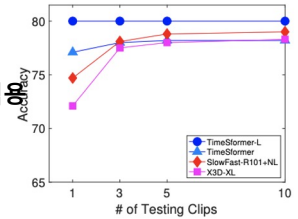


Figure 1. The video self-attention blocks that we investigate in this work. Each attention layer implements self-attention (Vaswani et al., 2017b) on a specified spatiotemporal neighborhood of frame-level patches (see Figure 2 for a visualization of the neighborhoods). We use residual connections to aggregate information from different attention layers within each block. A 1-hidden-layer MLP is applied at the end of each block. The final model is constructed by repeatedly stacking these blocks on top of each other.

Method	Top-1	Top-5
I3D-R50+Cell (Wang et al., 2020c)	79.8	94.4
LGD-3D-101 (Qiu et al., 2019)	81.5	95.6
SlowFast (Feichtenhofer et al., 2019b)	81.8	95.1
X3D-XL (Feichtenhofer, 2020)	81.9	95.5
TimeSformer	79.1	94.4
TimeSformer-HR	81.8	**95.8**
TimeSformer-L	**82.2**	95.6

Table 6. Video-level accuracy on Kinetics-600.



- Transformer와 Vision 연관, Action Recognition 실험
  - 2월부터 꾸준히 Top hype
- Video Classification 문제에서 Convolution free 아키텍처 접근방식 제시
- **TimeSformer**
  - 공간과 시간 (Space and Time) 에 기반한 self-attention
  - **Frame-Level patch로부터 직접 spatiotemporal feature learning을 가능하게 하여 Standard Transformer architecture 에 적용**
- Experiments
  - self-attention 과 비교.
  - divided attention : temporal attention 과 spatial attention 을 각 block 에 별도로 적용
  - self-attention과 비교하여 divided attention 이 best video classification 정확도
- 몇몇의 Action Recognition benchmark에서 SOTA 성능 도달
  - Kinetics-400과 600에서 best
- Github : https://github.com/facebookresearch/TimeSformer

LocalViT: Bringing Locality to Vision Transformers

Yawei Li[1] Kai Zhang[1] Jiezhang Cao[1] Radu Timofte[1] Luc Van Gool[1,2]
[1]Computer Vision Lab, ETH Zurich, Switzerland [2]KU Leuven, Belgium
{yawei.li, kai.zhang, jiezhang.cao, timofter, vangool}@vision.ee.ethz.ch

https://arxiv.org/pdf/2104.05707.pdf

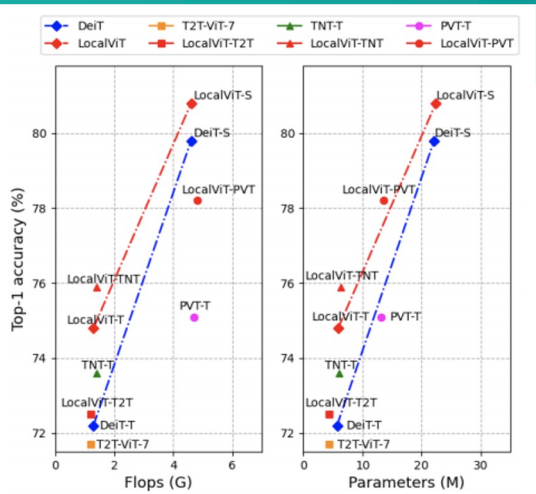


Figure 1: Comparison between LocalViT and the baseline transformers. The transformers enhanced by the proposed locality mechanism outperform their baselines.

- Luc Van Gool
  - h-index : 158
- vision transformers 에 locality mechanisms 적용.
  - 기존 transformers 의 self-attention 메커니즘으로 토큰 간 global interaction 이 잘 생성될 수 있었지만, local region 내에서의 information 교환을 위한 locality mechanism은 부족했음.
  - 이미지에서는 locality가 필수적임. (lines, edges, shapes, objects)
- vision transformers feed-forward network에 depth-wise convolution을 적용함으로써 locality 추가.
- locality mechanism은 두 가지 방법으로 검증됨.
  1) 넓은 범위의 design choices (activation function, layer placement, expansion ratio) 가 사용 가능하고, baseline 보다 더 좋은 성능을 보임.
  2) 동일한 locality 메커니즘이 4개의 vision transformer에 적용되었을 때, 모두 locality concept에서 좋은 일반화 (generalization) 를 보임.
- ImageNet2012 classification 에서 locality-enhanced transformers 는 baselines DeiT-T, PVT-T 보다 2.6%, 3.1% 좋은 성능을 보임.
- github : https://github.com/ofsoundof/LocalViT

End of the Document