UFO-ViT: High Performance Linear Transformer without Softmax

School of Industrial and Management Engineering, Korea University

Young Jae Lee

KOREA UNIVERSITY

DMQA

Contents

Research Purpose

❖ UFO-ViT: High Performance Linear Vision Transformer without Softmax (arXiv, 2021)

- Kakao에서 연구하였고 2021년 11월 05일 기준으로 0회 인용

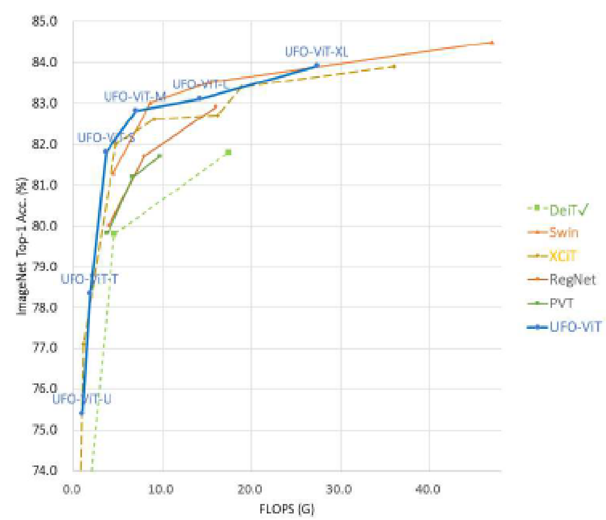
UFO-ViT: High Performance Linear Vision Transformer without Softmax

Jeong-geun Song
Kakao Enterprise
po.ai@kakaoenterprise.com

PREPRINT

Abstract

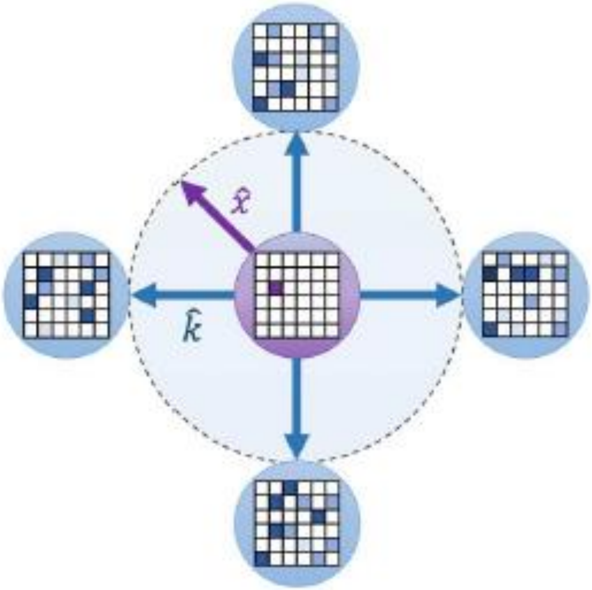
Vision transformers have become one of the most important models for computer vision tasks. While they outperform earlier convolutional networks, the complexity quadratic to N is one of the major drawbacks when using traditional self-attention algorithms. Here we propose the UFO-ViT(Unit Force Operated Vision Trnasformer), novel method to reduce the computations of self-attention by eliminating some non-linearity. Modifying few of lines from self-attention, UFO-ViT achieves linear complexity without the degradation of performance. The proposed models outperform most transformer-based models on image classification and dense prediction tasks through most capacity regime.



Research Purpose

❖ UFO-ViT: High Performance Linear Vision Transformer without Softmax (arXiv, 2021)

- Vision Transformer에서 기존 Self-Attention의 연산 복잡도와 Data-Efficiency를 향상시키기 위한 방법을 제안
  - ✓ Matrix Multiplication을 사용하기 위해 Softmax 연산을 제거
  - ✓ Linear Complexity의 Self-Attention 연산을 위한 Xnorm 제안(Constraint)
- Unit Hypersphere 상에서 Feature들 간의 관계를 추출하도록 함



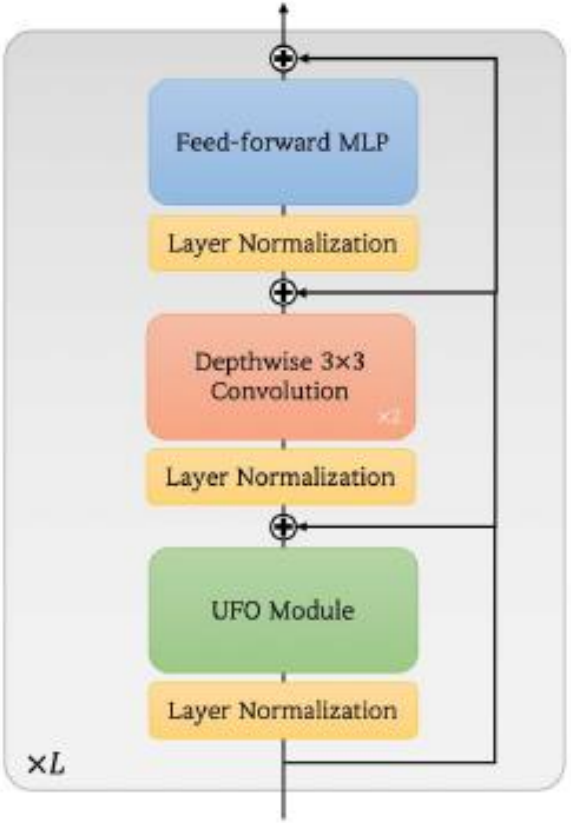
DMQA

UFO-ViT

❖ Overview of UFO-ViT Module

- Convolutional Layers, UFO Module, Simple Feed-Forward MLP Layer, Residual Connection으로 구성

- Patch Embedding with Convolutions
  - ✓ Linear Projection 대신 합성곱 Patch Embedding Layers를 차용
- Positional Encoding
  - ✓ 학습 가능한 매개변수로 Positional Encoding 사용
- Multi-Headed Attention
- Local Patch Interaction
  - ✓ 3 × 3 Depth-wise Convolution 사용
- Feed-Forward Network
- Class Attention
  - ✓ Spatial Information을 모으기 위한 CLS Token 예측



DMQA

UFO-ViT

❖ UFO Module

- 기존 Self-Attention 연산에서는 Softmax의 비선형성으로 인하여 분리가 불가능함
- 제안하는 방법은 K^TV를 먼저 계산하기 위해 Softmax 연산을 제거(간단한 Constraint 포함)
  - ✓ Cross-Normalization or Xnorm
  - ✓ 간단한 L_2-norm이지만 공간 차원 K^TV와 채널 차원 Q가 있어 Cross-Normalization으로도 부름
  - ✓ 연산 법칙으로 K^TV를 먼저 계산한 후에 Q를 곱함
  - ✓ γ: 학습가능한 파라미터 / h: embedding 차원 수

XNorm

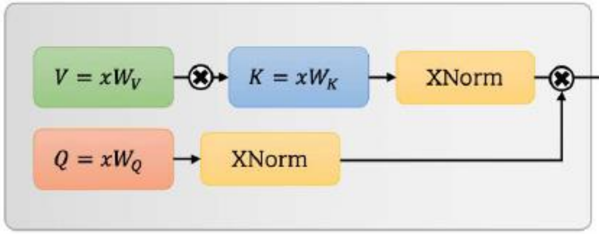
기존 Self-Attention

$$A(\mathbf{x}) = \sigma(QK^T/\sqrt{d_k})V \quad (1)$$

$$Q = \mathbf{x}W_Q, K = \mathbf{x}W_K, V = \mathbf{x}W_V \quad (2)$$

$$A(\mathbf{x}) = \mathrm{XN}_{\mathrm{dim=filter}}(Q)(\mathrm{XN}_{\mathrm{dim=space}}(K^TV)) \quad (3)$$

$$\mathrm{XN}(\mathbf{a}) := \frac{\gamma\mathbf{a}}{\sqrt{\sum_{i=0}^{h}||\mathbf{a}||^2}} \quad (4)$$



UFO-ViT

❖ UFO Module

- 연산 법칙으로 $K^T V$를 먼저 계산한 후에 Q를 곱함
- 위 연산으로 $O(hNd)$ Complexity를 가지며 N에 Linear함

Module Type	Complexity
ViT[10]	$O(N^2d)$
Linformer[40]	$O(kNd)$
Efficient Attention[31]	$O(hNd)$
Axial[20]	$O(N\sqrt{N}d)$
XCiT[12]	$O(Nd^2)$
UFO-ViT	$O(hNd)$

DMQA

UFO-ViT

❖ XNorm

- Softmax 연산을 XNorm으로 대체
  - ✓ Key와 Value 계산(5)
  - ✓ $K^T V$와 Q 에 Xnorm 적용(7, 8) 및 Attention Operator 계산(6)
  - ✓ Projection Weight Scales를 Weight Sum에 의해 계산(9)

$$[K^T V]_{ij} = \sum_{k=1}^{n} K^T_{ik} V_{kj} \qquad (5)$$

$$A(\mathbf{x}) = \begin{bmatrix} \hat{q}_0 \cdot \hat{k}_0 & \hat{q}_0 \cdot \hat{k}_1 & \cdots & \hat{q}_0 \cdot \hat{k}_h \\ \hat{q}_1 \cdot \hat{k}_0 & \hat{q}_1 \cdot \hat{k}_1 & \cdots & \hat{q}_1 \cdot \hat{k}_h \\ \vdots & \vdots & \ddots & \vdots \\ \hat{q}_N \cdot \hat{k}_0 & \hat{q}_N \cdot \hat{k}_1 & \cdots & \hat{q}_N \cdot \hat{k}_h \end{bmatrix} \qquad (6)$$

$$\hat{q}_i = \mathrm{XN}[(Q_{i0}, Q_{i1}, \cdots, Q_{ih})] \qquad (7)$$

$$\hat{k}_i = \mathrm{XN}[([K^T V]_{0i}, [K^T V]_{1i}, \cdots, [K^T V]_{hi})] \qquad (8)$$

$$[W_{\text{proj}} A(\mathbf{x})]_{ij} = \sum_{m=1}^{h} w_{mj} \hat{q}_i \cdot \hat{k}_j \qquad (9)$$

Experiments

❖ Image Classification Metric

- Top-1 Accuracy: Softmax의 Output에서 제일 높은 수치를 가지는 값이 정답일 경우에 대한 지표
- Float Point Operations Per Second (FLOPs): 컴퓨터의 성능을 표현하는 지표
- Parameters: Model의 Weight 또는 Parameter 수

❖ Object Detection Metric

- Average Precision (AP): IoU 계산 결과 값이 0.5 이상이면 True Positive (TP), 0.5 미만이면 False Positive (FP)로 판단하고 검출 결과들 중 옳게 검출한 비율을 의미(정확도)

Experiments

Image Classification Results

❖ ImageNet-1K Dataset

Hyperparam	Model	Value
learning rate	UFO-ViT-S, L, XL	5e-4
	UFO-ViT-M	4e-4
	UFO-ViT-B	3.5e-4
weight decay[27]	UFO-ViT-S, L	0.05
	UFO-ViT-M, XL	0.07
	UFO-ViT-B	0.09
drop path[21]	UFO-ViT-S, L	0.1
	UFO-ViT-M, XL	0.15
	UFO-ViT-B	0.2
grad clip[28]	UFO-ViT-S/L	1.0
	UFO-ViT-M, XL	0.7
	UFO-ViT-B	0.5

Table 3: **Hyperparameters for image classification.** All the other hyperparameters are same as DeiT[37].

Method	Top-1 Acc. (%)
Baseline(Linear Embed+XNorm)	81.8
XNorm → LN[1], GN[43]	Failed
XNorm → Learnable p-Norm	81.8
XNorm → Single L2Norm	Failed
Linear Embed[10] → Conv Embed	82.0
+Tuned Hyperparameter	82.8

Table 4: **Ablation study on ImageNet1k classification.** The results of ablation study on UFO-ViT-M. Note that single L2Norm means applying L2Norm to only one of query and key-value interaction. The learnable parameter p of p-norm is initialized by 2.

Model	Top-1 Acc	Res	Params (M)	FLOPs (G)
RegNetY-1.6G[30]	78.0	224	11	1.6
DeiT-Ti[37]	72.2	224	5	1.3
XCiT-T12/16[12]	77.1	224	26	1.2
UFO-ViT-T	78.3	224	10	1.9
ResNet-50[17]	75.3	224	26	3.8
RegNetY-4G[30]	80.0	224	21	4.0
DeiT-S[37]	79.8	224	22	4.6
Swin-T[26]	81.3	224	29	4.5
XCiT-S12/16[12]	82.0	224	26	4.8
UFO-ViT-S	81.8	224	21	3.7
ResNet-101[17]	75.3	224	47	7.6
RegNetY-8G[30]	81.7	224	39	8.0
Swin-S[26]	83.0	224	50	8.7
XCiT-S24/16[12]	82.6	224	48	9.1
UFO-ViT-M	82.8	224	37	7.0
RegNetY-16G[30]	82.9	224	84	16.0
DeiT-B[37]	81.8	224	86	17.5
Swin-B[26]	83.5	224	88	15.4
XCiT-S12/8[12]	83.4	224	26	18.9
UFO-ViT-L	83.1	224	21	14.3
EfficientNet-B7[34]	84.3	600	66	37.0
XCiT-S24/8[12]	83.9	224	48	36.0
UFO-ViT-XL	83.9	224	37	27.4

Table 5: **Comparison with the state of the art models.** Note that the properties of the other models are taken from original papers.

Experiments

Object Detection Results

❖ Object Detection on COCO

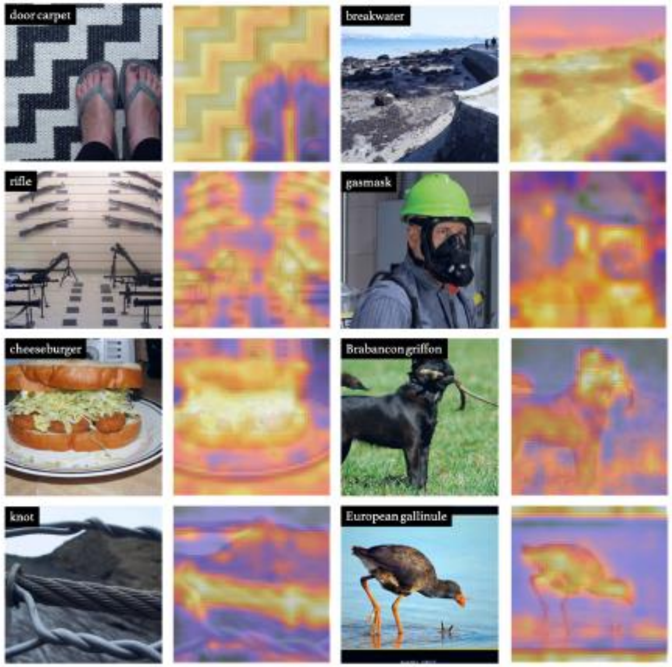


Figure 6: **Visualized UFO-module outputs.** UFO schemes cannot be visualized like traditional $N \times N$ self-attention maps. Instead, we visualize tensor length map of UFO-module output from UFO-ViT-M pretrained on ImageNet1k. All maps are extracted from first layer which gathers most of spatial information. This map is scaled at $(0, 1)$. Red means the value is close to 1.

Backbone	Params (M)	AP^b	AP^b_{50}	AP^b_{75}	AP^m	AP^m_{50}	AP^m_{75}
ResNet50[17]	44.2	41.0	61.7	44.9	37.1	58.4	40.1
PVT-Small[41]	44.1	43.0	65.3	46.9	39.9	62.5	42.8
Swin-T[26]	47.8	46.0	68.1	50.3	41.6	65.1	44.9
XCiT-S12/16[12]	44.3	45.3	67.0	49.5	40.8	**64.0**	43.8
UFO-ViT-S	39.7	44.6	66.7	48.7	40.4	63.6	42.9
ResNet101[17]	63.2	42.8	63.2	47.1	39.2	60.1	41.3
PVT-Medium[41]	63.9	44.2	66.0	48.2	40.5	63.1	43.5
Swin-S[26]	69.0	48.5	70.2	53.5	43.3	67.3	46.6
XCiT-S24/16[12]	65.8	46.5	68.0	50.9	41.8	65.2	45.0
UFO-ViT-M	56.4	46.0	68.2	50.0	41.0	64.6	43.7
ResNeXt101-64[45]	101.9	44.4	64.9	48.8	39.7	61.9	42.6
PVT-Large[41]	81.0	44.5	66.0	48.3	40.7	63.4	43.7
XCiT-M24/16[12]	101.1	46.7	68.2	51.1	42.0	65.6	44.9
UFO-ViT-B	82.4	45.8	67.4	50.1	41.2	64.5	44.1

Table 6: **Object detection performance on the COCO val2017.**

Conclusion

- Softmax 제거 및 행렬 연산 법칙을 사용하여 Self-Attention가 Linear Complexity를 가지도록 UFO-ViT를 제안
- 굉장히 간단한 방법으로 ViT에서 발생하는 Complexity와 Data-Efficiency를 해결
- Image Classification과 Object Detection Task에서 우수한 성능 달성
  - 후기: 최근 Self-Attention의 연산 복잡도를 줄이기 위해 다양한 방법론들이 제시되어 왔고 Data-Efficiency 문제와 함께 해결하였음. 연구 트렌드에 맞게 매우 간단한 연산 방법으로 Complexity 와 Efficiency를 해결하는 부분이 인상 깊었음.
    코드도 공개가 되어있었다면 이해하기가 쉬웠을 듯!

- Reference

  - Song, J. G. (2021). UFO-ViT: High Performance Linear Vision Transformer without Softmax. arXiv preprint arXiv:2109.14382.

Thank You