Towards Robust Visual Transformer Networks via K-Sparse Attention

IEEE ICASSP 2022 Paper #4604

Sajjad Amini, Shahrokh Ghaemmaghami

Electronics Research Institute
Sharif University of Technology

May 11, 2022

Table of Contents

Deep Learning Architectures [2]

Strengths

- Capable of Feature Engineering
- Unstructured data accepted
- Self-supervised Efficiency
- Multimodality

Challenges

- Data Hunger
- Loosely Interpretable
- Low Robustness
- Computational Complexity

Robustness

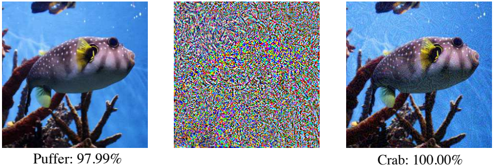
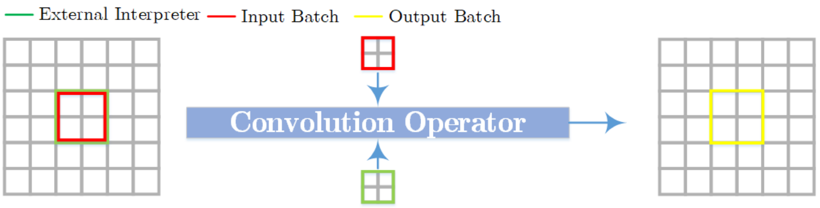
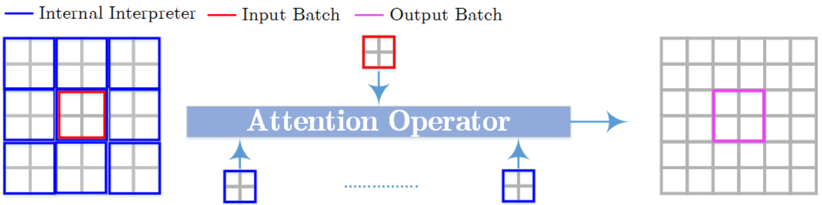
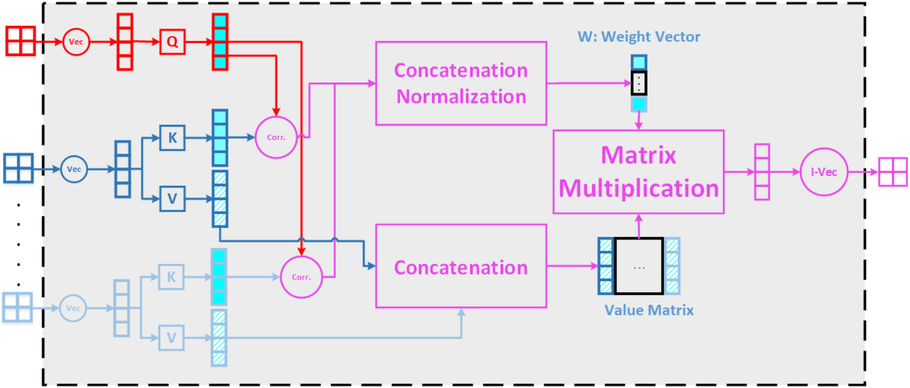


Figure: Sample Adversarial attack [1]

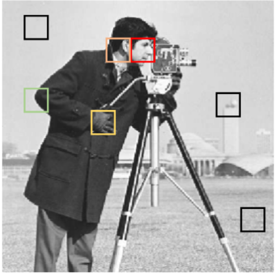
Convolution vs. Attention

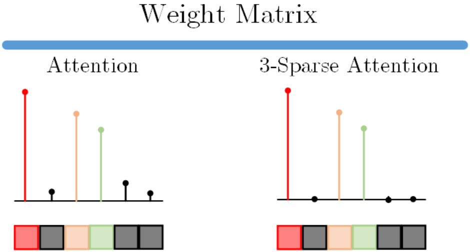





Vec
Q
K
V
Corr.
Vec
K
V
Corr.
Vec
W: Weight Vector
Concatenation
Normalization
Matrix
Multiplication
I-Vec
Concatenation
...
Value Matrix

K-Sparse Attention Justification





Justifications

- Improve accuracy by blocking the propagation of irrelevant information
- Improve robustness via blocking back-propagation through irrelevant paths

Vision Transformers (ViT) [4]

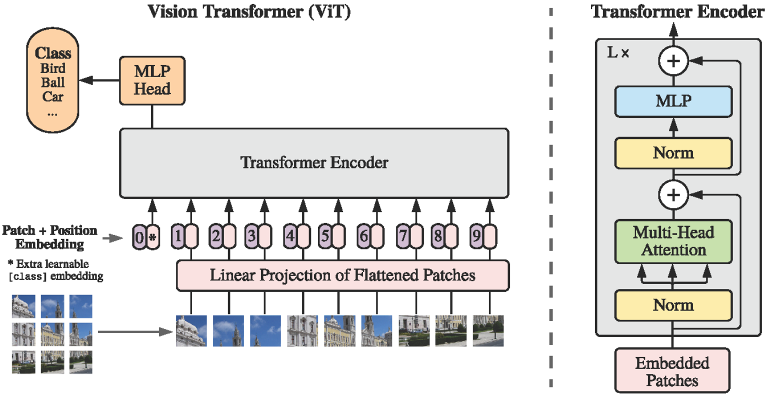


Figure: Visual Transformer Architecture (Photo from [4])

K-Sparse Attention Formulation

Basic Constrained Formulation

$$\mathcal{P} : \min_{\mathbf{p}} \sum_{j=1}^{N} D(\mathbf{y}_j, \widehat{\mathbf{y}}_j) \ w.r.t \ \|\mathbf{w}_{i,l}^j\|_0 \leq K_{i,l}^j, \begin{cases} 0 \leq i \leq I_l \\ l \in \mathcal{S} \end{cases}$$

where:

$\mathbf{p}$	Vector of transformer parameters
N, j	Number of training samples, Training set index
$\mathbf{y}_j$, $\widehat{\mathbf{y}}_j$	Network and target output for j-th sample
i, l	Sequence position, layer index
$\mathbf{w}_{i,l}^j$, $K_{i,l}^j$	Weight vector and corresponding sparsity level
I_l	Sequence length for l-th layer
$\mathcal{S}$	Regularized attention module set

K-Sparse Attention Formulation

Unconstrained Formulation

$$\min_{\mathbf{p}} \sum_{j=1}^{N} \left[D(\mathbf{y}_j, \widehat{\mathbf{y}}_j) + \sum_{l \in \mathcal{S}} \sum_{i=1}^{I_l} \mathcal{I} \left\{ \|\mathbf{w}_{i,l}^{j}\|_0 \leq K_{i,l}^{j} \right\} \right]$$

where:

$$\mathcal{I}\left\{\|\mathbf{x}\|_0 \leq \delta\right\} = \begin{cases} 0 & \textit{if} \quad \|\mathbf{x}\|_0 \leq \delta \\ \infty & \textit{if} \quad \|\mathbf{x}\|_0 > \delta \end{cases}$$

K-Sparse Attention Formulation

Using Penalty method [5]

$$\mathcal{P}_\mu : \min_{\mathbf{p},\{\mathbf{s}^j_{i,l}\}} \sum_{j=1}^{N} \left[D(\mathbf{y}_j, \widehat{\mathbf{y}}_j) + \sum_{l\in\mathcal{S}} \sum_{i=0}^{I_l} \left(\mathcal{I}\left\{ \|\mathbf{s}^j_{i,l}\|_0 \le K^j_{i,l} \right\} + \frac{1}{2\mu^j_{i,l}} \|\mathbf{s}^j_{i,l} - \mathbf{w}^j_{i,l}\|_2^2 \right) \right]$$

For $\mu^j_{i,l} \to 0$, $\mathcal{P}_\mu$ can approximate $\mathcal{P}$.

Using proximal mapping:

$$\mathbf{s}(k+1) = \arg\min_{\mathbf{s}} \ \mathcal{I}\{\|\mathbf{s}\|_0 \le K\} + \frac{1}{2\mu}\|\mathbf{s} - \mathbf{w}(k)\|_2^2 = Prox_{\mathcal{I}}(\mathbf{w}(k)) = [\mathbf{w}(k)]_K$$

Using gradient based optimization methods:

$$\mathbf{p}(k+1) = \arg\min_{\mathbf{p}} \ \sum_{j=1}^{N} \left[D(\mathbf{y}_i, \widehat{\mathbf{y}}_i) + \sum_{l\in\mathcal{S}} \sum_{0\le i\le I_l} \frac{1}{2\mu^j_{i,l}} \|\mathbf{s}^j_{i,l}(k+1) - \mathbf{w}^j_{i,l}\|_2^2 \right]$$

Final Algorithm

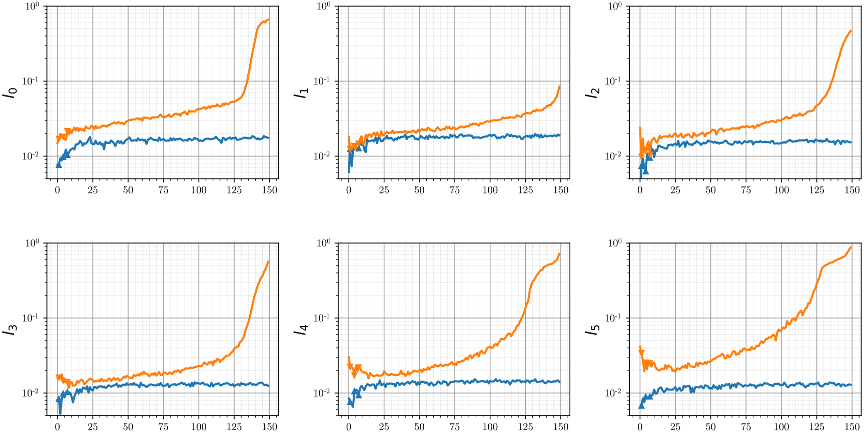
Algorithm pseudocode for the calculation of

Input: Training patterns $(\{\mathbf{X}_i, \mathbf{y}_i\})$, N_1, N_2, c, μ.
Output: Network parameters vector $\mathbf{p}_{final}$

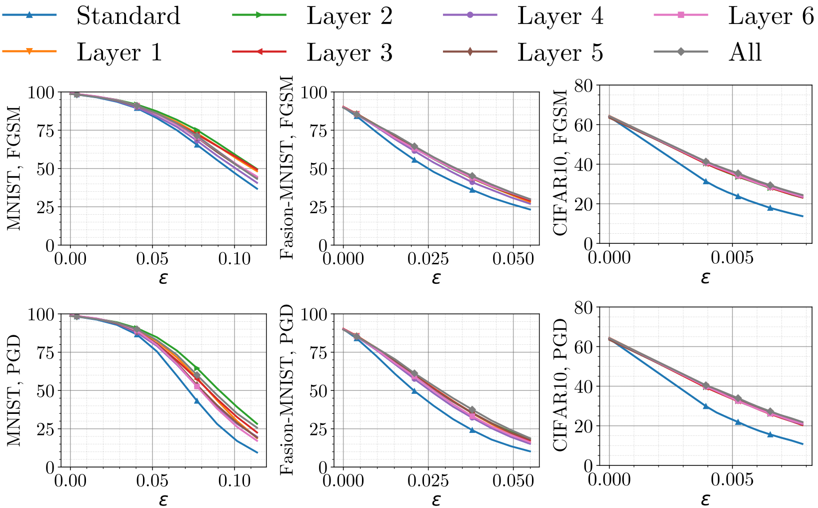
$\mathbf{p}_0$, $k = 0$, $m = 0$
while $m \leq N_1$ **do**
  while $k \leq N_2$ **do**
    $\mathbf{s}^j_{i,l}(k+1) = [w^j_{i,l}(k)]_{K^j_{i,l}}$, for i, l and j
    $\mathbf{p}(k+1) = \arg\min_{\mathbf{p}} \sum_{j=1}^{N} \left[D(\mathbf{y}_i, \widehat{\mathbf{y}}_i) + \sum_{l \in \mathcal{S}} \sum_{0 \leq i \leq I_l} \frac{1}{2\mu^j_{i,l}} \|\mathbf{s}^j_{i,l}(k+1) - \mathbf{w}^j_{i,l}\|_2^2 \right]$
    $k \leftarrow k + 1$
  end while
  $\mu \leftarrow c \cdot \mu$
  $m \leftarrow m + 1$
  $\mathbf{p}_0 \leftarrow \mathbf{p}$
  $k = 0$
end while
$\mathbf{p}_{final} \leftarrow \mathbf{p}$

Sparsity Comparison

Hoyer measure vs. Epochs (Blue: ViT - Red: KSA-ViT)



Untargeted Adversarial Attacks



Targeted Adversarial Attacks

Type	CW-L2				CW-Linf			
	ASR	L_1	L_2	L_∞	ASR	L_1	L_2	L_∞
Satndard	0.62	27.29	0.74	0.08	,0.81	,37.70	,0.81	,0.033
Layer 1	0.60	32.71	0.88	0.09	,0.78	,46.08	,0.97	,0.036
Layer 2	0.57	32.38	0.87	0.09	,0.76	,45.82	,0.96	,0.036
Layer 3	0.61	30.91	0.83	0.09	,0.76	,46.13	,0.97	,0.036
Layer 4	0.58	33.46	0.90	**0.09**	,0.79	,46.14	,0.97	,0.036
Layer 5	0.56	33.46	0.90	0.09	,0.77	,48.06	,1.00	,0.036
Layer 6	0.58	32.95	0.88	0.09	,0.77	,46.98	,0.98	,0.036
All	**0.56**	**34.83**	**0.93**	0.09	,**0.75**	,**49.51**	,**1.03**	,**0.036**

Conclusions

Dense weight vector in the attention module

- Lower the generalization of architecture
- Provide space for adversarial attacks

K-Sparse attention

- Formulation Based on ℓ_0 norm regularizer
- Solve the problem using penalty method
- Limit the weight matrix in an unstructured manner
- Improve the generalization performance
- Improve adversarial robustness

References I

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li,
"Boosting adversarial attacks with momentum,"
in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9185–9193.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton,
"Deep learning,"
Nature, vol. 521, no. 7553, pp. 436–444, 2015.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin,
"Attention is all you need,"
in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

References II

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al.,
"An image is worth 16x16 words: Transformers for image recognition at scale,"
arXiv preprint arXiv:2010.11929, 2020.

Sajjad Amini and Shahrokh Ghaemmaghami,
"A new framework to train autoencoders through non-smooth regularization,"
IEEE Transactions on Signal Processing, vol. 67, no. 7, pp. 1860–1874, 2019.