Rényi Attention Entropy for Patch Pruning

Hiroaki Aizawa* and Yuki Igaue*

Hiroshima University, Japan

* Equal contribution

Vision Transformer and patch attention

A Vision Transformer (ViT) divides an image into patches, embeds each patch as a token, and processes the token sequence using Transformer blocks. Multi-head self-attention models the relations among all patch tokens.

For patch-token matrix $\boldsymbol{X}_{\text{patches}}$, the softmax yields the attention distribution of patch $\boldsymbol{x}_i$:

$$\underbrace{\boldsymbol{a}(\boldsymbol{x}_i)}_{\text{distribution over patches}} := \underbrace{\sigma\left(\frac{\boldsymbol{X}_{\text{patches}}^{\top}\boldsymbol{W}_K\boldsymbol{W}_Q^{\top}\boldsymbol{x}_i}{\sqrt{d'}}\right)}_{\text{softmax of query--key similarities}}$$

$$\boldsymbol{a}(\boldsymbol{x}_i) = [a_1(\boldsymbol{x}_i), \ldots, a_{n-1}(\boldsymbol{x}_i)]^{\top}, \qquad \sum_{j=1}^{n-1} a_j(\boldsymbol{x}_i) = 1$$

- Computing all query-key relations produces a cost quadratic in the number of tokens.
- **Patch pruning** reduces this cost by removing redundant tokens at early blocks.

Related Work: Limitations of attention magnitude

- Attention-magnitude methods such as EViT use $a_{i,j}$ to estimate patch importance, but attention magnitude does not directly quantify how concentrated the complete distribution is, i.e., the model's certainty across patches.
- A large attention weight does not always imply a large contribution to the self-attention output because the output also depends on the value vector:

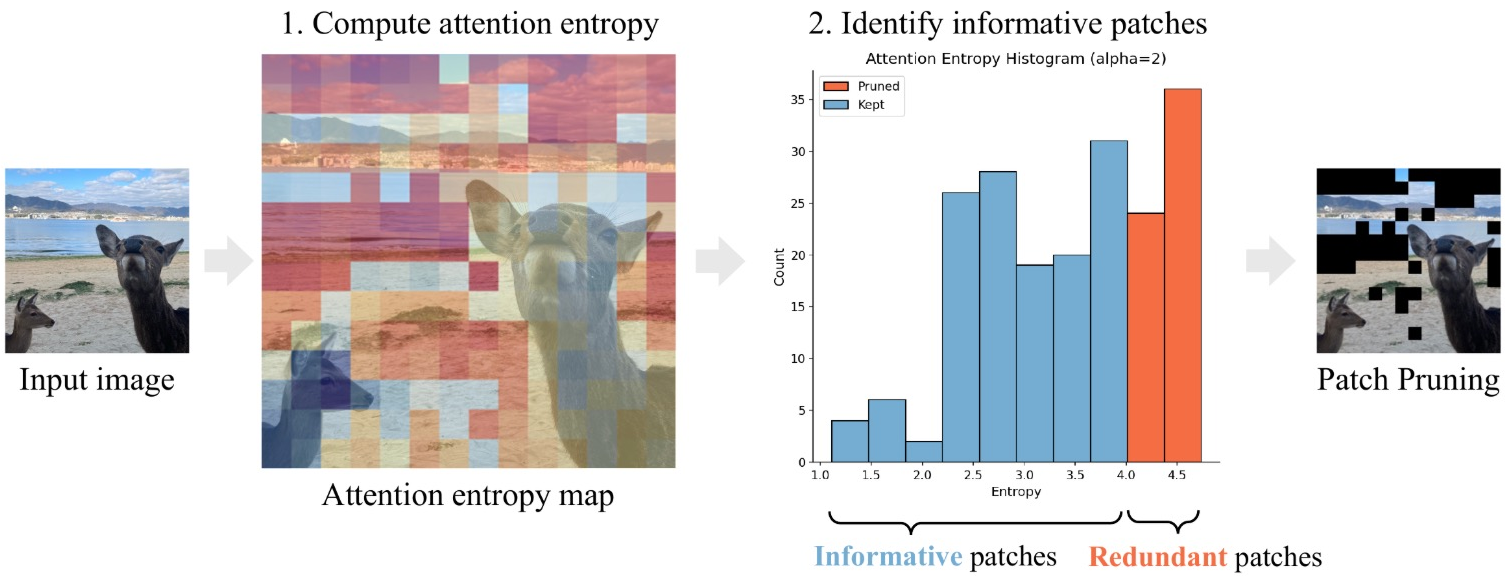
$$\boldsymbol{x}_i' = \sum_{j=1}^{n} \underbrace{a_{i,j}}_{\text{attention weight}} \cdot \underbrace{v(\boldsymbol{x}_j)}_{\text{value vector}}$$

- Even when $a_{i,j}$ is large, patch j contributes little if $v(\boldsymbol{x}_j)$ is small. Magnitude-based pruning may therefore retain redundant patches or mistakenly remove useful ones.

Our idea: use attention entropy to quantify concentration and score patch importance.

Proposal 1: Identify patches with attention entropy

We introduce attention entropy as a patch-importance criterion.



- Measure the concentration of each patch's complete attention distribution and rank patches by entropy.
- Lower entropy indicates higher importance; no additional importance-prediction module is needed.

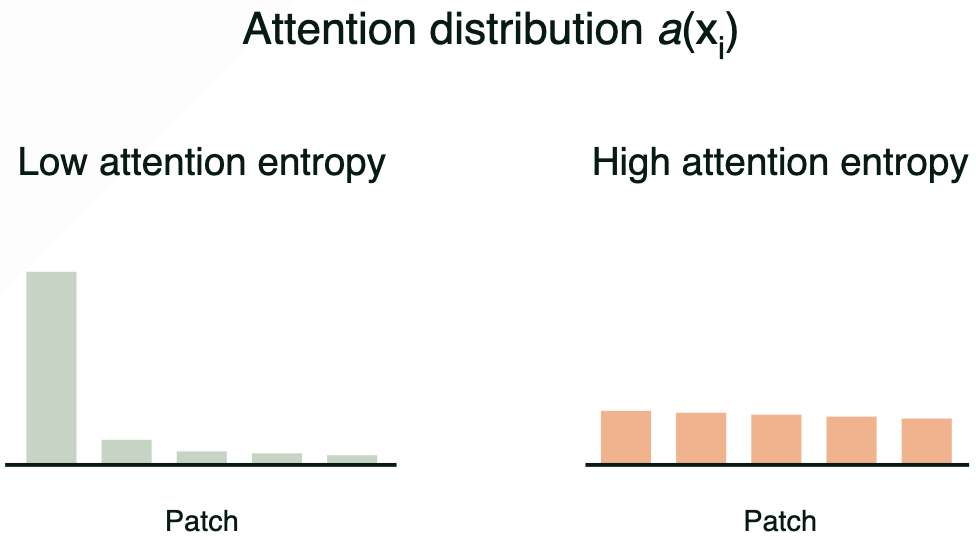
Low entropy marks concentrated, informative attention

For the attention distribution $\boldsymbol{a}(\boldsymbol{x}_i)$:

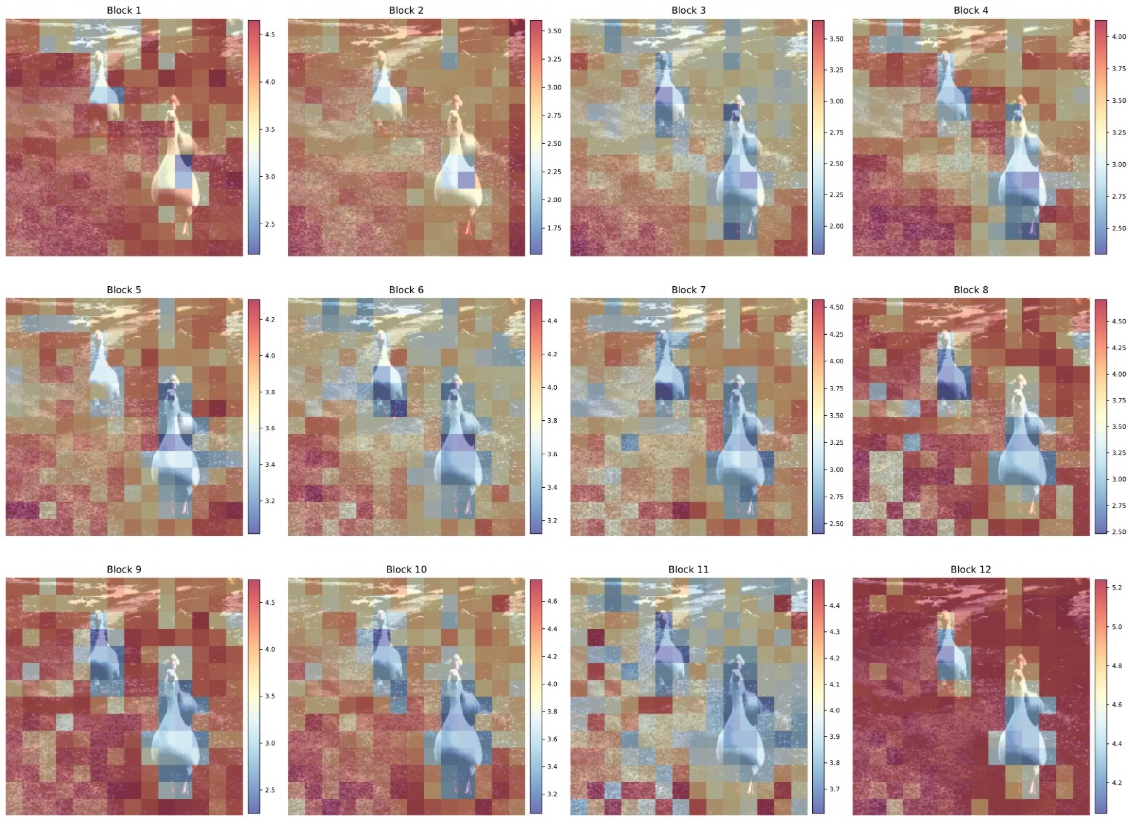
$$\mathcal{H}(\boldsymbol{x}_i) := -\sum_{j=1}^{n-1} a_j(\boldsymbol{x}_i) \log a_j(\boldsymbol{x}_i)$$

Low entropy: attention is concentrated; retain as informative.

High entropy: attention is dispersed; treat as a removal candidate.



Low entropy tracks objects across layers



Low entropy often lies on the foreground object
→ informative patch

High entropy often lies in the background
→ pruning candidate

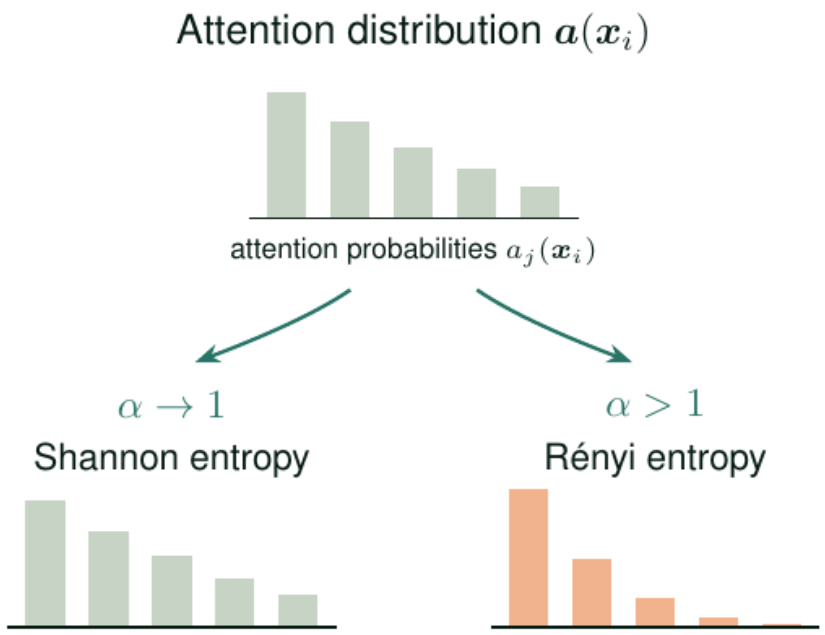
Proposal 2: Rényi entropy tunes peak emphasis

We extend the criterion from Shannon entropy to Rényi entropy.

Rényi entropy generalizes Shannon entropy by introducing an order parameter α that controls peak emphasis.

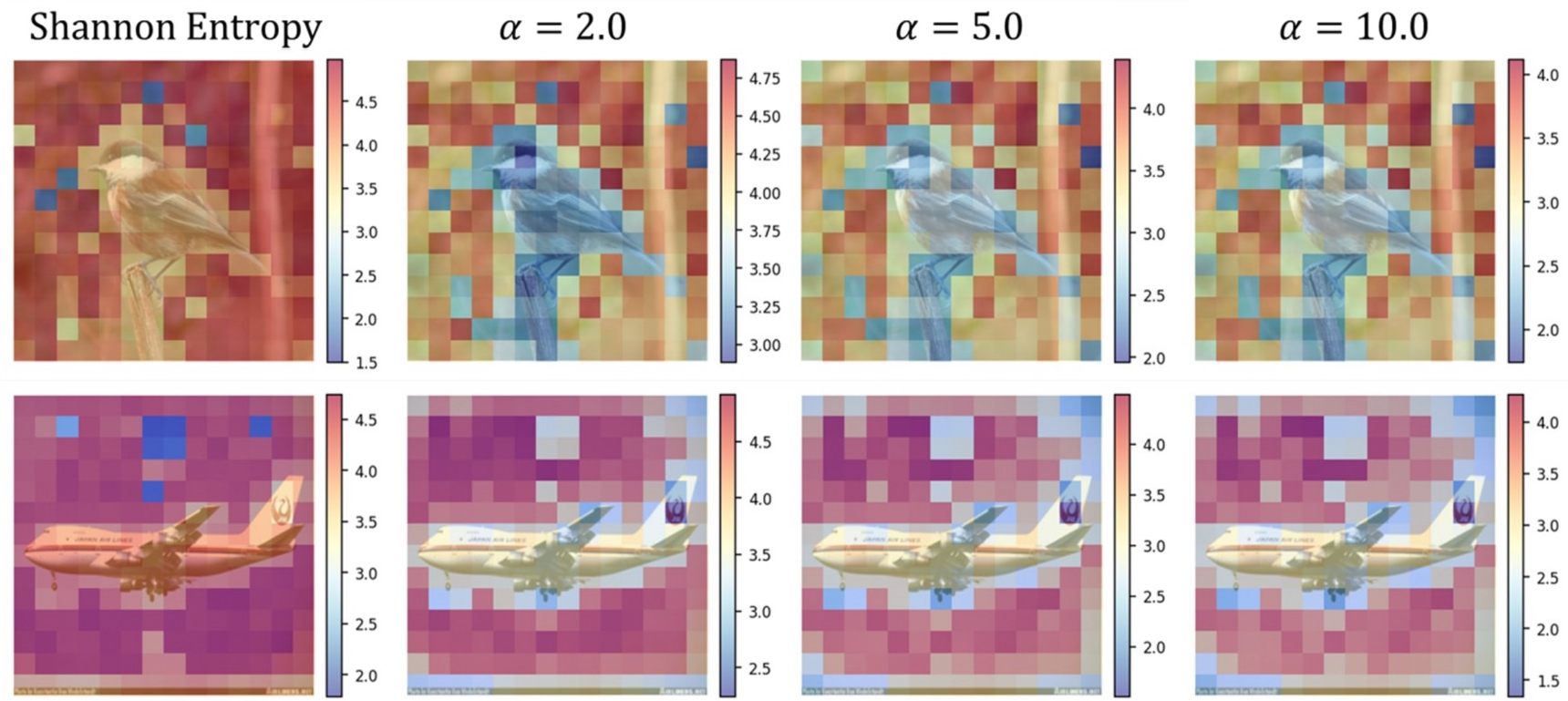
$$\mathcal{H}_\alpha(\boldsymbol{x}_i) := \frac{1}{1-\alpha}\log\sum_{j=1}^{n-1}\underbrace{a_j(\boldsymbol{x}_i)^\alpha}_{\text{peak emphasis controlled by }\alpha}$$

- $\alpha \to 1$: Rényi entropy converges to Shannon entropy.
- $\alpha > 1$: higher-probability events receive more emphasis, so dominant peaks have a stronger effect on the score.



Schematic effect of the Rényi order

Rényi order can change patch ranking

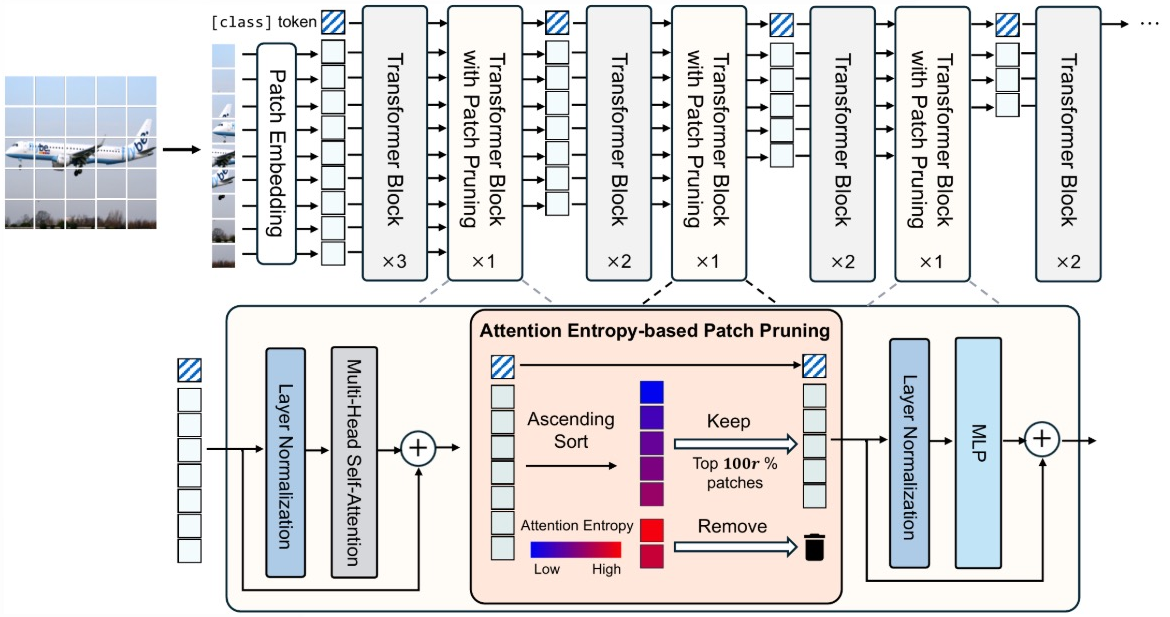


Larger α shifts entropy scores toward lower values

Re-ranking can change which patches are retained at the same keep rate

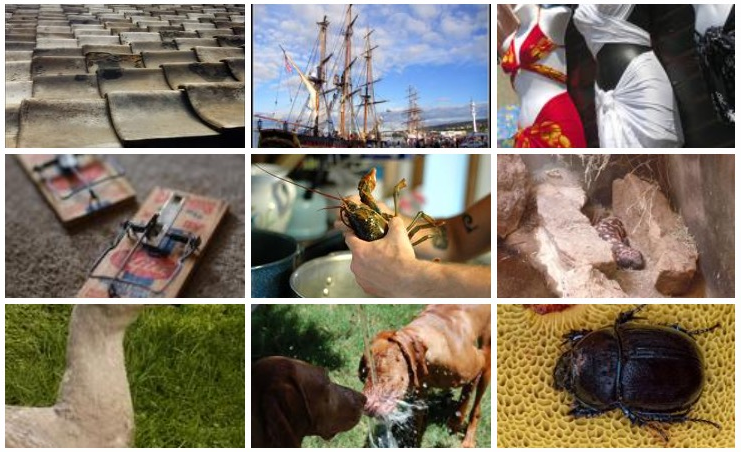
Our method: Rényi attention entropy pruning

We apply the proposed entropy criterion directly to patch pruning.



- Compute Shannon or Rényi attention entropy per head and average across heads.
- Sort entropy scores; retain low-entropy patches and the class token.
- Propagate selected tokens and repeat at Blocks 4, 7, and 10.

Overall experimental setup



ImageNet-100

FGVC Aircraft

Oxford Flowers102

We evaluate general image classification on **ImageNet-100** and fine-grained classification on **FGVC Aircraft** and **Oxford Flowers102**. For every dataset, we fine-tune an ImageNet-1k-pretrained DeiT-S for 100 epochs with AdamW and cross-entropy loss. Inputs are resized to 224 × 224 with a batch size of 256; Mixup, CutMix, and RandomErasing are applied during training.

Pruning settings

Item	Setting
Pruning blocks	4, 7, and 10
Keep rate	$r \in \{0.9, 0.8, \ldots, 0.3\}$
Entropy	Shannon ($\alpha = 1.0$); Rényi $\alpha \in \{2, 5, 10\}$
Head aggregation	Average entropy across attention heads
Selection	Retain the lowest-entropy $100r\%$ of patch tokens
Special token	Always retain the class token
Baseline	EViT under the same keep rate and pruning blocks

The keep rate is fixed before inference; the entropy score itself needs no additional training.

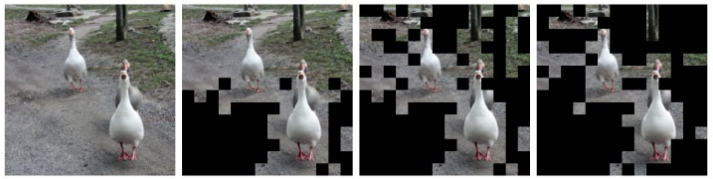
EViT serves as the **attention-magnitude baseline** under the same keep rate and pruning blocks.

ImageNet-100 classification results

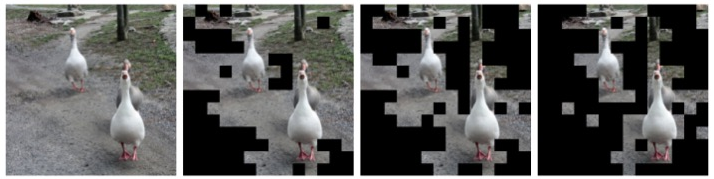
Unpruned reference: DeiT-S (r = 1.0), 86.74% Top-1

Method	$r = .9$	$r = .8$	$r = .7$	$r = .6$	$r = .5$	$r = .4$	$r = .3$
EViT	86.06	86.04	86.46	85.68	85.32	84.50	83.92
Shannon ($\alpha = 1.0$)	**86.40**	85.74	86.14	85.66	85.54	84.70	84.16
Rényi $\alpha = 2$	86.34	**86.46**	85.60	85.50	85.36	84.78	**84.30**
Rényi $\alpha = 5$	85.88	86.20	86.30	**85.78**	85.44	**85.18**	84.00
Rényi $\alpha = 10$	86.26	86.02	**86.72**	85.72	**85.68**	84.82	84.20

Each sequence: Input → Block 4 → Block 7 → Block 10 · Black regions are pruned patches



EViT



Ours, Rényi α = 10

Shannon is effective under mild pruning; higher-order Rényi entropy is stronger at moderate keep rates.

FGVC Aircraft classification results

Unpruned reference: DeiT-S (r = 1.0), 79.72% Top-1

Method	$r = .9$	$r = .8$	$r = .7$	$r = .6$	$r = .5$	$r = .4$	$r = .3$
EViT	71.29	72.10	78.88	76.30	78.82	73.69	76.96
Shannon ($\alpha = 1.0$)	77.53	**77.71**	72.91	76.06	68.44	74.68	64.45
Rényi $\alpha = 2$	71.56	67.54	77.11	71.50	73.87	80.02	70.15
Rényi $\alpha = 5$	78.94	76.75	**80.14**	77.77	78.16	**80.29**	74.26
Rényi $\alpha = 10$	**79.84**	74.32	78.55	**78.70**	**80.29**	77.38	**77.08**

EViT

Ours, Rényi α = 5

Higher-order Rényi entropy is consistently effective, aligning with the need to emphasize sharp, part-level cues for fine-grained recognition.

Oxford Flowers102 classification results

Unpruned reference: DeiT-S (r = 1.0), 89.04% Top-1

Method	$r = .9$	$r = .8$	$r = .7$	$r = .6$	$r = .5$	$r = .4$	$r = .3$
EViT	90.58	88.47	**90.81**	87.71	**89.67**	**88.81**	**86.19**
Shannon ($\alpha = 1.0$)	90.23	84.99	88.91	87.64	86.86	86.39	84.52
Rényi $\alpha = 2$	89.64	88.26	90.63	88.49	88.76	87.04	83.61
Rényi $\alpha = 5$	**90.71**	**90.14**	90.45	**89.15**	88.32	87.10	84.00
Rényi $\alpha = 10$	88.37	90.11	89.75	87.93	87.97	86.21	67.26

EViT

Ours, Rényi α = 2

Rényi entropy is competitive under mild and moderate pruning, while EViT is stronger under aggressive pruning; the order must be adapted to the dataset and keep rate.

Computational cost comparison

Method	r	Top-1 (Δ)	GFLOPs (Δ)	images/s (Δ)
DeiT-S	1.0	86.74	4.61	1111.4
EViT	0.7	86.46 (-0.28)	3.04 (-34.1%)	1947.1 (+75.2%)
	0.5	85.32 (-1.42)	2.31 (-49.8%)	2535.8 (+128.2%)
	0.3	83.92 (-2.82)	1.82 (-60.5%)	3150.1 (+183.4%)
Ours	0.7	**86.72 (-0.02)**	**3.00 (-34.8%)**	1730.3 (+55.7%)
	0.5	**86.68 (-0.06)**	**2.29 (-50.4%)**	2257.5 (+103.1%)
	0.3	**84.30 (-2.44)**	**1.80 (-60.8%)**	2790.2 (+151.1%)

All measurements use ImageNet-100 and one NVIDIA RTX A6000. The deltas are relative to DeiT-S without pruning.

Ours improves the accuracy–GFLOPs trade-off over EViT; throughput improves over DeiT-S but remains below EViT.

Discussion: Interpretation and current limitations

Why entropy helps
It evaluates the complete attention distribution instead of relying on a single large attention weight.

Why the Rényi order is tuned
The appropriate peak emphasis depends on the dataset and keep rate; no single order is universally optimal.

Current limitation
The order and keep rate are fixed before inference. Adapting them to each block and input remains future work.

Rényi entropy makes peak emphasis controllable, but α remains task- and budget-dependent.

Summary: From attention entropy to patch pruning

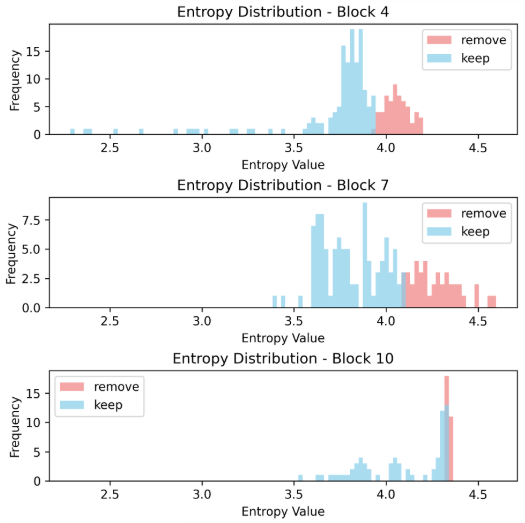
- We introduce attention entropy to identify informative patches from complete attention distributions.
- Low and high entropy often align with foreground and background regions, respectively.
- We extend the criterion to Rényi entropy; order α can change patch ranking for the task and keep rate.
- We apply the criterion to patch pruning and obtain a favorable accuracy–GFLOPs trade-off.
- The score uses one forward pass, needs no additional training, and integrates with existing ViTs.

Interested in this work?

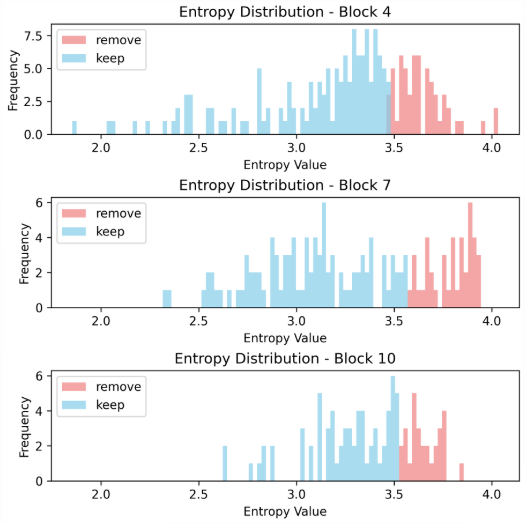
Read the preprint: arxiv.org/abs/2604.03803

Please feel free to email me: hiroaki-aizawa@hiroshima-u.ac.jp

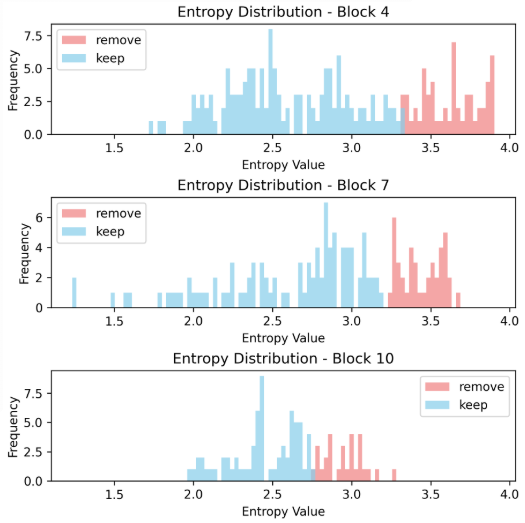
Rényi order shifts entropy distributions



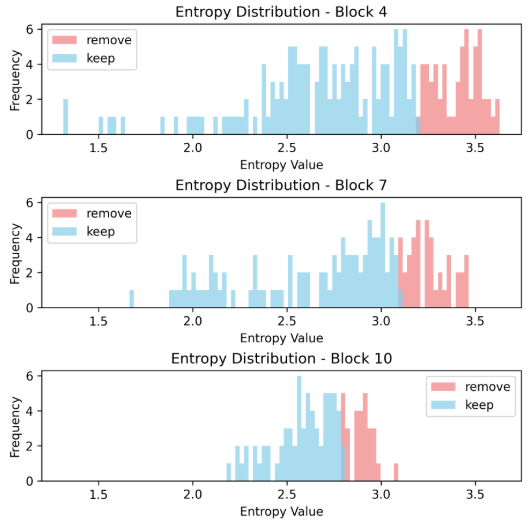
Shannon, α = 1.0



Rényi, α = 2



Rényi, α = 5



Rényi, α = 10

Increasing α shifts the entropy distribution toward lower values and can re-rank tokens under the same keep rate. The distribution also changes with block depth.

How should the Rényi order be selected?

Step	Practical rule
Candidates	Evaluate Shannon ($\alpha = 1$) and Rényi $\alpha \in \{2, 5, 10\}$.
Match the budget	Compare candidates at the target keep rate.
Select	Choose α on held-out validation data for the target dataset and keep rate.
Deploy	Fix the selected α during inference; no additional predictor is required.

The current method treats α as a task- and budget-dependent hyperparameter.

Dynamic selection for each block and input remains future work.

Scaling from DeiT-S to DeiT-B

Dataset	S-base	S-ours	α_S	B-base	B-EViT	B-ours	α_B
ImageNet-100	86.74	86.72	$\alpha = 10$	86.82	86.36	**86.44**	$\alpha = 2$
FGVC Aircraft	79.72	80.14	$\alpha = 5$	81.70	74.98	**83.32**	Shannon
Flowers102	89.04	90.63	$\alpha = 2$	93.10	91.88	**92.39**	Shannon

Results at $r = 0.7$. Scaling to DeiT-B gives the clearest gain on Aircraft, where Ours reaches 83.32.

Why can GFLOPs and throughput differ?

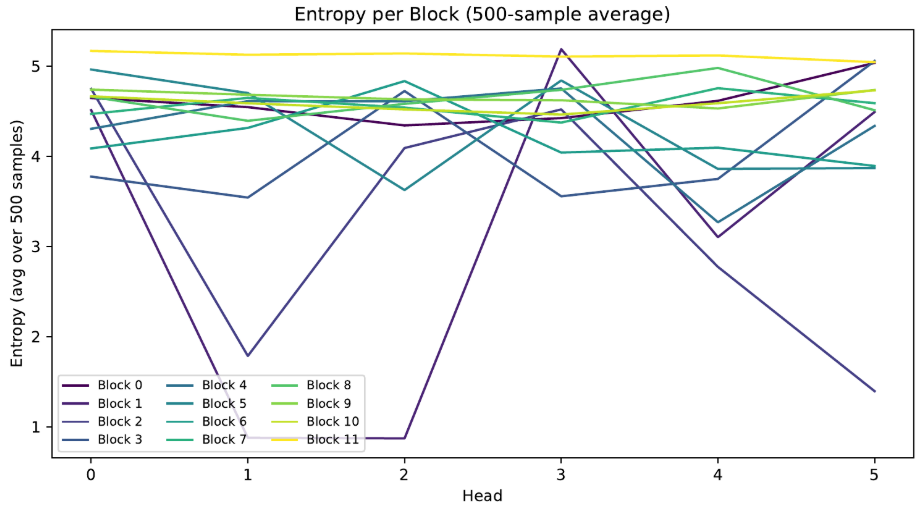
GFLOPs	Counts arithmetic operations, but omits memory traffic, kernel launches, and hardware utilization.
images/s	Measures end-to-end runtime and depends on the GPU and implementation.

Entropy scoring adds logarithms or powers and reductions whose runtime cost is not fully represented by GFLOPs.

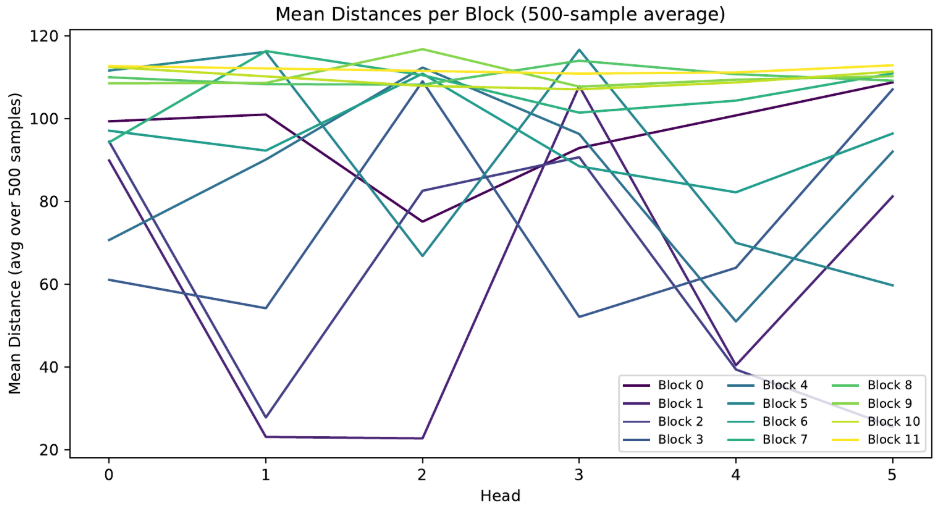
Our main gain is a better accuracy–GFLOPs trade-off, not maximum throughput.

Interpretation only: kernel-level profiling was not performed.

Entropy and attention distance



Attention entropy



Attention distance

Both measures correlate with head diversity and block depth. Curves show averages over 500 samples.

What may shape attention entropy?

Hypothesis: foreground size and spatial-frequency structure were not isolated in the current experiments.

Foreground size

Small or localized foreground → attention can concentrate on fewer related patches → lower entropy

Large or distributed foreground → attention may spread across more object patches → higher entropy

Spatial frequency

Distinctive edge or part → sharp patch similarity → lower entropy

Repeated texture → many similar candidate patches → dispersed attention and higher entropy

Attention entropy is not a direct foreground detector; it also reflects spatial extent and feature redundancy.

Future test: correlate entropy with foreground-area ratio and local spectral energy while controlling for block depth.

Patch scoring and propagation

Stage	Implementation
Distribution	Softmax yields $\boldsymbol{a}(\boldsymbol{x}_i)$ for each patch and head.
Score	Compute Shannon or Rényi attention entropy.
Aggregate	Average head-wise entropy into a per-layer score.
Select	Sort scores; retain low-entropy patches and the class token.
Propagate	Send selected tokens to the next layer and repeat.

The method uses a single forward pass, requires no additional training, and integrates with existing ViTs.

EViT uses class attention to reorganize tokens

Importance score	Average the attention weights from the class token to image tokens across heads.
Token reorganization	Preserve the top-ranked attentive tokens and fuse the remaining tokens into one representative token using attention-weighted aggregation.
Computation	Process fewer tokens in subsequent MHSA and FFN modules without introducing additional parameters.

In our experiments, EViT is the attention-magnitude baseline under matched keep rates and pruning blocks.

Liang et al., EViT, ICLR 2022.

Token merging and entropy-guided future work

General idea	Merge redundant tokens instead of removing them, retaining their information in a smaller token set.
ToMe	Use lightweight bipartite matching based on key similarity, then combine matched token features with token-size-weighted averaging.
Deployment	Apply token merging to existing ViTs with or without additional training.

Future work: use entropy to decide which tokens remain as informative anchors and which tokens should be merged.

Low-entropy tokens could remain as anchors, while high-entropy tokens could be merged into similar anchors selected by key similarity.

Not evaluated in this work. Reference: Bolya et al., ToMe, ICLR 2023.