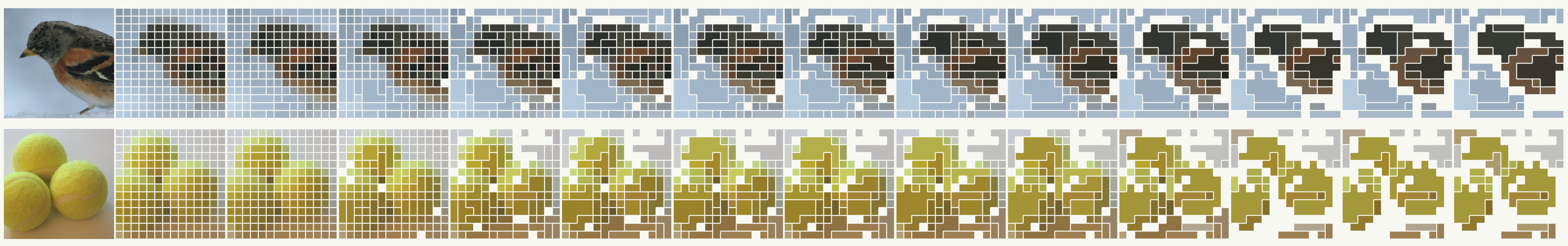
Learned Threshold Token Merging and Pruning for Vision Transformers

AUTHORS
Maxim Bonnaerens
Joni Dambre

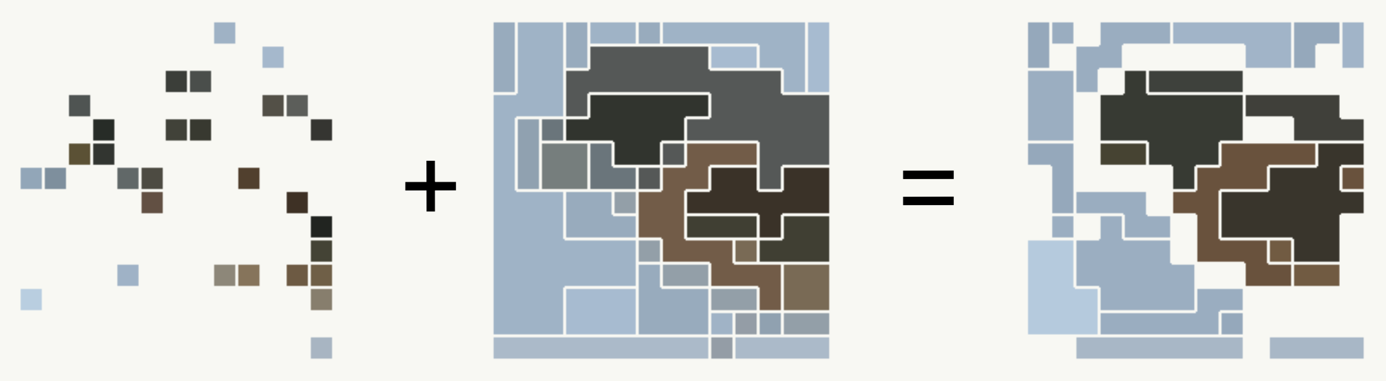
AFFILIATIONS
IDLab - AIRO
Ghent University - imec



SUMMARY

Learned Thresholds token Merging and Pruning (**LTMP**) for Vision Transformers makes it possible to **reduce the computational cost of ViTs to any reduction target** with minimal loss in accuracy. **LTMP adaptively merges similar tokens and prunes unimportant ones**. Our implementation uses **learned thresholds** which enable different merging and pruning rates per image and allows the model to learn the optimal trade-off between merging and pruning across layers. As LTMP only introduces two learnable parameters per transformer block, our method is able to **converge within a single epoch**, which is an order of magnitude quicker than other approaches.

TL;DR



LEARNED THRESHOLDS

Given importance scores, learned threshold masking modules learn which tokens to keep.

$$M(\mathbf{s}_i^l,\theta^l) = \begin{cases}1, & \text{if } \mathbf{s}_i^l > \theta^l\\ 0, & \text{otherwise}\end{cases} \qquad M(\mathbf{s}_i^l,\theta^l) = \sigma\left(\frac{\mathbf{s}_i^l-\theta^l}{\tau}\right)$$

$\xrightarrow{\hspace{2cm}}$ Forward pass $\qquad \xleftarrow{\hspace{2cm}}$ Backward pass

Construct pruning masks:

$$\mathbf{m}_i^l = \begin{cases}M(\mathbf{s}_i^l,\theta^l), & \text{if } \mathbf{m}_i^{l-1} = 1\\ \mathbf{m}^{l-1}, & \text{otherwise}\end{cases}$$

Modify attention function:

$$\text{Attention\_with\_mask}(\mathbf{Q},\mathbf{K},\mathbf{V},\mathbf{m}) = \mathbf{SV}$$

where,

$$\mathbf{S}_{ij} = \frac{\exp(\mathbf{A}_{ij})\mathbf{m}_j}{\sum_{k=1}^N \exp(\mathbf{A}_{ik})\mathbf{m}_k}, 1 \le i,j,k \le n$$

$$\mathbf{A} = \mathbf{QK}^T/\sqrt{d_k} \in \mathbb{R}^{n\times n}$$

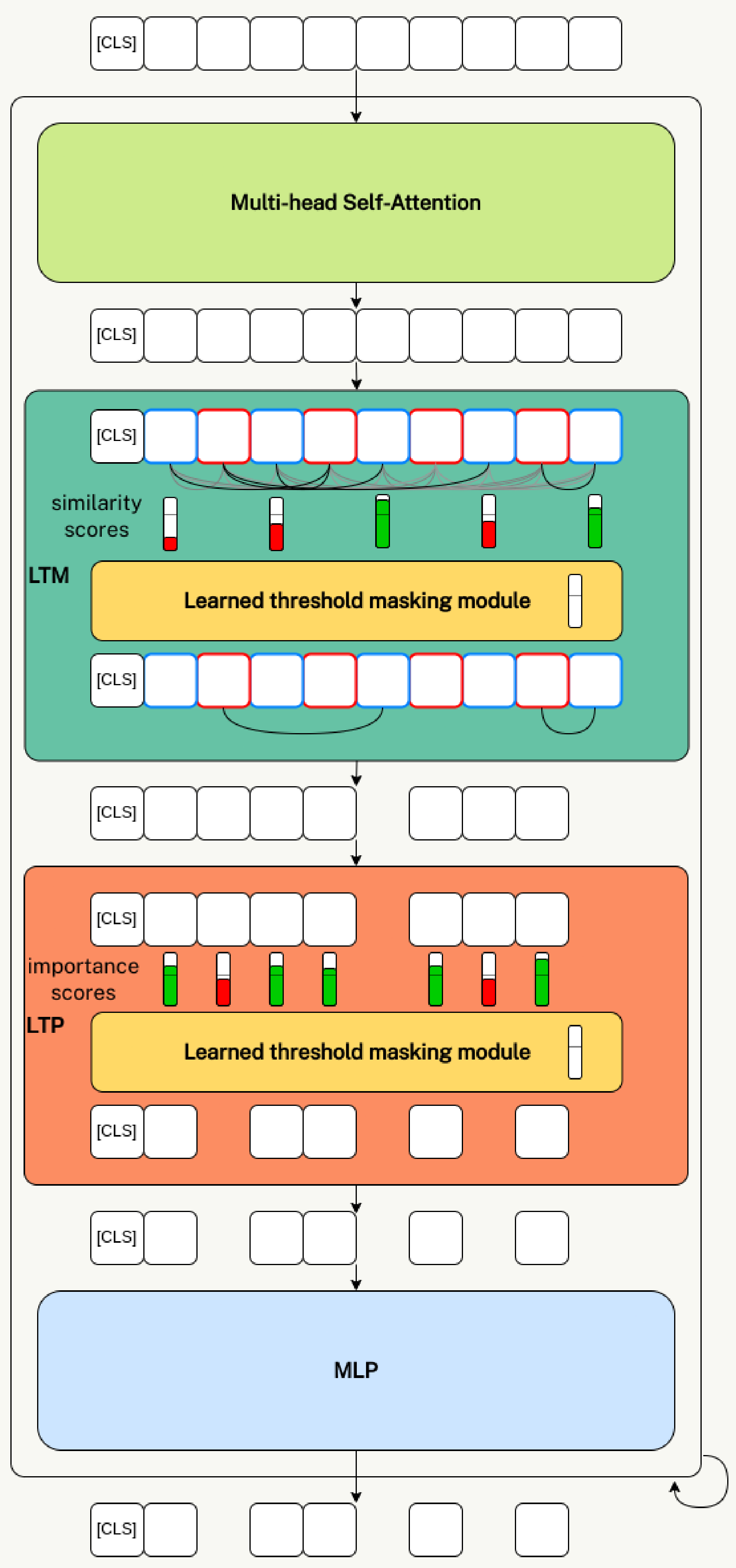
TRAINING OBJECTIVE

We optimize the thresholds using a novel budget-aware training loss for which we introduce a reduction target and an actual FLOPs reduction factor.
This allows us to create models of any size and allows the model to freely distribute the reduction operations across layers.

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda\overbrace{(r_{\text{target}} - r_{\text{FLOPs}})^2}^{\mathcal{L}_{\text{reg}}}$$

$$r_{\text{FLOPs}} \approx \sum_{l=1}^{L} \frac{1}{L}\left(\frac{2\bar{\mathbf{m}}^{l-1}nd^2+(\bar{\mathbf{m}}^{l-1}n)^2d+4\bar{\mathbf{m}}^{l}nd^2}{6nd^2+n^2d}\right)$$

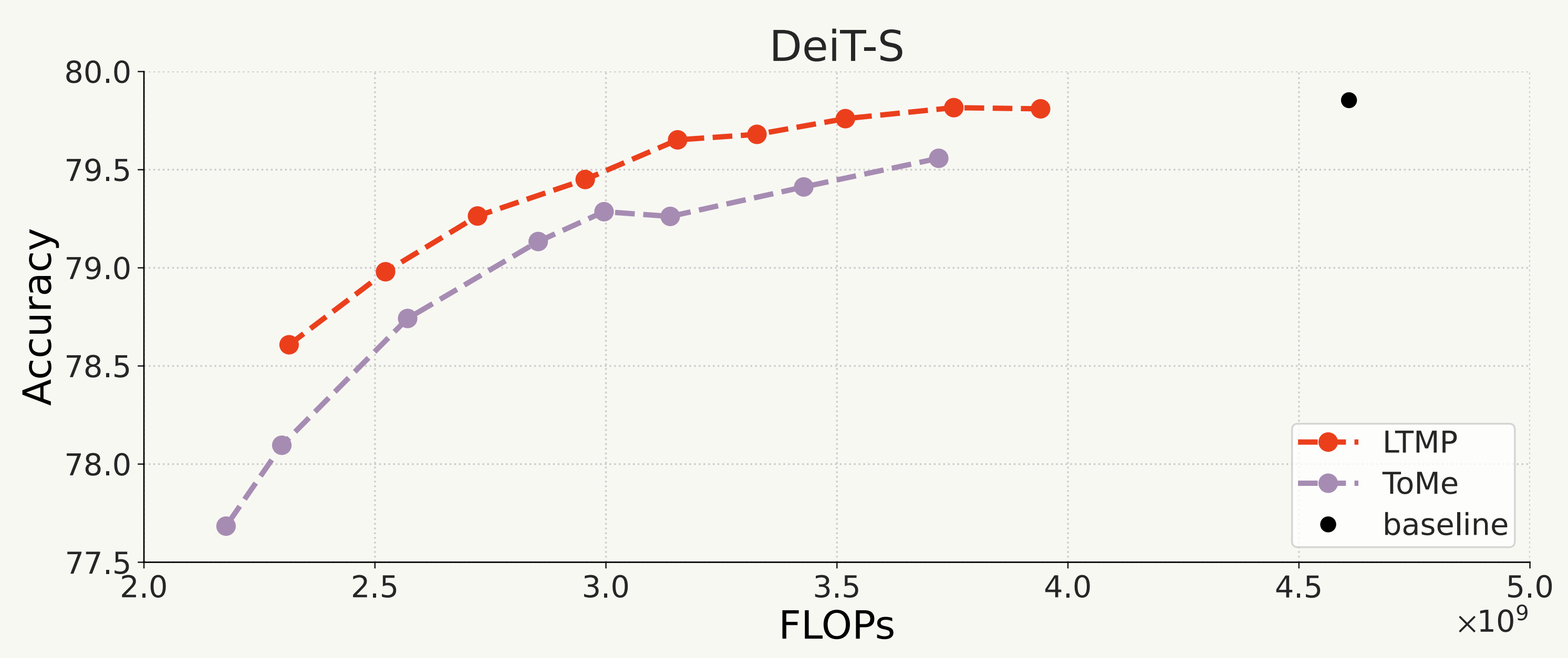
FRAMEWORK

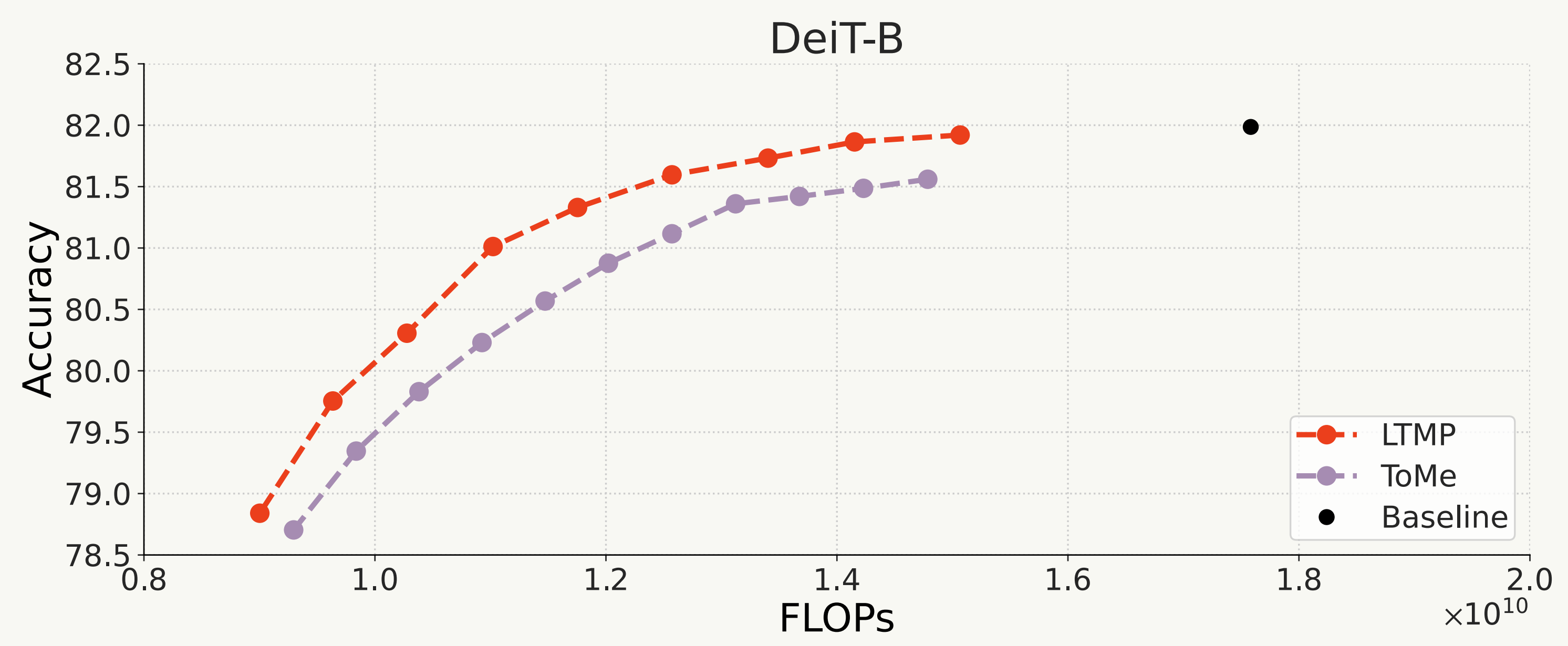

[CLS]
Multi-head Self-Attention
similarity scores
LTM
Learned threshold masking module
importance scores
LTP
Learned threshold masking module
MLP

RESULTS

Method	FLOPs	Latency*	Accuracy	Fine-tune epochs
DeiT-S (Baseline)	4.6G	212 ms	79.8	-
DynamicViT	3.0G	N/A	79.3	30
EViT	3.0G	N/A	79.5	30
EViT	3.0G	N/A	79.8	100
Evo-ViT	3.0G	N/A	79.4	300
ToMe	3.0G	149 ms	79.3	0
LTMP (Ours)	3.0G	138 ms	79.6	1

*Latency measured on a Google Pixel 7.





RELATED LITERATURE

Bolya, Daniel, et al. "Token Merging: Your ViT But Faster." The Eleventh International Conference on Learning Representations. 2022.
Kim, Sehoon, et al. "Learned token pruning for transformers." Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2022.
Rao, Yongming, et al. "Dynamicvit: Efficient vision transformers with dynamic token sparsification." Advances in neural information processing systems. 2021.
Liang, Youwei, et al. "EViT: Expediting Vision Transformers via Token Reorganizations." International Conference on Learning Representations. 2021.
Xu, Yifan, et al. "Evo-vit: Slow-fast token evolution for dynamic vision transformer." Proceedings of the AAAI Conference on Artificial Intelligence. 2022.

THE DETAILS

Full paper with more results and ablations, our *timm* compatible code base, and more can be found at:

GHENT UNIVERSITY imec

maxim.bonnaerens@ugent.be @Mxbonn