# Accelerating Transformers with Spectrum-Preserving Token Merging


Hoai-Chau Tran[1,2*], Duy M. H. Nguyen[1,3,4*], Duy M. Nguyen[5], TrungTin Nguyen[6], Ngan Le[7], Pengtao Xie[8,9], Daniel Sonntag[1,10], James Zou[11], Binh T. Nguyen[2], Mathias Niepert[3,4]

[1]German Research Center for Artificial Intelligence (DFKI), [2]University of Science - VNUHCM, [3]Max Planck Research School for Intelligent Systems (IMPRS-IS), [4]University of Stuttgart, [5]Dublin City University, [6]University of Queensland, [7]University of Arkansas, [8]MBZUAI, [9]UC San Diego,[10]Oldenburg University,[11]Stanford University.

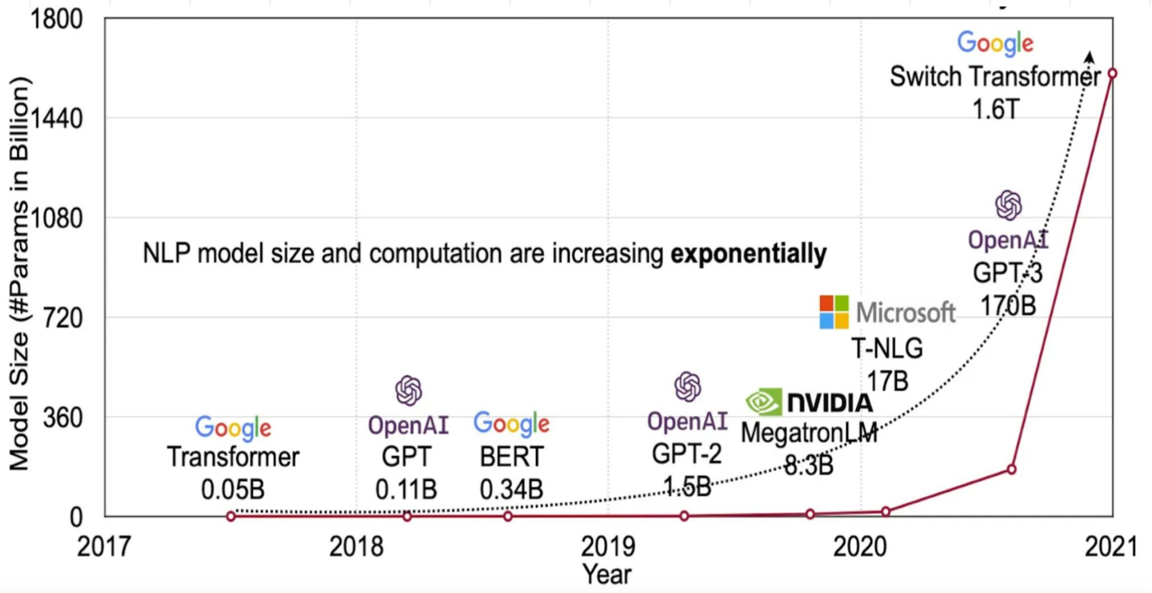
# Content

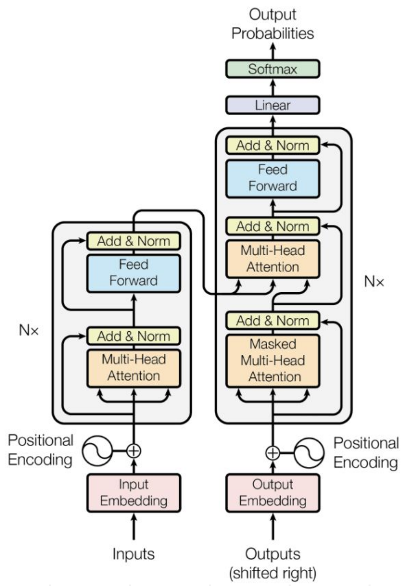
Part 1. Background and Motivation

Part 2. Our proposed method: Spectrum-preserving Token Merging

# Benefit of Token Merging

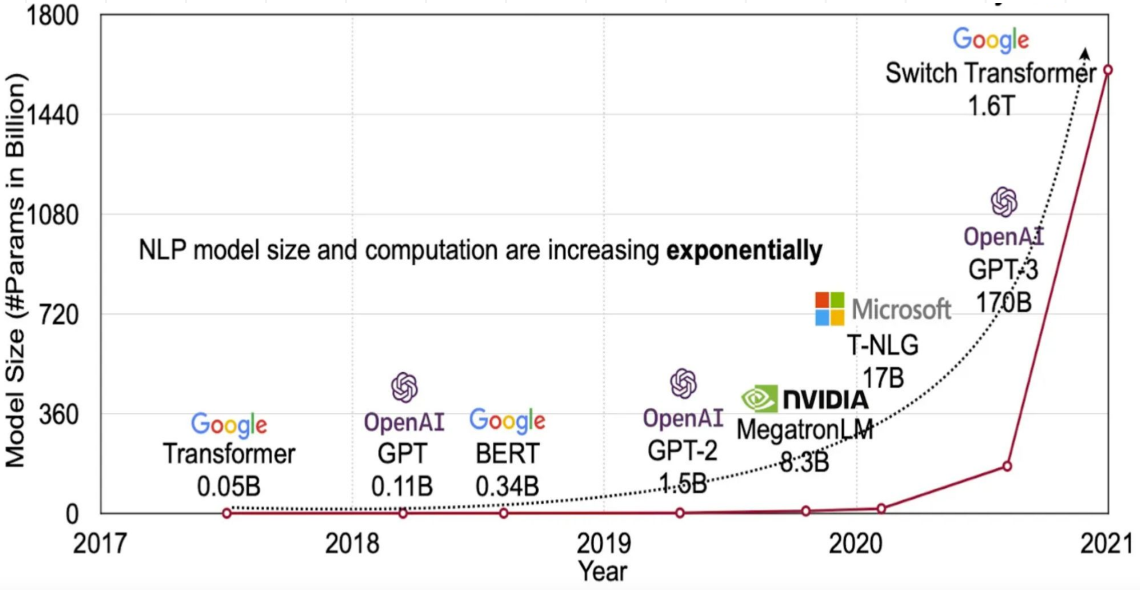
- Large Language Models (LLMs) and other SOTA architectures are based on **Transformer**.
- LLMs power is driven by volume of data and the number of parameters they are trained upon.
- LLMs model size is hugely increasing over year



# Benefit of Token Merging

- Large Language Models (LLMs) and other SOTA architectures are based on **Transformer**.
- LLMs power is driven by volume of data and the number of parameters they are trained upon.
- LLMs model size is hugely increasing over year




Requiring (i) large memory-GPU and (ii) higher computational costs for training/inference.

*Source: https://medium.com/@harishdatalab/unveiling-the-power-of-large-language-models-llms-e235c4eba8a9*

## Related Works on Token Reduction

- **Combining Tokens to a Fixed Size**

  - **None of previous methods (e.g. new efficient architecture, pruning, pooling, etc)** can **offer a reasonable speed-accuracy trade-off when combining tokens without training**

    E.g., **Token Pooling** drops of 10-40% accuracy when combining tokens without training.

**ToMe** is proposed (*Bolya, Daniel et al., 2023, ICLR 2023*) which is a **simple method** but **increase throughput ViT for both training or without training (off-the-shelf) settings.**

Experiments showed that **ToMe can 2 x throughput of state-of-the-art ViT-L @ 512 and ViT-H @ 518 models** on images and **2.2× the throughput of ViT-L on video** with only a **0.2-0.3% accuracy drop**

TOKEN MERGING: YOUR VIT BUT FASTER

Daniel Bolya[1,2*] Cheng-Yang Fu[2] Xiaoliang Dai[2] Peizhao Zhang[2]
Christoph Feichtenhofer[2] Judy Hoffman[1]
[1] Georgia Tech [2] Meta AI
{dbolya,judy}@gatech.edu, {chengyangfu,xiaoliangdai,stzpz,feichtenhofer}@meta.com

*Bolya, Daniel, et al. "Token merging: Your ViT but Faster." ICLR 2023, Top 5% paper.*

# 1. TOME - Method

- ToMe inserts a token merging module into an existing ViT (Figure 1.b)
- In each block of ViT, ToMe merges tokens to reduce by a number of r tokens.
  - ➔ Over **L blocks in the network**, merging rL tokens.
  - ➔ For e.g., on ViT-L/16, if we remove r = 8 tokens, at the final $24^{th}$ layer, we remove upto 98% tokens (Figure 1.a)

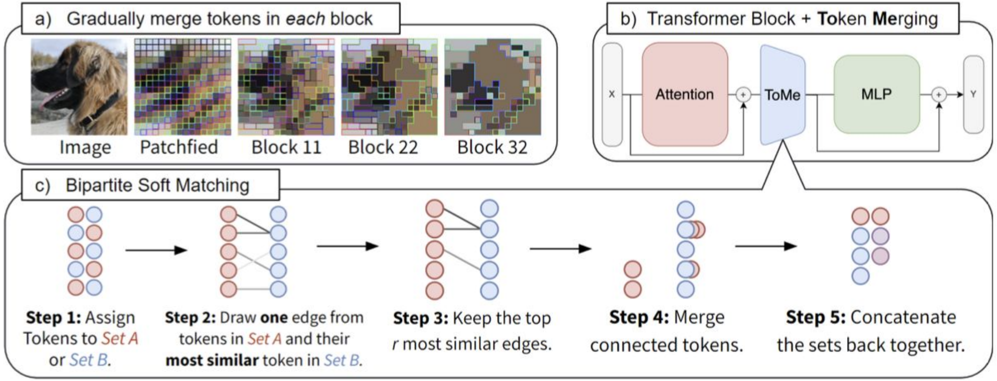



Figure 1: **Token Merging.** (a) With ToMe, similar patches are merged in each transformer block: for example, the dog's fur is merged into a single token. (b) ToMe is simple and can be inserted inside the standard transformer block. (c) Our fast merging algorithm, see Appendix D for implementation.

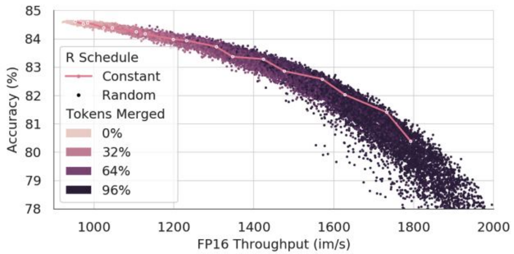



Figure 2: **Token Merging Schedule.** Our default constant merging schedule is close to optimal when compared to 15k randomly sampled merging schedules on an AugReg ViT-B/16.

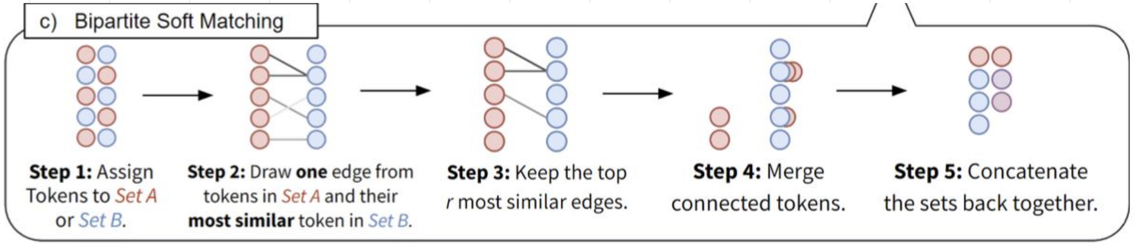
# Content

Part 1. Token Merging: Your ViT But Faster

**Part 2. Our proposed method: Energy-based Token Merging**

## 2. Energy-based Merging

- **ToMe and its variations (PuMer, LTMP, DiffRate, etc) have some significant drawbacks:**
- Firstly, the choice of a tokens-splitting strategy highly affects the performance of the algorithm.
  - ➔ ToMe **divided by odd and even indices**; therefore, unavoidable mis-merging occurs since tokens in set A perceive tokens in set B but not themselves
- Secondly, while the *bipartite soft matching algorithm* works effectively in the initial layers where redundant tokens for backgrounds and noise are abundant, **as tokens go deeper into the network, there is a risk of compromising informative tokens that represent the main object because of their high similarity.**




*Cao, Qingqing, Bhargavi Paranjape, and Hannaneh Hajishirzi. "Pumer: Pruning and merging tokens for efficient vision language models." ACL 2023*
*Bonnaerens, Maxim, and Joni Dambre. "Learned Thresholds Token Merging and Pruning for Vision Transformers." TMLR 2023*

## 2. Spectrum-preserving Token Merging

We address those problems by **prioritizes** the protection of **informative tokens** using an **additional criterion called energy score.**

Several experiments on two tasks, image classification, and image-text retrieval, using both large and small backbones models, our method demonstrates superior off-the-shelf performance.

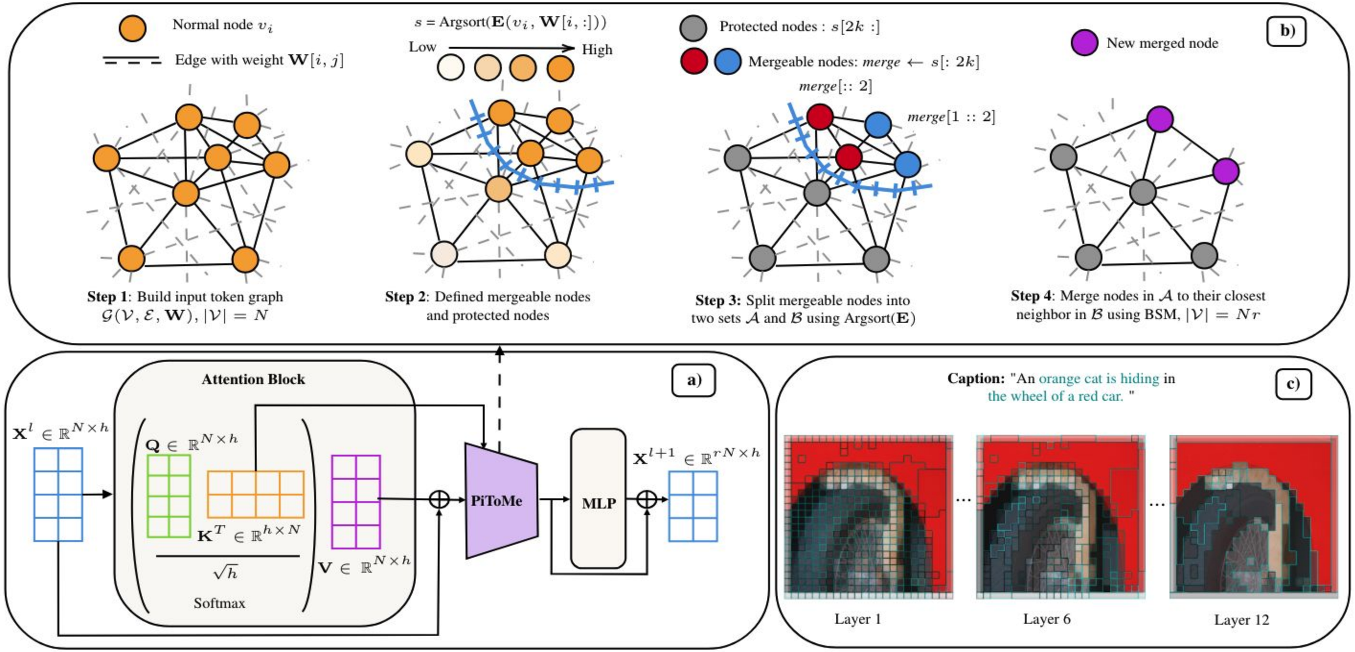


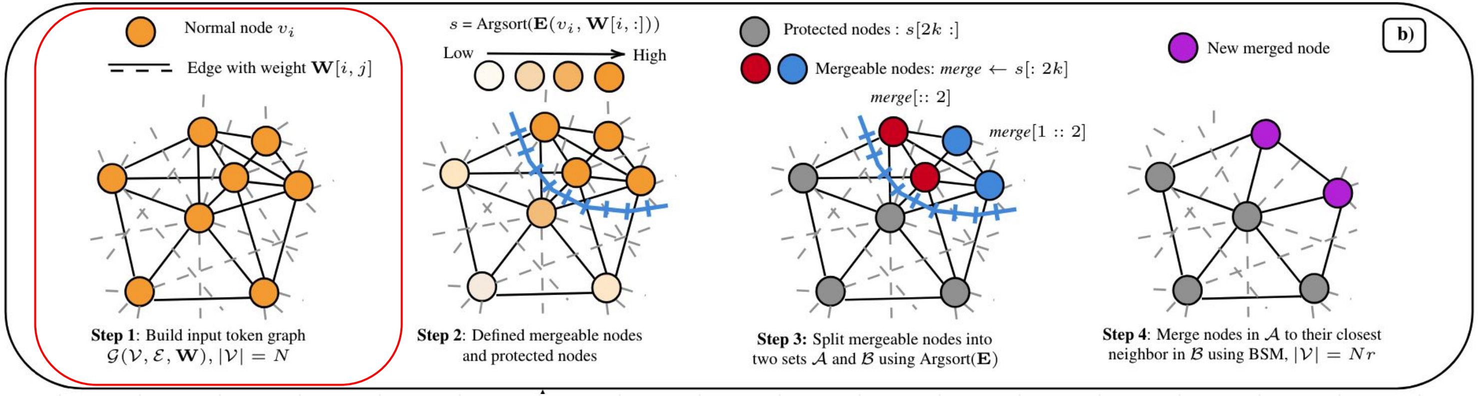

Figure 2: **a)** PiToMe can be inserted inside transformer block; **b)** Energy scores are computed to identify mergeable and protective tokens; **c)** Our algorithm gradually merges tokens in each block.

# 2. Energy-based Merging - Method

**Token Graph Construction**: Given a set of $N$ token inputs in $\hat{\mathbf{X}}^l$, we build a weighted graph $\mathcal{G}(\mathcal{V}, \mathcal{E}, \mathbf{W})$ with $\mathcal{V}$ a set of $N = |\mathcal{V}|$ nodes, $\mathcal{E}$ a set of $M = |\mathcal{E}|$ edges defined by connecting one token to the remaining ones in $\mathcal{G}$, $\mathbf{W} \in \mathbb{R}^{N \times N}$ be a weighted adjacency matrix. We opt for using the *key* vectors $\mathbf{K} = \mathbf{X}^l \mathbf{W}_K \in \mathbb{R}^{N \times h}$ as node features of $\mathcal{V}$, i.e., $v_i \in \mathcal{V}$ has $h$ feature dimensions. The weight $\mathbf{W}[i, j]$ assigned to an edge $e_{ij} \in \mathcal{E}$ connects $v_i$ and $v_j$ is computed by cosine distance:

$$\mathbf{W}[i, j] = 1 - \cos(v_i, v_j), \text{ where } \cos(v_i, v_j) = \frac{v_i \cdot v_j}{\|v_i\| \|v_j\|}, \quad \forall v_i \in \mathcal{V}, v_j \in \mathcal{V}. \tag{3}$$

For simplicity, $\mathbf{W}[i, :]$ and $\mathbf{W}[:, i]$ denote the i-th row and column, *resp.*; $[N]$ stands for $\{1, \ldots, N\}$.

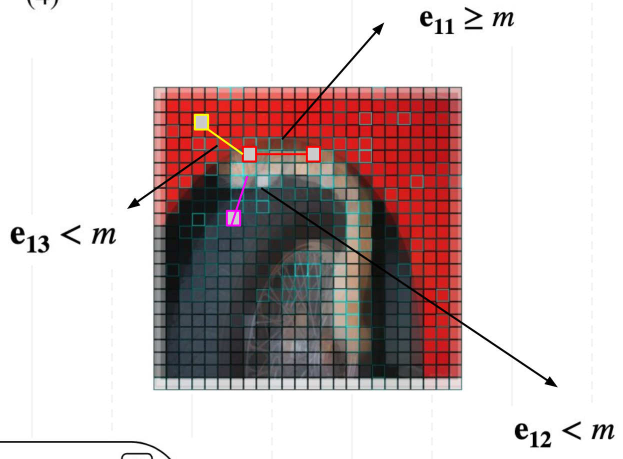


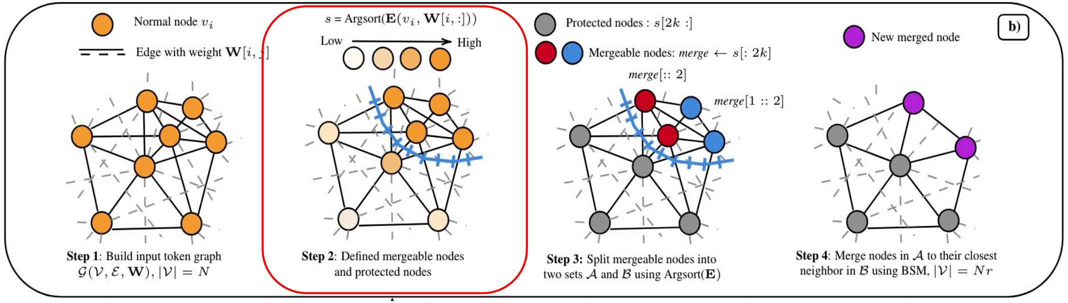
# 2. Spectrum preserving Token Merging - Method

Let $i$ be the index of the current node and $\mathcal{N}(i)$ represent the set of neighbor nodes. The energy score $E_i \equiv E_i(v_i, \mathbf{W}[i,:])$ of node $v_i$ is calculated using the following equation:

Adaptive changing

$$E_i(v_i, \mathbf{W}[i,:]) = \frac{1}{N} \sum_{j \in \mathcal{N}(i)} f_m(\cos(v_i, v_j)), \; f_m(x) = \begin{cases} x & \text{if } x \geq \boxed{m} \\ \alpha(\exp(x-m)-1) & \text{otherwise} \end{cases}. \quad (4)$$



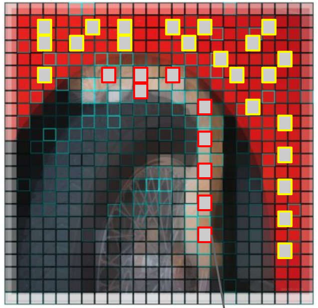



# 2. Spectrum preserving  Merging - Method

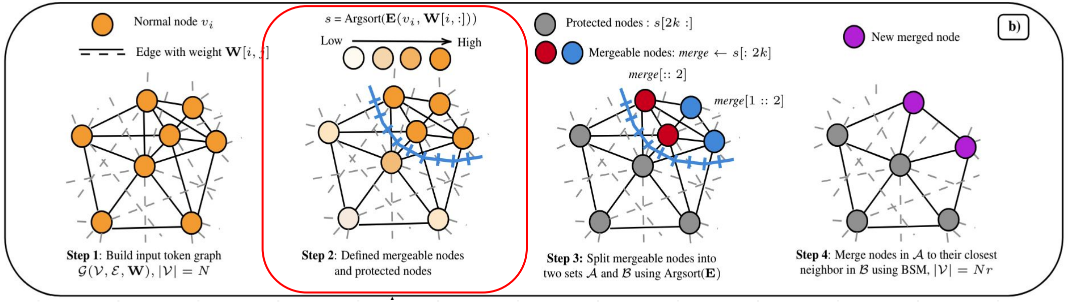
Let $i$ be the index of the current node and $\mathcal{N}(i)$ represent the set of neighbor nodes. The energy score $E_i \equiv E_i(v_i, \mathbf{W}[i,:])$ of node $v_i$ is calculated using the following equation:

$$E_i(v_i, \mathbf{W}[i,:]) = \boxed{\frac{1}{N}\sum_{j\in\mathcal{N}(i)} f_m(\cos(v_i, v_j))}, \quad f_m(x) = \begin{cases} x & \text{if } x \geq m \\ \alpha(\exp(x-m)-1) & \text{otherwise} \end{cases}. \qquad (4)$$

The summary term in the Energy score is designed to reflect the density of tokens potentially representing the same group, i.e., tokens of a smaller object will have smaller energy compared to the other. Energy scores are then estimated and sorted, and the top $2k$ nodes with the highest energy scores are selected for merging.

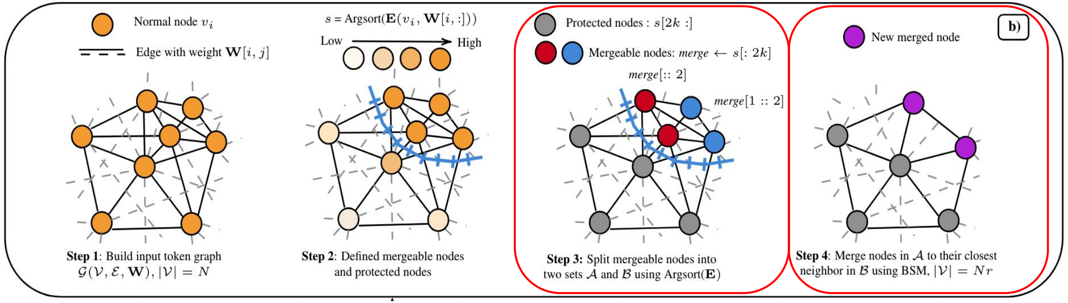


Tokens representing ‘cat’ objects will have smaller energy than other regions.

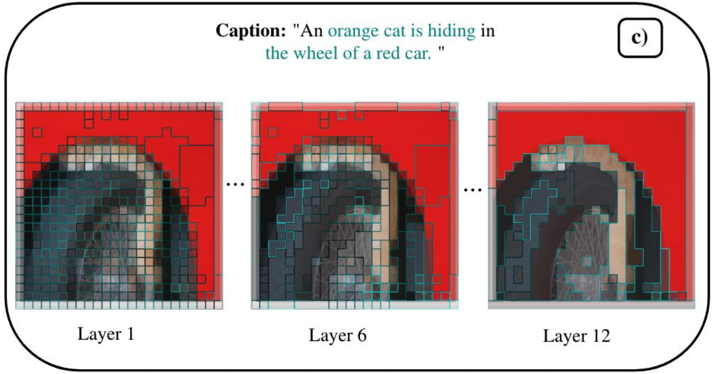


## 2. Spectrum preserving Merging - Method

**Step 3 & 4**: Having identified mergeable tokens, we partition them into two sets, denoted as $\mathcal{A}$ and $\mathcal{B}$, each containing $k$ nodes. All nodes in set $\mathcal{A}$ are merged with their nearest neighbors in set $\mathcal{B}$ through a weighted average procedure based on their energy scores.




**Low energy tokens are protected at any layer.**



# 2. Spectrum preserving Merging - Performance

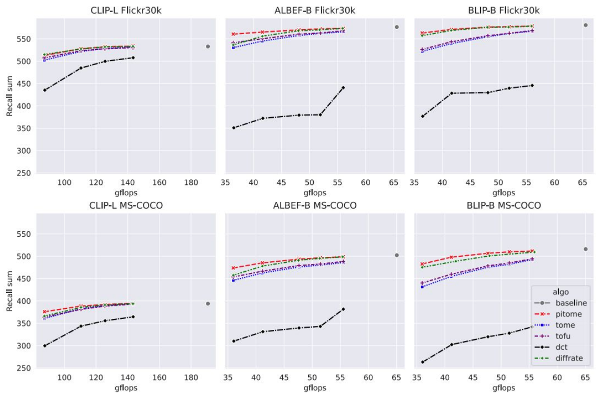



Figure 3: **Off-the-shell Image-Text Retrieval comparison** between PITOME v.s. merging/pruning methods on different backbones on tasks when varying the number of merged tokens. Here, Recall sum $= Rt@1 + Rt@5 + Rt@10 + Ri@1 + Ri@5 + Ri@10$ is close to 600, indicating recall scores at top 1,5, and 10 for retrieving image and text reached close to 100%. PITOME curves, in most cases, are above other baselines.

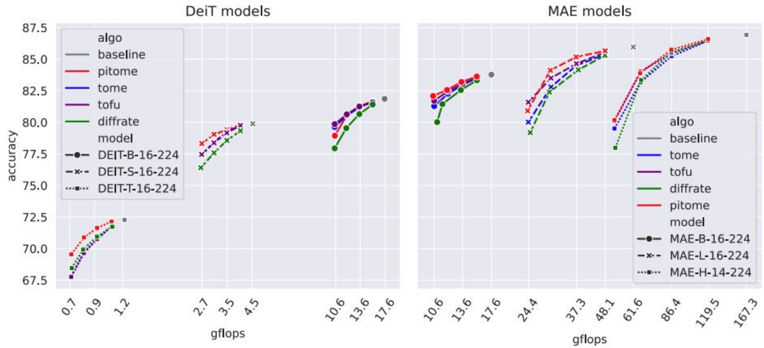



Figure 5: **Off-the-shelf results on Imagenet-1k. Zoom in for better view.**

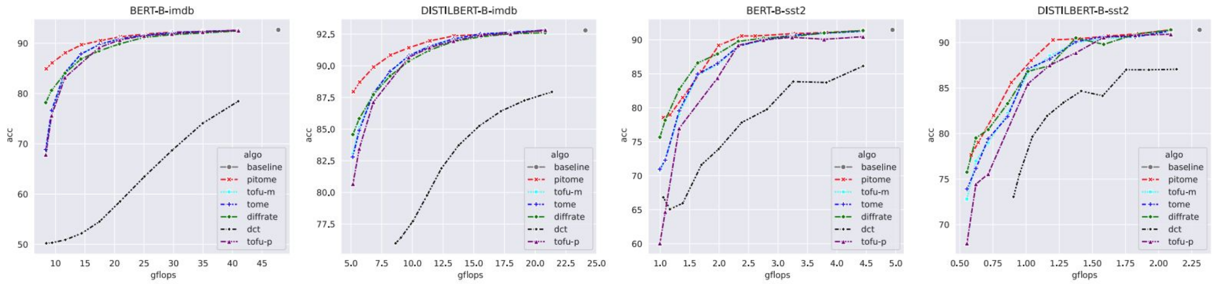



Figure 11: **Off-the-shelf** performance of various algorithms on the **text classification task**.

# 2. Spectrum preserving Merging - Experiments

## VQA with LLM

USER: Describe what you see.

USER: Describe what you see.

| Model | Answer |
|---|---|
| **LLaVA-1.5-7B** | In the image, a man is standing on a folding chair in the back of a yellow taxi cab. He is holding a blue shirt, possibly drying it or preparing to put it on. The taxi is driving down a busy street with multiple cars around it, including another taxi in front of it. The man's position on the chair and the presence of the shirt suggest that he might be a street vendor or a worker who is using the taxi as a mobile workspace. |
| **LLaVA-1.5-7B + PiToMe** | In the image, a man is standing on the back of a moving car, holding a clothesline with clothes hanging from it. The car appears to be a taxi cab, and the man is wearing a yellow shirt. The scene seems to be taking place in a busy city, as there are multiple cars visible in the background. The man's unusual activity of hanging clothes from the back of a moving car likely attracts attention and curiosity from passersby. |
| **LLaVA-1.5-7B +ToMe** | The image shows a man standing on the back of a taxi cab, holding a large umbrella. The taxi is driving down a street, and the man appears to be using the umbrella to shield himself from the elements. In addition to the taxi, there are several other vehicles in the scene, including a truck and a car. The man on the taxi is the main focus of the image, and his actions suggest that he is trying to protect himself from rain or other weather conditions. |
| **LLaVA-1.5-7B + ToFu** | The image depicts a busy city street scene with a man standing on the back of a truck, possibly performing a stunt or participating in a parade. The truck is positioned in the middle of the street, and the man appears to be the main focus of the scene. There are several other people in the image, some of whom are walking or standing near the truck. Additionally, there are multiple cars parked or driving along the street, and a traffic light can be seen in the background. The overall atmosphere suggests an urban environment with a lively and bustling atmosphere. |
| **LLaVA-1.5-7B + DiffRate** | The image features a man wearing a yellow shirt, standing next to a yellow car. He appears to be holding onto the car, possibly for support or to steady himself. The car is parked on a street, and there are other cars visible in the background. In addition to the man and the car, there are two other people in the scene, one standing further back and another closer to the foreground. A handbag can also be seen placed on the ground near the center of the image. |

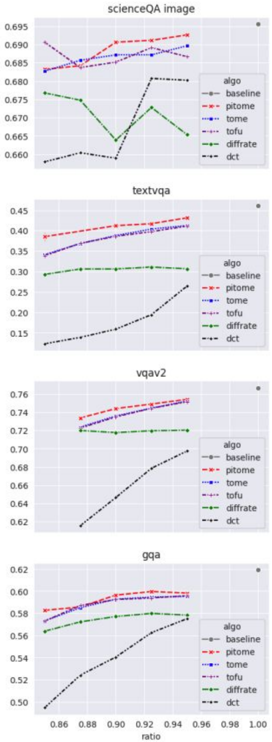
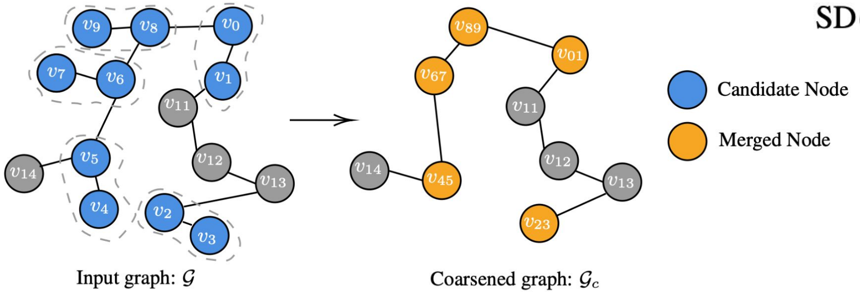



Figure 4: **Off-the-shelf** performance of PiToMe on **LLaVA-1.5-7B** with different compressing ratio $r$.

# PiToMe: Energy-based Merging - Connection to Spectral Properties




$$\mathrm{SD}(\mathcal{G}, \mathcal{G}_c) = \|\boldsymbol{\lambda} - \boldsymbol{\lambda}_l\|_1 = \sum_{i=1}^{N} |\lambda_i - \lambda_{li}| \text{ as a spectral distance.}$$

Figure 7: Token merging outputs can be seen as coarsened graph from an input graph.

**Theorem 1** (Spectrum Consistent of Token Merging). *Suppose the graphs $\mathcal{G}_0^{(s)}$, $\mathcal{G}_{\text{PITOME}}^{(s)}$, and $\mathcal{G}_{ToMe}^{(s)}$ are coarsened from the original graph $\mathcal{G}$ by iteratively merging pairs of nodes $v_{a_s}$ and $v_{b_s}$* w.r.t. *the true partition $\mathcal{P}_0^{(s)} = \{\mathcal{V}_{0i}^{(s)}\}_{i\in[s]}$, the* PITOME*-partition $\mathcal{P}_{\text{PITOME}}^{(s)} = \{\mathcal{V}_{\text{PITOME}i}^{(s)}\}_{i\in[s]}$, defined by* PITOME *Algorithm 1, and the ToMe-partition [15, 16], $\mathcal{P}_{ToMe}^{(s)} = \{\mathcal{V}_{ToMei}^{(s)}\}_{i\in[s]}$, for $s = N, \dots, n+1$. We assume some standard mild assumptions: (A1) $\mathbb{E}[\cos(v_{a_s}, v_{b_s})] \to 1, \quad \forall v_{a_s} \in \mathcal{V}_{0i}^{(s)}, \forall v_{b_s} \in \mathcal{V}_{0i}^{(s)}, i \in [s]$; (A2) there exists a margin $m$ s.t., $\cos(v_{a_s}, v_{b_s}) \geq m > \cos(v_{a_s}, v_{c_s}), \quad \forall v_{a_s} \in \mathcal{V}_{0i}^{(s)}, \forall v_{b_s} \in \mathcal{V}_{0i}^{(s)}, \forall v_{c_s} \in \mathcal{V}_{0j}^{(s)}, \forall i \neq j \in [s]$; and (A3) there is an order of cardinality in the true partition, without loss of generality, we assume $N_1^{(s)} \geq N_2^{(s)} \geq \ldots \geq N_s^{(s)}$, where $N_i^{(s)} = |\mathcal{V}_{0i}^{(s)}|, \forall i \in [s]$. Then it holds that:*

1. *The spectral distance between the original $\mathcal{G} \equiv \mathcal{G}_0^{(N)}$ and the* PITOME*-coarse $\mathcal{G}_{\text{PITOME}}^{(n)}$ graphs converges to* 0, i.e., *$SD(\mathcal{G}, \mathcal{G}_{\text{PITOME}}^{(n)}) \to 0$,*
2. *The spectral distance between the original $\mathcal{G}$ and the ToMe-coarse $\mathcal{G}_{ToMe}^{(n)}$ graphs converges to a non-negative constant $C$, with a high probability that $C > 0$.*



16

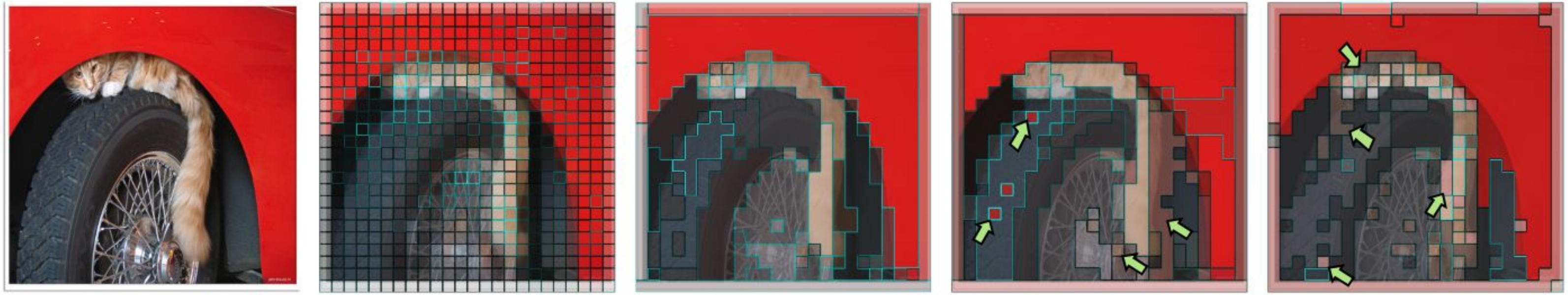
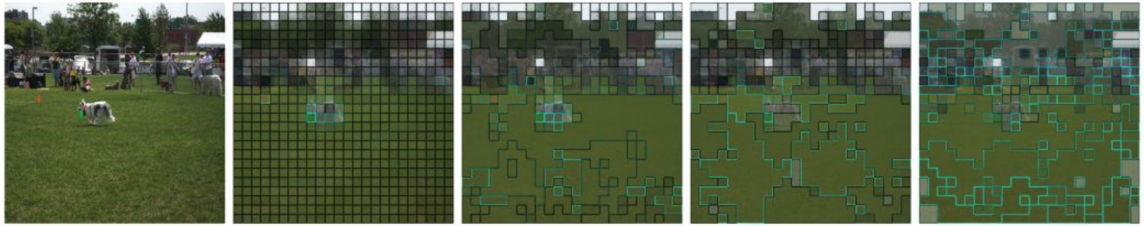



Figure 1: A comparison of token merging methods. Patches of the same color are merged. Green arrows highlight incorrect merges, avoided by PiToMe. Position of tokens with high attention scores (cyan borders, zoom for clarity) in PiToMe are maintained proportionality akin to ViT-base 384.

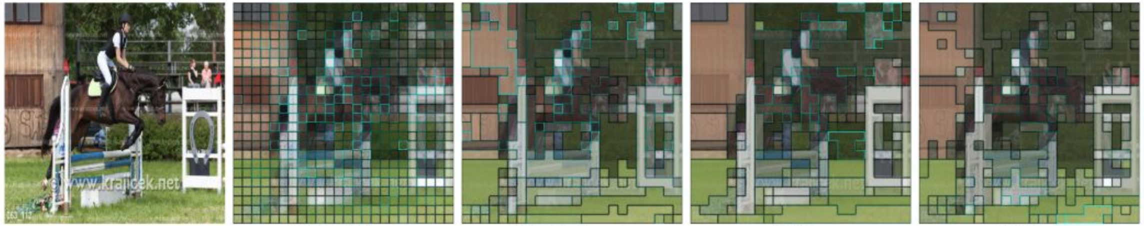
## 2. Spectrum preserving Merging - Performance

**Image-Text Retrieval**: Visualization



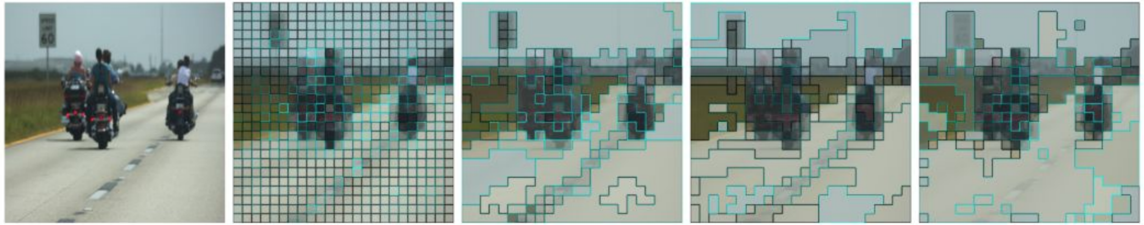
ViT-B 384 | **PiToMe (ours)** | ToMe | DiffRate

(a) A white dog catching a novelty flying disc in a competition.




ViT-B 384 | **PiToMe (ours)** | ToMe | DiffRate

(b) A woman riding a horse jumping it over obstacles.



ViT-B 384 | **PiToMe (ours)** | ToMe | DiffRate

(c) Three different motorcycle couples riding down a road.

Our implementation is available on GitHub

Thank you for listening!