MaMe: Matrix-Based Token Merging

Supplementary Material

7. Appendix

7.1. Computational Efficiency

Our token merging method aims to reduce the computational cost of self-attention by effectively shortening the sequence length. The overhead introduced by the merging process itself is analyzed as follows:

- **Similarity Matrix Calculation (S):** Computing cosine similarity 2 between M destination tokens and N source tokens involves a matrix multiplication of shape $(M \times d)$ with $(d \times N)$, resulting in $O(MNd)$ operations.
- **Adaptive Weight Pruning (W, ζ_j, $\tilde{W}$, $W^{\mathbf{F}}$):** These involve several matrix operations, primarily column-wise summations and element-wise operations. These steps are dominated by the $O(MN)$ complexity of iterating through the similarity matrix.
- **Token Aggregation ($\mathbf{X}'_{\mathbf{dst}}$):** The aggregation 9 involves a matrix multiplication of shape $(M \times N)$ with $(N \times d)$, resulting in $O(MNd)$ operations.
- **Token Preservation:** Identifying preserved tokens involves column-wise summation on W^F, which is $O(MN)$.

Given that M and N are fractions of the original sequence length L (i.e., $M \approx \alpha L$, $N \approx (1-\alpha)L$), this overhead scales approximately as $O(\alpha(1-\alpha)L^2 d)$, similar to self-attention. When merging is applied, the subsequent attention computation is reduced to $O(L'^2 d)$. Assuming $L' \approx \beta L$ with $\alpha \leq \beta \leq 1$, this becomes $O(\beta^2 L^2 d)$. Therefore, the total cost is $O\left(\left(\alpha(1-\alpha)+\beta^2\right)L^2 d\right)$. For more efficient than standard self-attention, it requires $\alpha(1-\alpha)+\beta^2 < 1$, which simplifies to the condition $\beta < \sqrt{\alpha^2-\alpha+1}$. To assess the strictness of this condition, we consider the case where α and β are uniformly distributed over $(0,1)$ with $\alpha \leq \beta$. The probability that $\beta < \sqrt{\alpha^2-\alpha+1}$ is given by: The area of the region $A = (\alpha,\beta) \mid 0 < \alpha \leq \beta \leq 1$ is $\frac{1}{2}$; the area of the region $B = (\alpha,\beta) \in A \mid \beta < \sqrt{\alpha^2-\alpha+1}$ is $\int_0^1 \left(\sqrt{\alpha^2-\alpha+1}-\alpha\right) d\alpha = \frac{3}{8}\ln 3$. Therefore, the probability is $P = B/A = \frac{3}{4}\ln 3 \approx 0.824$,, indicating that the condition holds in approximately 82.4% of cases. This means that for most parameter choices, it achieves computational efficiency. Moreover, even if the condition is not strictly met in the current block, the reduced sequence length L' propagates to subsequent blocks, ensuring that all following attention computations benefit from the shorter sequence, leading to overall computational savings across the network.

Algorithm 1 Matrix-Based Token Merging

Require: Input tokens $X \in \mathbb{R}^{L\times d}$, similarity threshold τ
Ensure: Reduced and fused token sequence $X' \in \mathbb{R}^{L'\times d}$

Note: X_{spec} refers to any special tokens (e.g., class tokens) that are preserved without modification.
function MaMe(X, τ)
  Partition X into destination tokens $X_{\text{dst}} \in \mathbb{R}^{M\times d}$ and source tokens $X_{\text{src}} \in \mathbb{R}^{N\times d}$
  Step 1. Compute Similarity Matrix
  Initialize similarity matrix $S \in \mathbb{R}^{M\times N}$
  for $i = 0 \to M-1$ **do**
    for $j = 0 \to N-1$ **do**
      $S_{ij} \leftarrow \frac{\mathbf{x}_i^{\text{dst}} \cdot \mathbf{x}_j^{\text{src}}}{\|\mathbf{x}_i^{\text{dst}}\| \cdot \|\mathbf{x}_j^{\text{src}}\|}$
    end for
  end for
  $\tilde{S} \leftarrow \text{ReLU}(S - \tau)$
  Step 2. Compute Fusion Weights
  Initialize $W, \tilde{W}, W^{\text{F}} \in \mathbb{R}^{M\times N}$ (all zeros initially)
  for $j = 0 \to N-1$ **do**
    for $i = 0 \to M-1$ **do**
      $W_{ij} \leftarrow \tilde{S}_{ij} / (\sum_{i=0}^{M-1} \tilde{S}_{ij} + \epsilon)$
    end for
    $C_j \leftarrow \sum_{i=0}^{M-1} \mathbb{I}(W_{ij} > 0)$
    $\zeta_j \leftarrow (\sum_{i=0}^{M-1} W_{ij}) / (C_j + \epsilon)$
    for $i = 0 \to M-1$ **do**
      $\tilde{W}_{ij} \leftarrow \text{ReLU}(W_{ij} - \zeta_j)$
    end for
    Mask $\leftarrow \sum_{i=0}^{M-1} \tilde{W}_{ij}$
    for $i = 0 \to M-1$ **do**
      $W_{ij}^{\text{F}} \leftarrow W_{ij} / (\text{Mask}_j + \epsilon)$
    end for
  end for
  Step 3. Compute Preserved and Fusion Tokens
  Initialize $X_{\text{pres}} \leftarrow \emptyset$
  Initialize fused destination tokens $X'_{\text{dst}} \in \mathbb{R}^{M\times d}$
  for $j = 0 \to N-1$ **do**
    if $\text{Mask}_j = 0$ **then**
      Add $\mathbf{x}_j^{\text{src}}$ to X_{pres}
    end if
  end for
  for $i = 0 \to M-1$ **do**
    $\mathbf{x}'_{\text{dst},i} \leftarrow \mathbf{X}_i^{\text{dst}} + \sum_{j=0}^{N-1} W_{ij}^{\text{F}} \mathbf{x}_j^{\text{src}}$
    $\mathbf{x}''_{\text{dst},i} = \mathbf{x}'_{\text{dst},i} / (1 + \sum_{j=1}^{N} W_{ij}^{\text{F}})$
  end for
  Step 4. Assemble the Final Tokens
  $X' \leftarrow \text{concat}(X_{\text{spec}}, X''_{\text{dst}}, X_{\text{pres}})$ ▷ Concatenate the special, fused destination, and preserved tokens
  return X'
end function

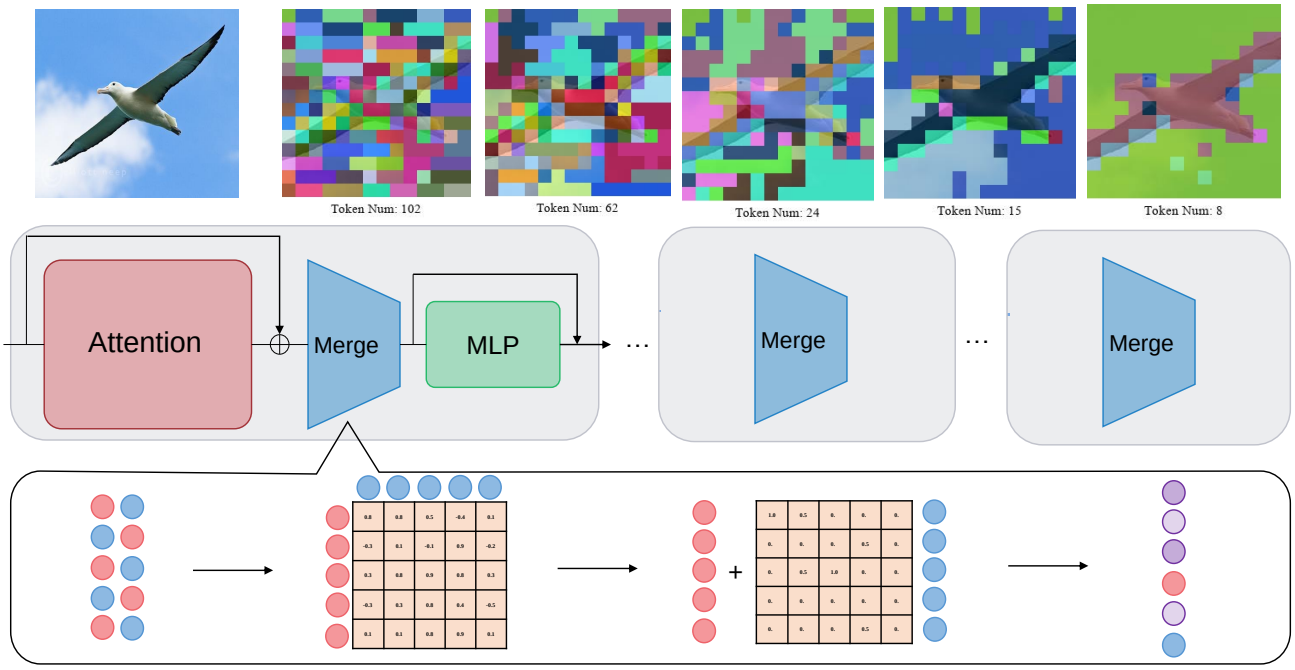


Figure 5. Illustration of MaMe Algorithm. The similarity threshold is 0.7 in this illustration.

COCO Caption Task	
	
Question:	Please describe this image in general. Directly provide the description, do not include prefix like "This image depicts".
LLaVA:	The image features a white plate with a variety of food items on it. There are two pieces of broccoli, one on the left side and the other on the right side of the plate. Additionally, there are two pieces of fish, one on the left side and the other on the right side of the plate. A fork is placed on the right side of the plate, ready to be used for eating. The plate is placed on a dining table, which is the main focus of the image.
+ToMe:	The image features a white plate with a variety of food items on it. There are two pieces of fish, one on the left side and the other on the right side of the plate. A generous portion of broccoli is spread across the plate, with some pieces located near the fish and others scattered around the plate. A fork and a knife are placed on the table, ready for use. The dining table occupies most of the image, with the plate and utensils placed on it.
+MaMe:	The image features a white plate with a variety of food items on it. There are two pieces of broccoli, one on the left side and the other on the right side of the plate. Two pieces of fish are also present on the plate, located towards the center. In addition to the food, there are utensils on the table. A fork is placed on the left side of the plate, while a knife is located on the right side. A cup can be seen on the left side of the table, and a bowl is placed on the right side. The dining table occupies the majority of the image, with the plate and utensils placed on top of it.

Table 7. MaMe correctly identified core components (plate, broccoli, fish, fork, knife) and their spatial arrangements, plus additional items (cup) that the other models miss. While it includes an erroneous "bowl," its output is structured with greater descriptive precision compared to the more generic and less detailed descriptions from LLaVA and ToMe.

Figure 6. The visualization in the 8th block of the AugReg ViT-B/16 using MaMe with different settings. Each color square represents a distinct type of token. Default is our default method.

Figure 7. The visualization illustrates the progression of token count reduction in the first 8 blocks of the AugReg ViT-B/16 with MaMe. Each color square represents a distinct type of token.

COCO Caption Task	
	
Question:	Please describe this image in general. Directly provide the description, do not include prefix like "This image depicts".
LLaVA:	2 people are standing in front of a table with a variety of wine bottles and a plate of food. The man is holding a wine bottle and the woman is holding a wine glass. They are both smiling and enjoying their time together. The table is filled with wine bottles and a few wine glasses, and there is a cake on the table as well.
+ToMe:	2 people are standing at a table with a variety of wine bottles and glasses. The man is wearing a yellow shirt and the woman is wearing a black shirt. They are both holding wine glasses and appear to be tasting different wines. There are several bottles of wine on the table, some of which are placed closer to the man and others near the woman. A plate of food is also present on the table, adding to the ambiance of the wine tasting event.
+MaMe:	2 people are standing at a table with a variety of food and drinks. The man is wearing a yellow shirt and the woman is wearing a black shirt. They are both holding wine glasses and appear to be enjoying a meal together. The table is filled with several bottles of wine, a plate of food, and a fork. The scene is set in a restaurant with a window view.

Table 8. MaMe provides the most accurate and comprehensive description, correctly identifying clothing details (yellow shirt, black shirt) and contextual elements (restaurant setting, window view). ToMe captures basic elements accurately but lacks environmental context. LLaVA omitting important visual details like clothing colors. MaMe demonstrates superior visual grounding and contextual understanding.

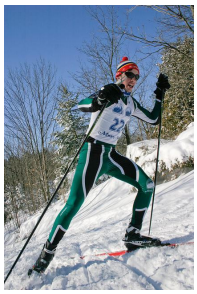
COCO Caption Task	
	
Question:	Please describe this image in general. Directly provide the description, do not include prefix like "This image depicts".
LLaVA:	A man wearing a green and white ski suit is skiing down a snowy hill. He is holding ski poles in his hands and appears to be enjoying the activity. The man is the main focus of the image, and the snowy hill and trees in the background provide a picturesque winter scene.
+ToMe:	24 is the number on the man's outfit.
+MaMe:	A man wearing a green and white ski suit is skiing down a snowy hill. He is holding ski poles in his hands and appears to be enjoying the activity. The man is wearing a hat and goggles, which are essential for protecting his eyes from the cold and snow. The scene captures the excitement and thrill of skiing in a beautiful winter landscape.

Table 9. MaMe delivers the most comprehensive and context-aware description, accurately identifying all visual elements (ski suit, poles, hat, goggles) and providing relevant contextual reasoning about protective gear. LLaVA covers basics but omits key details like headgear, while ToMe fails entirely, producing only an unrelated fragment ("number 24").