Divergence vs. Convergence: What Do We Need?
--Substitutes for "Attention"

Yi jin

(E-mail:yjauyb@gmail.com)

Abstract

This paper demonstrates that both similarity- and dissimilarity-weighted averages of features of elements (WAFE and anti-WAFE) in a sequence are effective for exchanging information among the elements. The paper also presents mean squared error and mean absolute error as effective substitutes for dot products in WAFE, effectively replacing the dot product-based "Attention" mechanism in Transformers. The code and weights are available for download from GitHub: https://github.com/yjauyb/WAFE-anti-WAFE.git.

Introduction

"Attention" based Transformers are dominate models for large language models, vision models, and other sequential and non-sequential models in the past few years. Their impact has been significant across various applications, most notably in natural language processing (NLP)[1], computer vision[2], time-series analysis[3], protein folding[4] and beyond. The core mechanism of "Attention"[5] is to assign varying levels of importance, or "weights" to different parts of the input data, allowing the model to focus more on the most relevant information. For example, in "Attention" based NLP models, each element in a sequence (like a word in a sentence) can attend to, or look at, other elements in the sequence. This enables the model to capture complex relationships and dependencies, both locally and globally, within the data.

The "Attention" mechanism is a type of similarity-weighted average of features of elements (WAFE) in a sequence. The similarity is calculated by the dot product of the query and key features. Specifically, for each element in the sequence, a query vector, a key vector, and a value vector are generated by applying a linear transformation to the feature of the element. The dot product of the query and key vectors determines the similarity or relevance between elements. This dot product is then scaled—typically by dividing by the square root of the dimension of the key vectors and passed through a softmax function to generate weights. These weights are used to compute a weighted sum of the value vectors, producing the output that aggregates information from other elements in the sequence.

Reasoning

Given a sequence $x \in \mathbb{R}^{L\times D}$, where L is the number of elements in the sequence and D is the dimensionality of the feature vector for an element $x_i \in \mathbb{R}^D$. Collecting features from similar elements during the forward pass of a model enables coherent gradient calculation in the backward pass and consistent parameter updates during training. This process allows the model to focus more on similarities among elements. When the model reaches an optimal statistical state, it relies on the similarity of elements in the sequence to generate results for the given tasks. The model process information to convergence. Extracting similarities among objects is a method humans use to classify them. In the classification process, we also consider the dissimilarities between objects.

From a statistical and recognition perspective, a model based on dissimilarity WAFE collects features from dissimilar elements and organizes dissimilar information during the forward process. In the backward process, the model calculates gradients and updates parameters coherently, similar to a model based on similarity WAFE. The key distinction in the training process is that this model focuses more on the dissimilarity of feature vectors rather than their similarity. Once trained to an optimal statistical state, the model relies on the dissimilarity of elements to predict class probabilities. This type of model

processes information toward divergence. Dot-product-based 'Attention' is an example of a similarity-based WAFE and has achieved significant success in building statistical models. Similarly, we can expect models based on dissimilarity WAFE to achieve comparable results.

The dot product is a method for calculating the similarity and dissimilarity between two vectors (Equations 1 & 2). Directly comparing two vectors using mean squared error (MSE) (Equations 3 & 4) or mean absolute error (MAE) (Equations 5 & 6) can also be used to measure similarity and dissimilarity. From this point onward, for simplicity in notation, labels or variables without the prefix "anti-" indicate calculations based on similarity-based WAFE. When the prefix "anti-" is added, the calculation is based on dissimilarity-based WAFE.

$$WAFE_{dot}(Q,K,V) = softmax(\frac{QK^T}{d_k^{0.5}})V \quad (1)$$

$$anti-WAFE_{dot}(Q,K,V) = softmax(-\frac{QK^T}{d_k^{0.5}})V \quad (2)$$

$$WAFE_{MSE}(Q,K,V) = softmax(-\frac{(Q-K)^2}{d_k})V \quad (3)$$

$$anti-WAFE_{MSE}(Q,K,V) = softmax(\frac{(Q-K)^2}{d_k})V \quad (4)$$

$$WAFE_{MAE}(Q,K,V) = softmax(-\frac{|Q-K|}{d_k})V \quad (5)$$

$$anti-WAFE_{MAE}(Q,K,V) = softmax(\frac{|Q-K|}{d_k})V \quad (6)$$

Experimental Verification

type	Acc-val (%)	Acc-train (%)
$WAFE_{dot}$ ("Attention")	82.47	85.58
anti-$WAFE_{dot}$	83.95	87.52
$WAFE_{MSE}$	82.98	86.16
anti-$WAFE_{MSE}$	82.27	86.03
$WAFE_{MAE}$	82.80	85.39
anti-$WAFE_{MAE}$	82.07	84.92

Table 1. Validation and training accuracies of WAFE- and anti-WAFE-based models with 0.18 million learnable parameters on CIFAR-10. The models have 4 layers and an embedding dimension of 64. The input image size is 112 × 112. Data augmentation is performed using RandAug (9, 0.5), and the training process lasts for 300 epochs. The validation dataset accuracy (acc-val) represents the maximum accuracy achieved during training, while the training dataset accuracy (acc-train) refers to the accuracy at the end of the training process, obtained using the same image preprocessing as for the validation dataset.

The WAFE and anti-WAFE methods were first tested on a small model with 0.18 million learnable parameters using the CIFAR-10 dataset. A small model and RandAug were selected to evaluate the effects of these methods on the capacity and generalization capability of the models. As shown in Table 1, the WAFE-dot-based model achieved accuracy very similar to that of the $WAFE_{MSE}$ and $WAFE_{MAE}$ based models on both validation and training datasets. The epoch-wise trajectories of validation accuracy and training loss for dot, MSE, and MAE-based WAFE were also very similar, as shown in Figures 1a and 2a.

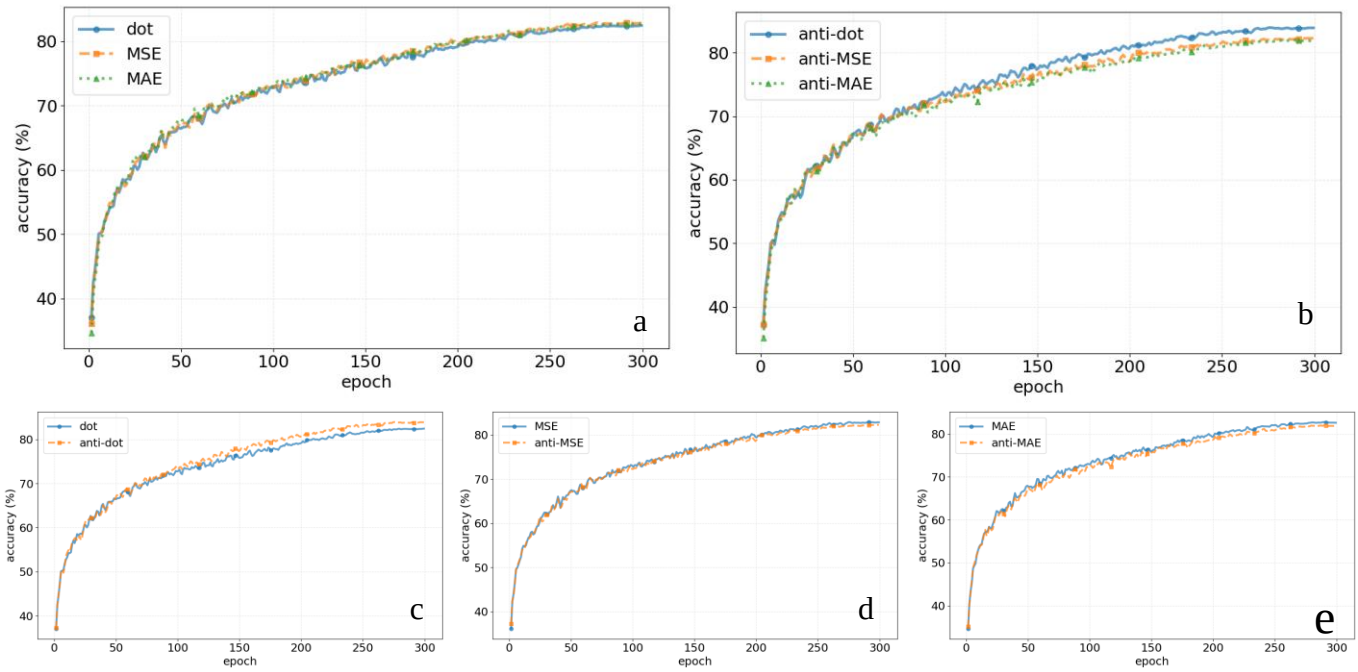

Figure 1. Validation accuracy vs. training epochs for WAFE- and anti-WAFE-based models with 0.18 million learnable parameters on CIFAR-10. The models consist of 4 layers with an embedding dimension of 64. The input image size is 112 × 112. Data augmentation is performed using RandAug (9, 0.5), and the training duration is 300 epochs.

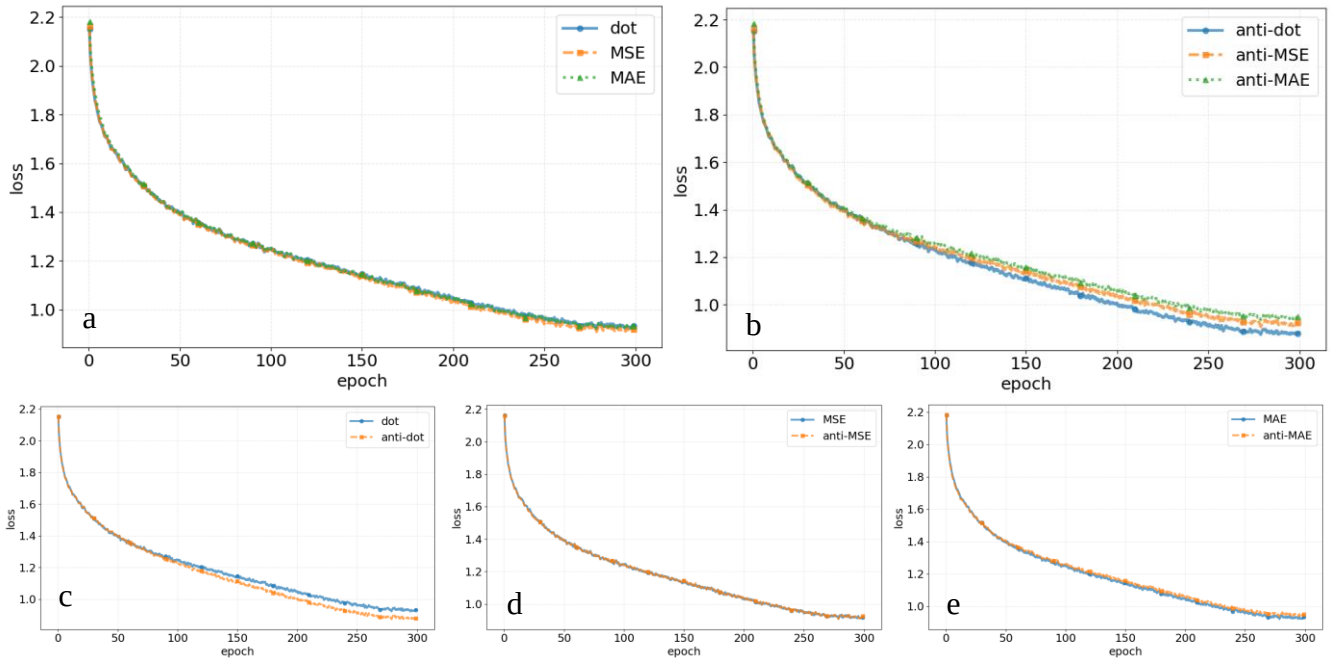

Figure 2. Training loss vs. epochs for WAFE- and anti-WAFE-based models with 0.18 million learnable parameters on CIFAR-10. The models consist of 4 layers with an embedding dimension of 64. The input image size is 112 × 112. Data augmentation is performed using RandAug (9, 0.5), and the training duration is 300 epochs.

The anti-WAFE$_{dot}$ based model demonstrated slightly better accuracy than the anti-WAFE$_{MSE}$ and anti-WAFE$_{MAE}$ based models on both validation and training datasets (Table 1). This improved performance

of the anti-$WAFE_{dot}$ based model over its anti-$WAFE_{MSE}$ and anti-$WAFE_{MAE}$ counterparts was also evident from the validation accuracy trajectories and training loss, showing higher accuracy and lower training loss over a significant portion of the training epochs (Figures 1b & 2b). The accuracy and training loss were very similar between the MSE or MAE-based WAFE and anti-WAFE models (Figures 1d, 1e, 2d, and 2e). However, the anti-$WAFE_{dot}$ based model exhibited slightly better performance than the $WAFE_{dot}$ based model (Figures 1c & 2c).

These experiments indicate that anti-WAFE-based models have a capacity similar to WAFE-based models. Replacing the dot product with MSE or MAE in the calculation of similarity or dissimilarity did not lead to significant changes in model capacity or generalization capability.

type	Acc-val (%)	Acc-train (%)
$WAFE_{dot}$ ("Attention")	93.51	99.99
anti-$WAFE_{dot}$	94.07	99.99
$WAFE_{MSE}$	93.48	100.0
anti-$WAFE_{MSE}$	93.19	99.98
$WAFE_{MAE}$	91.86	99.99
anti-$WAFE_{MAE}$	91.65	99.95

Table 2. Validation and training accuracies of WAFE- and anti-WAFE-based models with 15.05 million learnable parameters on CIFAR-10. The models consist of 12 layers with an embedding dimension of 320. The input image size is 224 × 224. Data augmentation is performed using RandAug (9, 0.5), and the training duration is 300 epochs. The accuracy for the validation dataset (acc-val) represents the maximum accuracy achieved during training. The accuracy for the training dataset (acc-train) is the accuracy at the end of the training process, obtained using the same image preprocessing as for the validation dataset.

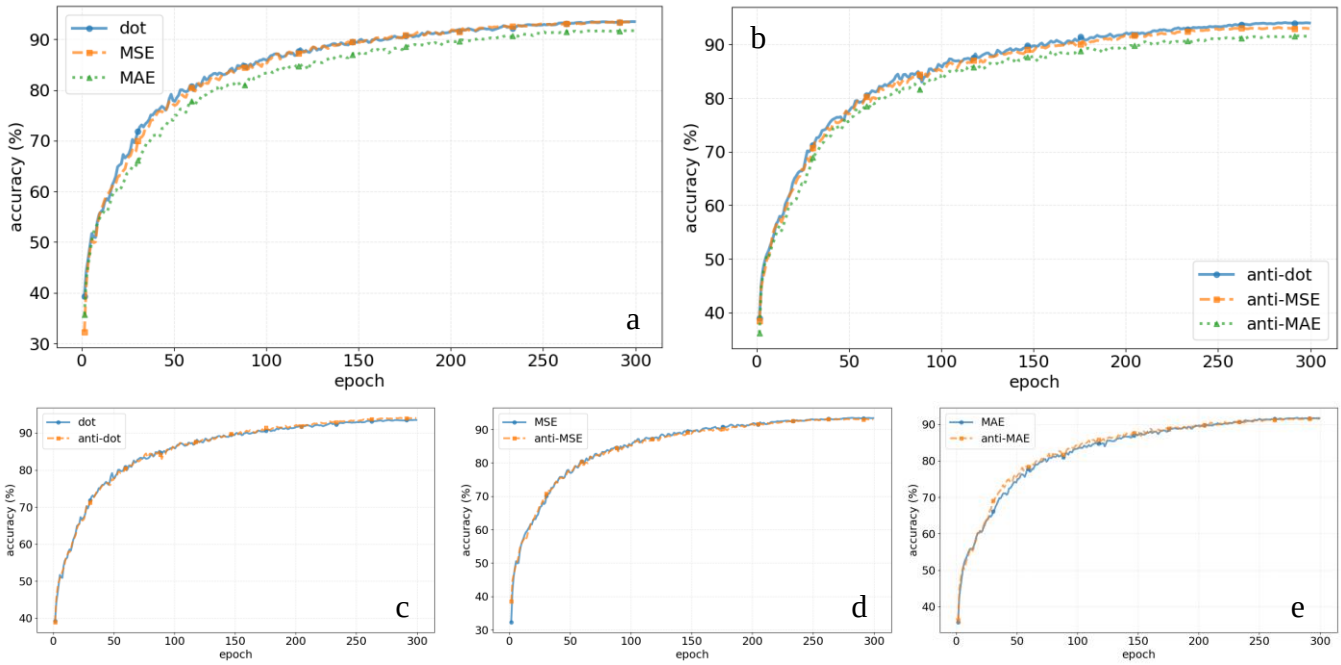

Figure 3. Validation accuracy vs. training epochs for WAFE- and anti-WAFE-based models with 15.05 million learnable parameters on CIFAR-10. The models consist of 12 layers with an embedding dimension of 320. The input image size is 224 × 224. Data augmentation is performed using RandAug (9, 0.5), and the training duration is 300 epochs.

The WAFE and anti-WAFE methods were further tested on highly over-parameterized models with 15.05 million learnable parameters on CIFAR-10. These highly over-parameterized models were selected to further evaluate the generalization capabilities of the WAFE-based and anti-WAFE-based models. The $WAFE_{MSE}$ and $WAFE_{dot}$ based models achieved very similar accuracy on both the validation and training datasets (Table 2). The increase in validation accuracy and the decrease in training loss were very similar during the training process (Figures 3a and 4a). Their performance was slightly better than the $WAFE_{MAE}$ based model. A similar trend was observed in the anti-$WAFE_{dot}$, anti-$WAFE_{MSE}$, and anti-$WAFE_{MAE}$ based models. Compared to the 0.18 million parameter models, accuracy and training loss were very similar between the MSE or MAE-based WAFE and anti-WAFE models with 15.05 million parameters (Figures 3d, 3e, 4d, and 4e). The anti-$WAFE_{dot}$ based model also exhibited slightly better performance than the $WAFE_{dot}$ based model (Table 2, Figures 3c & 4c). These results indicate that the anti-WAFE models have similar generalization capabilities to WAFE models in highly over-parameterized settings. The replacement of the dot product with MSE or MAE in the calculation of similarity or dissimilarity did not result in significant changes to the generalization capability of the models.

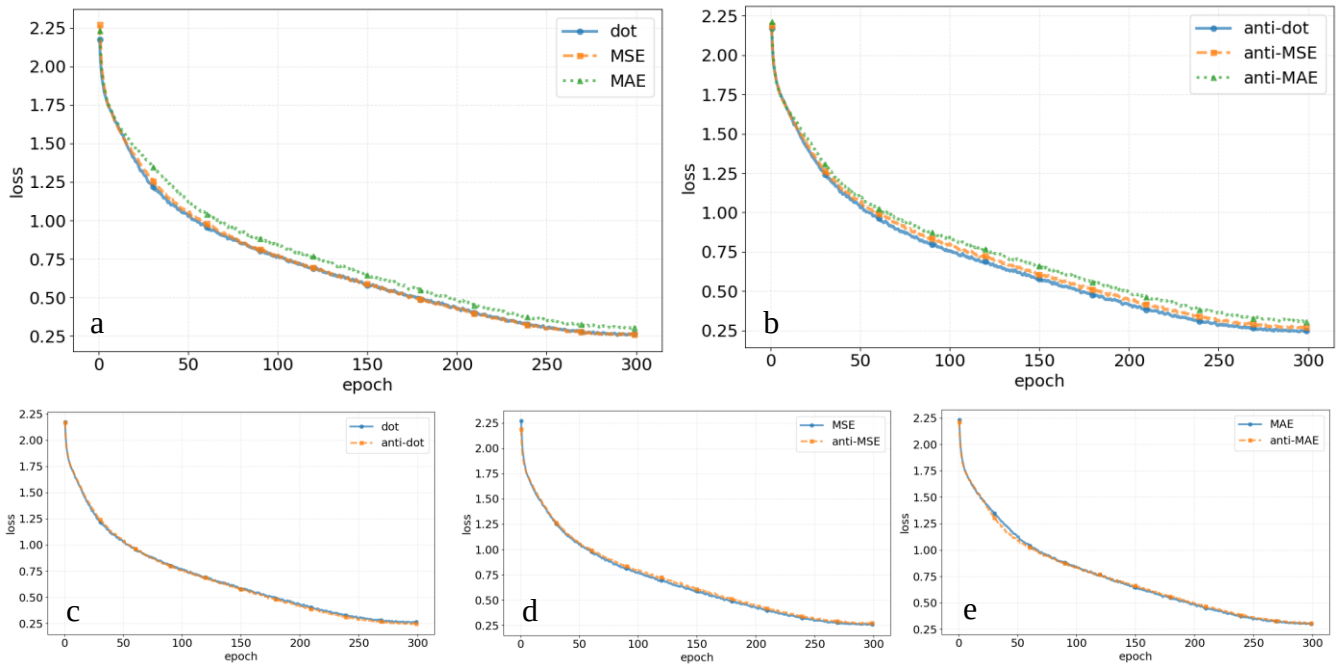

Figure 4. Training loss vs. epochs for WAFE- and anti-WAFE-based models with 15.05 million learnable parameters on CIFAR-10. The models consist of 12 layers with an embedding dimension of 320. The input image size is 224 × 224. Data augmentation is performed using RandAug (9, 0.5), and the training duration is 300 epochs.

type	Acc-val (%)	Acc-train (%)
$WAFE_{dot}$ ("Attention")	94.01	98.24
anti-$WAFE_{dot}$	93.88	97.89
$WAFE_{MSE}$	94.49	98.89
anti-$WAFE_{MSE}$	93.35	97.93
$WAFE_{MAE}$	92.26	97.53
anti-$WAFE_{MAE}$	91.47	96.40

Table 3. Validation and training accuracies of WAFE- and anti-WAFE-based models with 15.05 million learnable parameters on CIFAR-10. The models consist of 12 layers with an embedding dimension of

320. The input image size is 224 × 224. Data augmentation is performed using RandAug (9, 0.5), and the training duration is 300 epochs. Label smoothing (0.1), mixup (0.8), and cutmix (1.0) are also applied during training. The accuracy for the validation dataset (acc-val) represents the maximum value obtained during training. The accuracy for the training dataset (acc-train) is the accuracy at the end of the training process, obtained using the same image preprocessing as for the validation dataset.

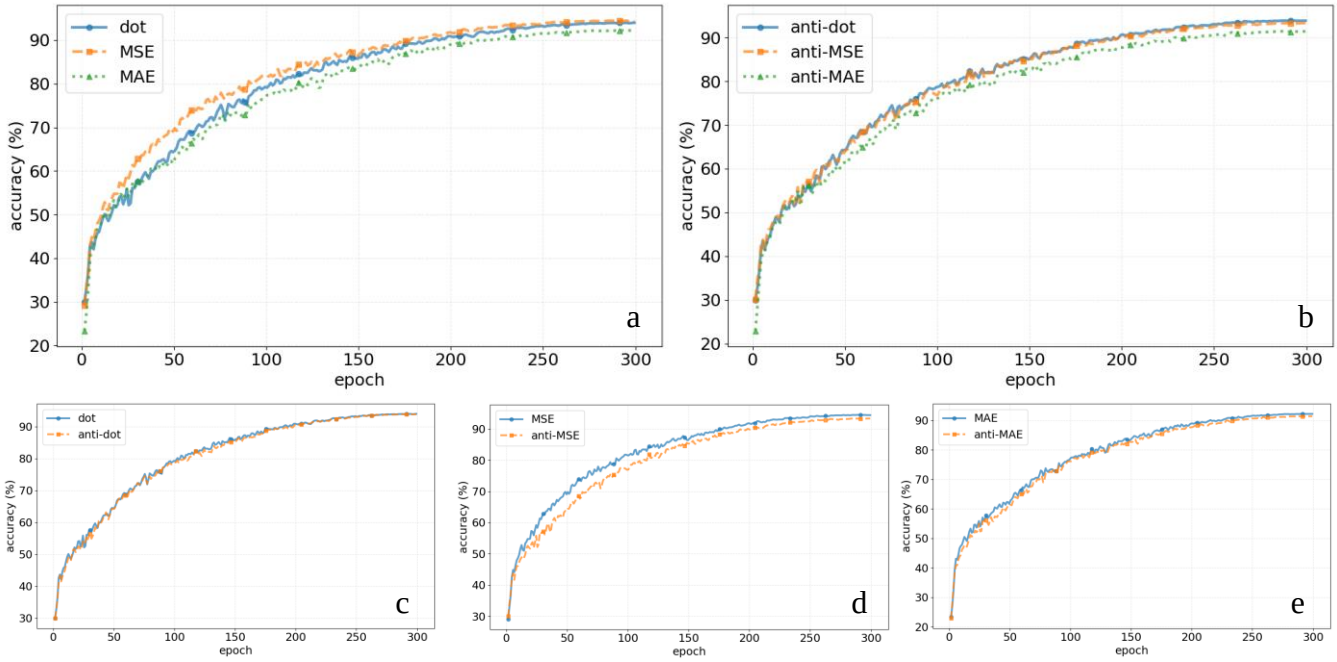

Figure 5. Validation accuracy vs. training epochs for WAFE- and anti-WAFE-based models with 15.05 million learnable parameters on CIFAR-10. The models consist of 12 layers with an embedding dimension of 320. The input image size is 224 × 224. Data augmentation is performed using RandAug (9, 0.5), and the training duration is 300 epochs. Label smoothing (0.1), mixup (0.8), and cutmix (1.0) are also applied during training.

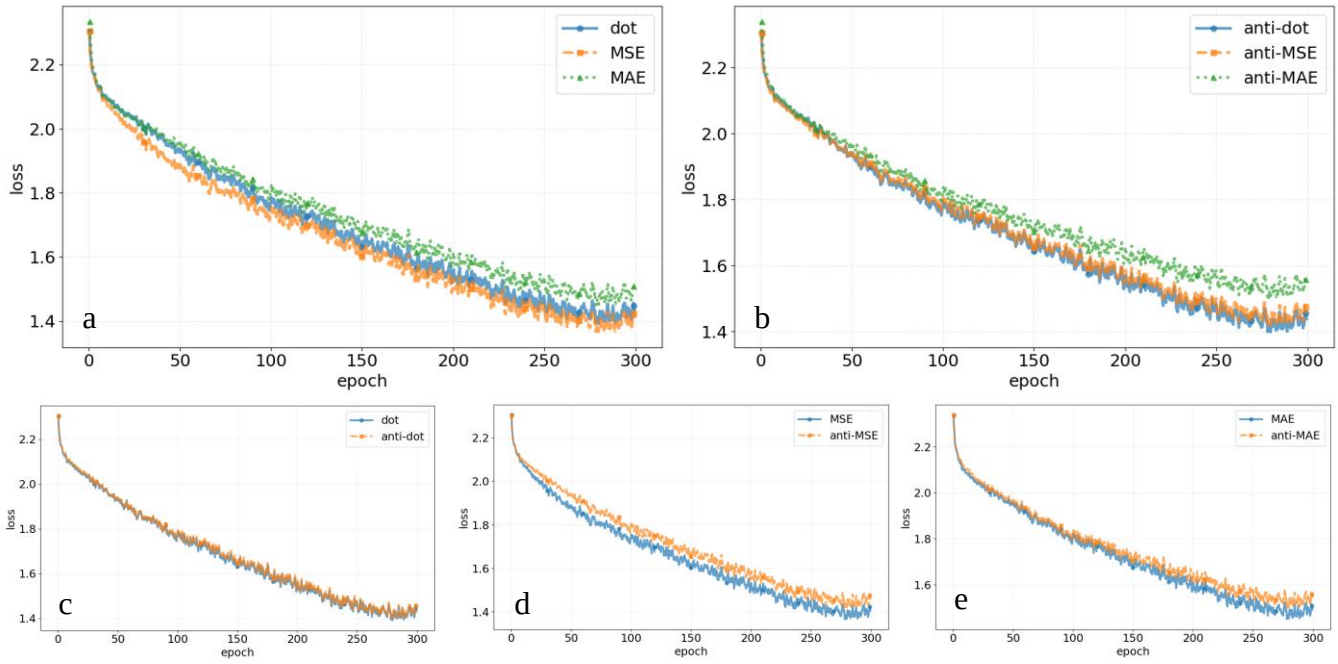

Figure 6. Training loss vs. epochs for WAFE- and anti-WAFE-based models with 15.05 million learnable parameters on CIFAR-10. The models consist of 12 layers with an embedding dimension of 320. The input image size is 224 × 224. Data augmentation is performed using RandAug (9, 0.5), and the training duration is 300 epochs. Label smoothing (0.1), mixup (0.8), and cutmix (1.0) are also applied during training.

The effect of mixing images and labels on WAFE- and anti-WAFE-based models was also evaluated on highly over-parameterized models with 15.05 million learnable parameters on CIFAR-10, using label smoothing (0.1), mixup (0.8), and cutmix (1.0). As shown in Table 3, Figure 5, and Figure 6, the combination of label smoothing, mixup, and cutmix did not result in significant performance enhancement on the over-parameterized models. The relative performance of anti-WAFE-based models compared to WAFE-based models decreased when trained with these techniques, as compared to models trained with only RandAug data augmentation. The manually assigned label smoothing ratio cannot accurately reflect the true correlation between classes. Additionally, the methods for calculating mixup and cutmix labels generated more noise, making it harder for the models to capture the true relationships among classes. As a result, label smoothing, mixup, and cutmix made it more difficult for anti-WAFE-based models to structure dissimilar information from elements, thereby decreasing their relative performance compared to WAFE-based models. In other words, anti-WAFE is more sensitive to label errors than WAFE-based models.

type	Acc-val (%)	Acc-train (%)
$WAFE_{dot}$ ("Attention")	76.39	90.49
anti-$WAFE_{dot}$	76.01	90.55

Table 4. Validation and training accuracies of $WAFE_{dot}$ and anti-$WAFE_{dot}$ based models with 10.43 million learnable parameters on the ImageNet1K dataset. The models consist of 8 layers with an embedding dimension of 320. The input image size is 224 × 224. Data augmentation is performed using RandAug (9, 0.5), and the training duration is 200 epochs. The accuracy for the validation dataset (acc-val) represents the maximum value obtained during training. The accuracy for the training dataset (acc-train) is the accuracy at the end of the training process, obtained using the same image preprocessing as for the validation dataset.

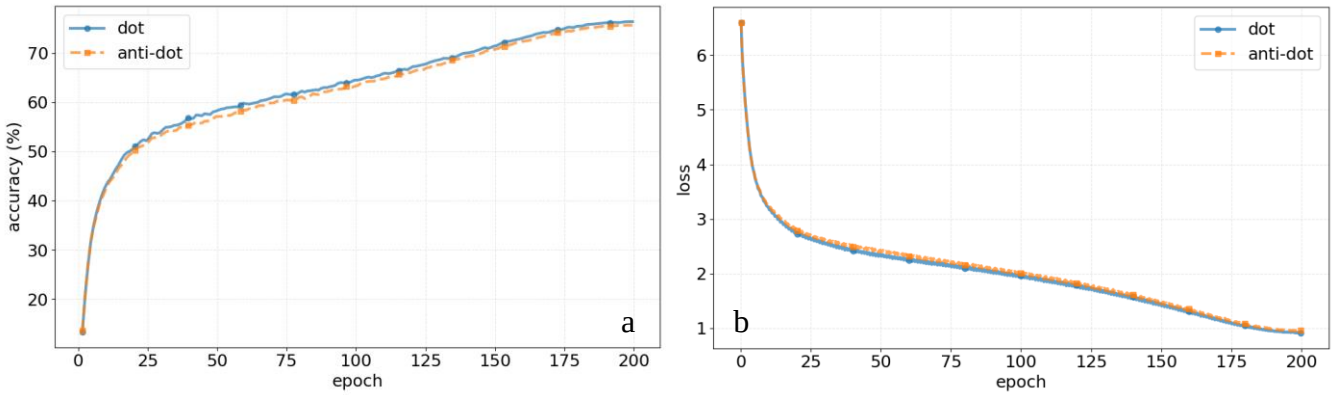

Figure 7. Validation accuracy and training loss vs. training epochs for $WAFE_{dot}$ and anti-$WAFE_{dot}$ based models with 10.43 million learnable parameters on the ImageNet1K dataset. The models consist of 8 layers with an embedding dimension of 320. The input image size is 224 × 224. Data augmentation is performed using RandAug (9, 0.5), and the training duration is 200 epochs.

$WAFE_{dot}$ and anti-$WAFE_{dot}$ methods were also evaluated on models with 10.43 million learnable parameters on the ImageNet1K dataset[6]. These two models exhibited similar validation and training

accuracies, as well as similar loss trajectories (Table 4, Figure 7). The overall performance of the $WAFE_{dot}$ -based model is slightly better than the anti-$WAFE_{dot}$-based model, which is the opposite of what was observed in the experiments with small models on the smaller CIFAR-10 dataset (Table 1, Figure 1, Figure 2). The ImageNet dataset is well known to contain approximately 6% label errors, which is much higher than the ~0.5% label errors in CIFAR-10.[7] As analyzed in the results from Table 3 and Figures 5 & 6, label errors make the anti-$WAFE_{dot}$-based models struggle to structure dissimilar information from elements, thus decreasing their relative performance compared to the $WAFE_{dot}$-based models.

Conclusion

Anti-WAFE-based models have similar capacity to WAFE-based models. Replacing the dot product with MSE or MAE in the calculation of similarity or dissimilarity did not result in significant changes to the model's capacity or generalization capability. Since $WAFE_{dot}$-based models are a minimized version of the ViT model[8], the experiments clearly show that anti-$WAFE_{dot}$, $WAFE_{MSE}$, anti-$WAFE_{MSE}$, $WAFE_{MAE}$, and anti-$WAFE_{MAE}$ can replace the dot product-based "Attention" mechanism in Transformers under certain conditions.

Limitations and Future Release Plan

This paper primarily focuses on verifying the reasoning behind WAFE and anti-WAFE and presents alternatives for substituting the dot product-based "Attention." The experiments were mainly conducted on the relatively simple CIFAR-10 classification task. As the results met the design expectations, ablation studies on ImageNet1K classification and NLP applications were not performed. An elegant combination of WAFE and anti-WAFE, named Comprehensive-WAFE, demonstrates much better performance than individual WAFE or anti-WAFE based models and will be released soon.

Experiment Implementation Details

The model is built by stacking several layers of building blocks. Each building block consists of a WAFE or anti-WAFE based sub-block and an MLP block. LayerNorm is applied at the beginning of each sub-block. All models use the same sine-cosine positional embeddings, which are added after the 2D convolution embedding. The final features are extracted from the last MLP block using average pooling and are further projected via a linear function for classification. Cross-entropy loss is used as the loss function when label smoothing, mixup, and cutmix are not applied; otherwise, soft target cross-entropy is used. Details of the training parameters can be found in Table 5.

config name	value
optimizer	AdamW
learning rate	0
weight decay	0.05
optimizer momentum	β_1, β_2 = 0.9, 0.999
batch size	128**
learning rate schedule	cosine decay with linear warmup
warmup steps	50
training epochs	300**
data augmentation	RandAug (9, 0.5)[9]
label smoothing	0.1*
mixup[10]	0.8*
cutmix[11]	1*

Table 5. Parameters for training WAFE or anti-WAFE based models from scratch. **The default training epochs are 200, and the default batch size is 256 on ImageNet1K. *Label smoothing, mixup, and cutmix are not used unless specified in the model.

Author information

Yi Jin received his Ph.D. in Chemistry from Clemson University. His current research focuses on the intersection of artificial intelligence and chemistry, exploring innovative applications of AI to advance the field.

References

1. Llama Team AI @ Meta, The Llama 3 Herd of Models. *arXiv:2407.21783* **2024**.
2. Beyer, L.; et al., Paligemma: A Versatile 3B VLM for Transfer. *arXiv:2407.07726* **2024**.
3. Wen, Q.; Zhou, T.; Zhang, C.; Chen, W.; Ma, Z.; Yan, J.; Sun, L., Transformers in Time Series: A Survey. *arXiv:2202.07125* **2022**.
4. Abramson, J., et al., Accurate Structure Prediction of Biomolecular Interactions with Alphafold 3. *Nature* **2024**, *630*, 493-500.
5. Vaswani, A., et al., Attention Is All You Need. *arXiv:1706.03762* **2017**.
6. Russakovsky, O., et al., Imagenet Large Scale Visual Recognition Challenge. *arXiv:1409.0575* **2015**.
7. Northcutt, C. G.; Athalye, A.; Mueller, J., Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks. *arXiv:2103.14749* **2021**.
8. Dosovitskiy, A., et al., An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv:2010.11929* **2020**.
9. Cubuk, E. D.; Zoph, B.; Shlens, J.; Le, Q. V., Randaugment: Practical Automated Data Augmentation with a Reduced Search Space. *arXiv:1909.13719* **2020**.
10. Zhang, H.; Cisse, M.; N. Dauphin, Y.; Lopez-Paz, D., Mixup: Beyond Empirical Risk Minimization. *arXiv:1710.09412* **2017**.
11. Yun, S.; Han, D.; Oh, S. J.; Chun, S.; Choe, J.; Yoo, Y., Cutmix: Regularization Strategy to Train Strong Classifiers with Localizable Features. *arXiv:1905.04899* **2019**.