Effort Model for Generalizable Deepfake Detection with Orthogonal Subspace Decomposition

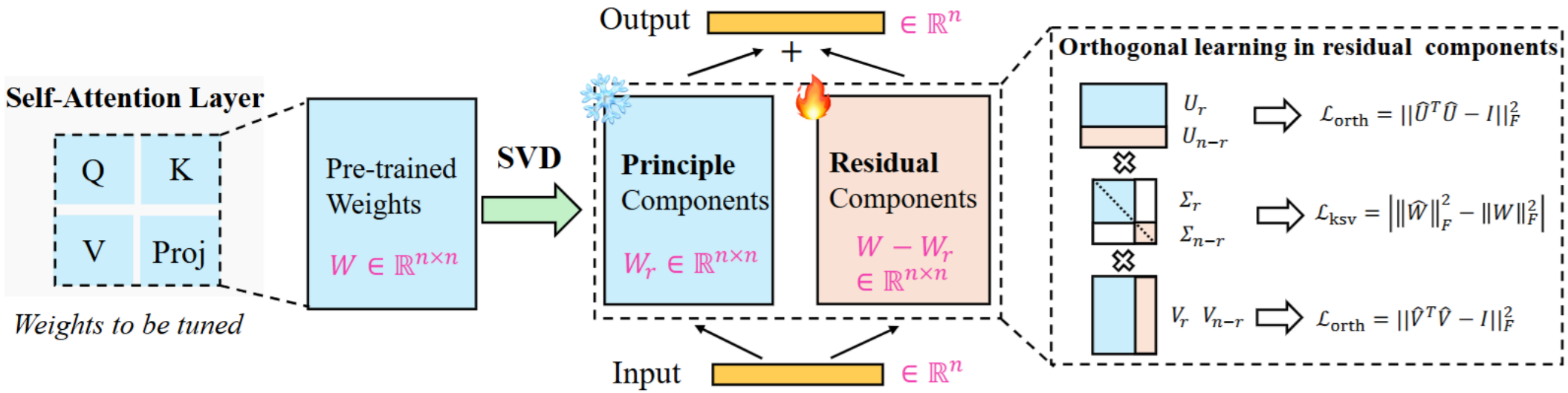
Pre-trained model

Effort: Efficient Orthogonal Modeling for Generalizable AI-Generated Image Detection

Zhiyuan Yan[1,2*], Jiangming Wang[2*], Zhendong Wang[3*], Peng Jin[1], Ke-Yue Zhang[2], Shen Chen[2],
Taiping Yao[2], Shouhong Ding[2†], Baoyuan Wu[4], Li Yuan[1†]
School of Electronic and Computer Engineering, Peking Univerisity[1],
Tencent Youtu Lab[2], University of Science and Technology of China[3],
School of Data Science, The Chinese University of Hong Kong, Shenzhen[4]

- Yan et al., "Effort: Efficient Orthogonal Modeling for Generalizable AI-Generated Image Detection", ICML 2025 (arXiv, Nov 2024).

Method


Self-Attention Layer
Q
K
V
Proj
Weights to be tuned
Pre-trained Weights
$W \in \mathbb{R}^{n\times n}$
SVD
Output
$\in \mathbb{R}^n$
+
Principle Components
$W_r \in \mathbb{R}^{n\times n}$
Residual Components
$W - W_r \in \mathbb{R}^{n\times n}$
Input
$\in \mathbb{R}^n$
Orthogonal learning in residual components
U_r U_{n-r}
$\mathcal{L}_{\text{orth}} = ||\hat{U}^T\hat{U} - I||_F^2$
Σ_r Σ_{n-r}
$\mathcal{L}_{\text{ksv}} = \left|\|\hat{W}\|_F^2 - \|W\|_F^2\right|$
V_r V_{n-r}
$\mathcal{L}_{\text{orth}} = ||\hat{V}^T\hat{V} - I||_F^2$

Method

- In this approach, the pre-trained weights of the self-attention layer are decompose using **Singular Value Decomposition (SVD)** to separate the knowledge inoto two distinct parts:
  1. **Freezing Principle Components**: The Principle Components (W_r) which contain the general-purpose knowledge of the pre-trained model, are frozen to preserve the original foundation features.
  2. **Updating Residual Components**: Only the Residual Components ($W - W_r$) which are hypothesized to capture features specifically relevant to Deepfakes, are targeted for updates during the tuning process.
- **Orthogonal Constraint for Generalization:** A key aspect of this method is maintaining **orthogonality** between the Residual and Principle components.
- By enforcing this orthogonal relationship, the model prevents the updated deepfake-specific information from interfering with the original knowledge, thereby ensuring robust **generalization performance** on unseen test datasets.

Train dataset & Weights

- We utilized weights pre-trained on the **FF++ dataset.**
- The pre-trained weights were further trained on the FF++ dataset, excluding the DeepFakeDetection fake data. We determined that the quality of images generated by DeepFakeDetection is **subpar compared to current state-of-the-art deepfakes**, making them less representative of modern deepfake characteristics.

Face Cropping

- Face cropping was performed using a pre-trained YOLOv8 model based on a rule-based approach:
  ① **Primary Rule:** Detect faces with confidence ≥0.75.
  ② **Fallback Rule:** If no face exceeds 0.75, select the single box with the maximum confidence.
  ③ **Scope:** This procedure was identical for both train and test sets."

Fine-tuning

- The model was trained by updating only the residual weights ($W - W_r$)
- During optimization, we utilized a **loss function designed to preserve orthogonality** between the principle and residual components as defined by SVD.

Inference

1. **Preprocessing:** Faces are converted to RGB, resized to the target resolution, and normalized using CLIP-based mean/std.
2. **Aggregation:** For videos, probabilities from sampled frames are averaged to produce a final authenticity score.
3. **Output:** Results are mapped to a baseline CSV, ensuring 100% coverage by filling failure cases with a neutral 0.5 probability.