NeurIPS 2024
The 38th Annual Conference on Neural Information Processing Systems

Provable Benefit of Cutout and CutMix for Feature Learning

Junsoo Oh and Chulhee Yun
KAIST AI

OptiML Optimization & Machine Learning Laboratory
KAIST AI Kim Jaechul Graduate School

Cutout and CutMix

(cat,dog) = (0,1)

Original

(cat,dog) = (0,1)

Cutout

(cat,dog) = (0.4,0.6)

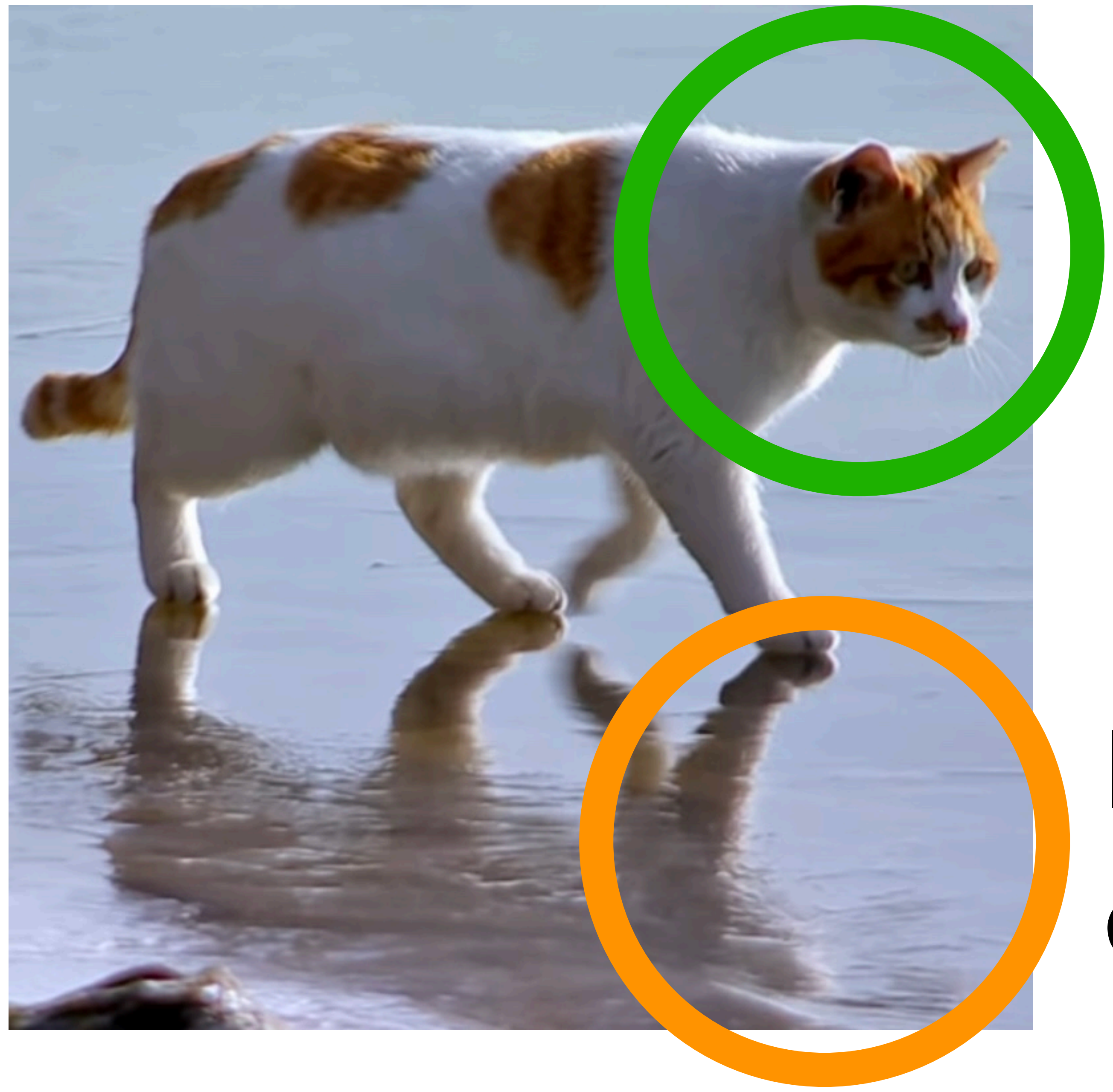
CutMix

[DeVries and Taylor (2017); Yun et al. (2019)]

TL;DR

We investigate the benefit of
Cutout and CutMix for learning features from data, and
show ERM < Cutout < CutMix in "extracting" rare features

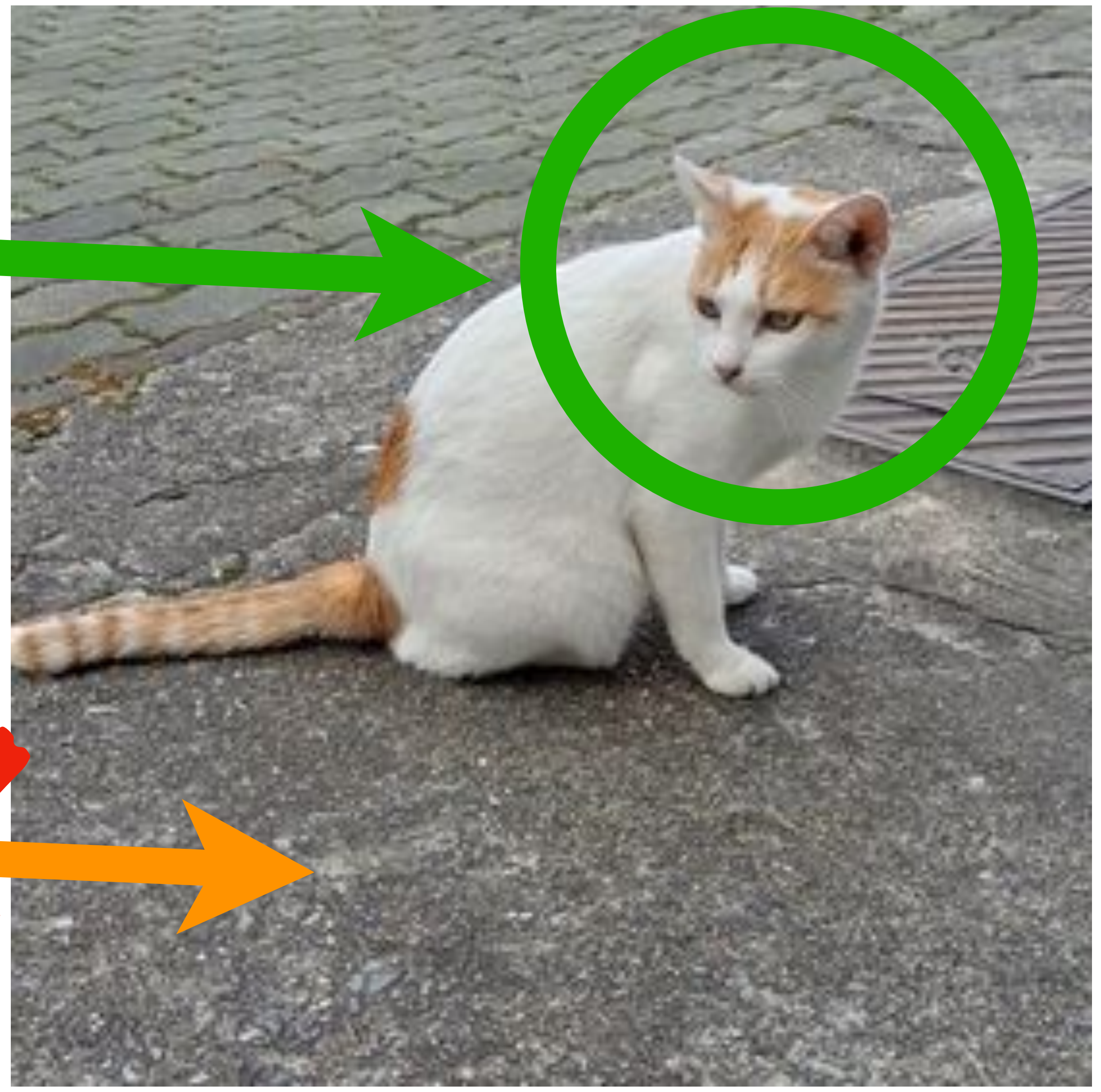
Characteristics of Images



Label-dependent **feature**
e.g. Cat's face

generalize

Label-independent **noise**
e.g. background



Training Data

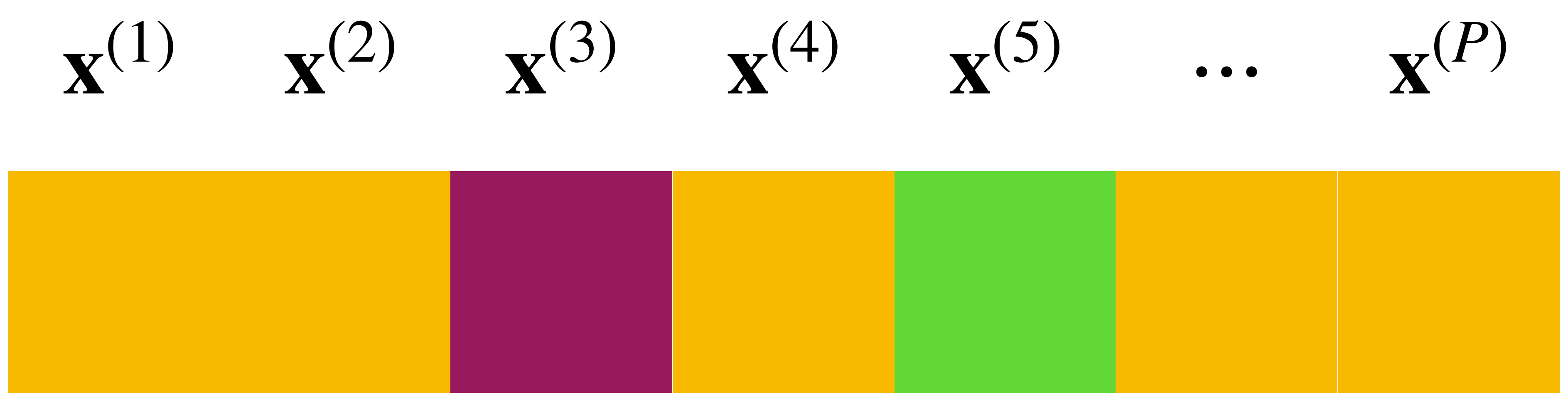
Test Data

Data Distribution

We now define our feature-noise data distribution $(\mathbf{X}, y) \sim \mathscr{D}$.

Label $y \in \{\pm 1\}$ is sampled uniformly at random, and
data point $\mathbf{X} = (\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(P)})$ consists of P "patches" of three different kinds:

$\mathbf{X} \in \mathbb{R}^{d \times P}$



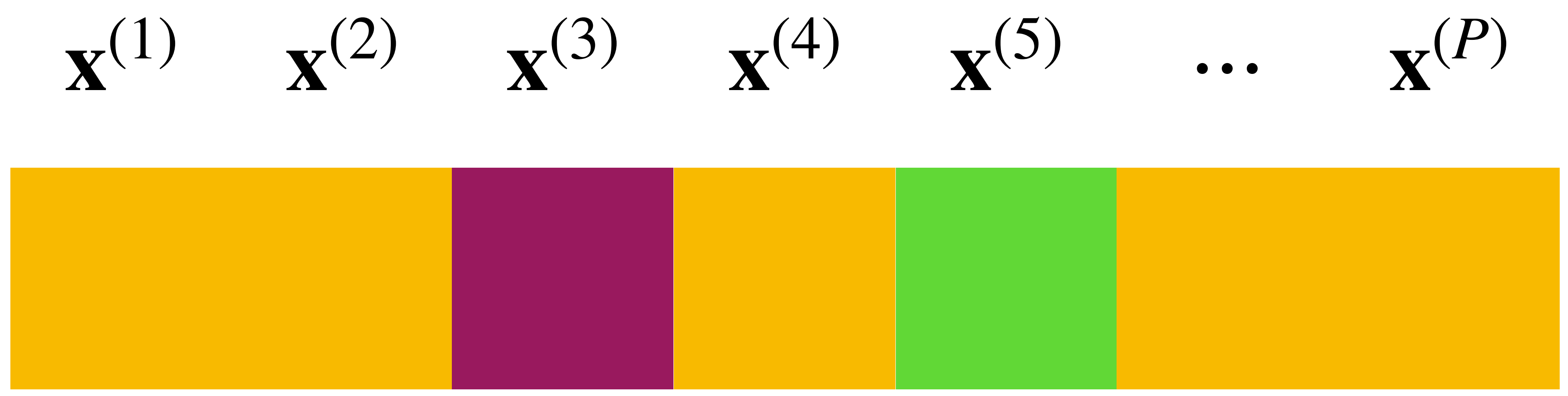
One **Feature** Patch

One **Dominant Noise** Patch

$P - 2$ **Background Noise** Patches

Data Distribution

Feature Patch. For each given label $y \in \{\pm 1\}$, there are K feature vectors $\{\mathbf{v}_{y,k}\}_{k\in[K]}$ which occur with conditional probabilities $\{\rho_k\}_{k\in[K]}$.

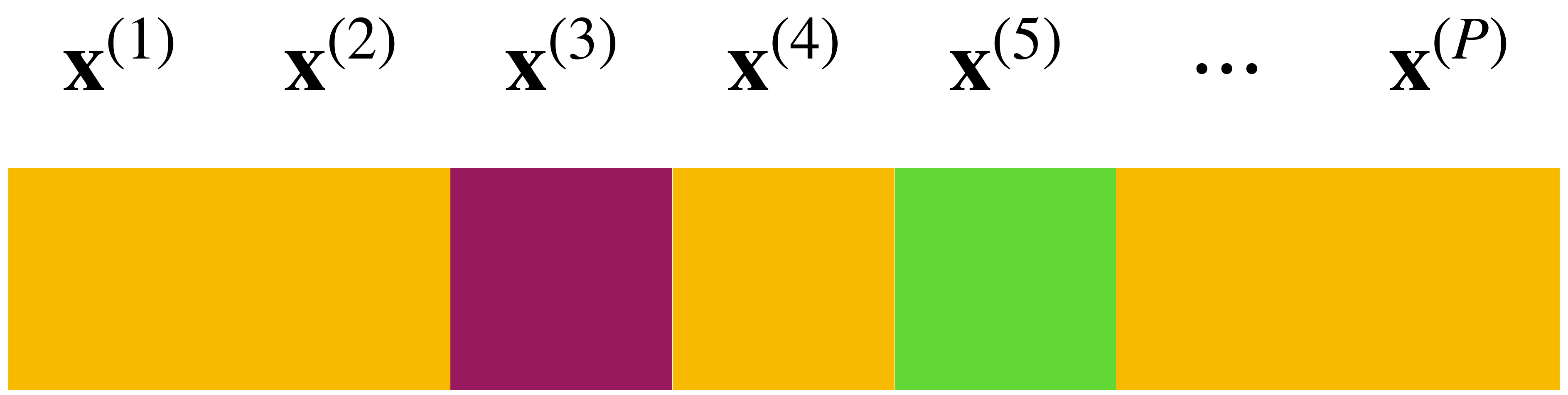


There are three kinds of features, with different levels of **rarity** (small ρ_k means rare)
Common $\mathscr{K}_C \subset [K]$, **Rare** $\mathscr{K}_R \subset [K]$, and **Extremely Rare** $\mathscr{K}_E \subset [K]$.

Given the choice of y, choose $\mathbf{v}$ from $\{\mathbf{v}_{y,k}\}_{k\in[K]}$ with probability $\{\rho_k\}_{k\in[K]}$ and position $p^* \in [P]$ uniformly at random, set $\mathbf{x}^{(p^*)} = \mathbf{v}$.

Here, $\{\mathbf{v}_{s,k}\}_{s\in\{\pm 1\},k\in[K]}$ **is orthonormal,** $\rho_1 \geq \rho_2 \geq \ldots \geq \rho_K$ **, and** $\sum_{k=1}^{K} \rho_k = 1$**.**

Data Distribution



Dominant Noise Patch.
Sample patch index $\tilde{p} \neq p^*$. Set

$$\mathbf{x}^{(\tilde{p})} = \alpha \mathbf{u} + \xi^{(\tilde{p})},$$

where $\alpha \mathbf{u}$ is "feature noise" and $\xi^{(\tilde{p})} \sim N(\mathbf{0}, \sigma_{\mathrm{d}}^2 \boldsymbol{\Lambda})$.

The feature noise is drawn $\mathbf{u} \sim$ Unif$\{\mathbf{v}_{+1,1}, \mathbf{v}_{-1,1}\}$ to model "confusing" features.

Background Noise Patch. The remaining $P-2$ patches $p \in [P] \setminus \{p^*, \tilde{p}\}$ are filled with independent and identically distributed Gaussian noise $\mathbf{x}^{(p)} = \xi^{(p)} \sim N(\mathbf{0}, \sigma_{\mathrm{b}}^2 \boldsymbol{\Lambda})$.

Here, $\boldsymbol{\Lambda} = \mathbf{I} - \sum \mathbf{v}_{s,k} \mathbf{v}_{s,k}^\top$ **and** $\sigma_{\mathrm{d}} \gg \sigma_{\mathrm{b}}$**.**

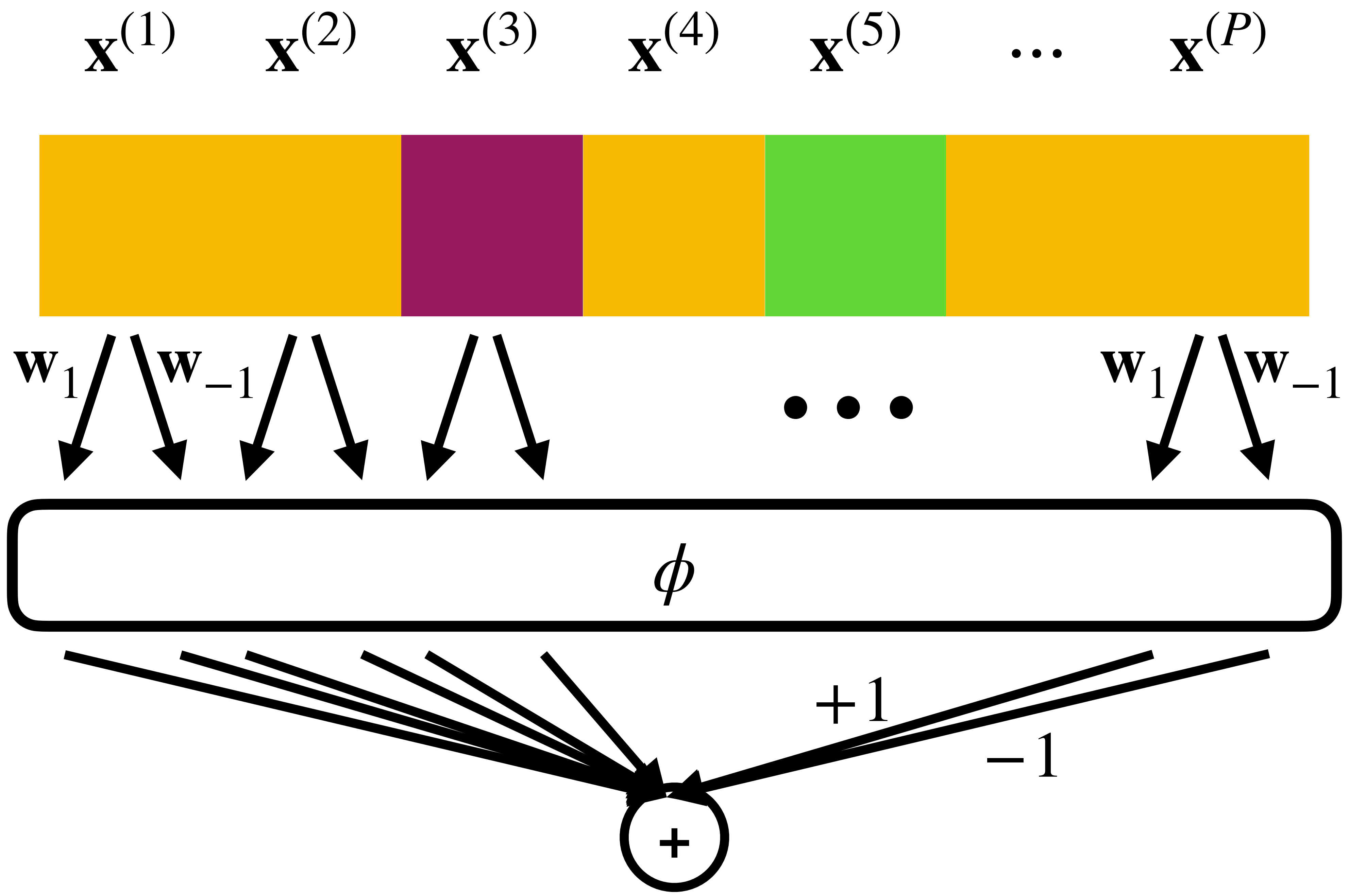
Network Architecture

We define **2-Layer CNN** $f_{\mathbf{W}} : \mathbb{R}^{d\times P} \to \mathbb{R}$**,** parameterized by $\mathbf{W} = \{\mathbf{w}_1, \mathbf{w}_{-1}\} \in \mathbb{R}^{d\times 2}$.

For input $\mathbf{X} = (\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(P)}) \in \mathbb{R}^{d\times P}$,
we define

$$f_{\mathbf{W}}(\mathbf{X}) = \sum_{p\in[P]} \phi\left(\left\langle \mathbf{w}_1, \mathbf{x}^{(p)} \right\rangle\right) - \sum_{p\in[P]} \phi\left(\left\langle \mathbf{w}_{-1}, \mathbf{x}^{(p)} \right\rangle\right).$$

If $f_{\mathbf{W}}(\mathbf{X}) \geq 0$, predict $y = +1$, and vice versa.



The activation function ϕ is a smoothed leaky ReLU activation.

Training Procedure 1: ERM

Training Data: $\{\mathbf{X}_i, y_i\}_{i\in[n]} \overset{\text{i.i.d}}{\sim} \mathscr{D}$

We define **ERM loss** as

$$\mathscr{L}_{\text{ERM}}(\mathbf{W}) := \frac{1}{n}\sum_{i\in[n]} \ell(y_i f_{\mathbf{W}}(\mathbf{X}_i)),$$

where $\ell(\,\cdot\,)$ is the logistic loss $\ell(z) = \log(1 + \exp(-z))$.

We consider GD on ERM loss $\mathscr{L}_{\text{ERM}}(\mathbf{W})$ with learning rate η.

Training Procedure 2: Cutout

Augmented Data: For each $i \in [n]$ and $\mathscr{C} \in \binom{[P]}{C}$

$$\mathbf{X}_{i,\mathscr{C}} = (\mathbf{x}^{(1)}_{i,\mathscr{C}}, \ldots, \mathbf{x}^{(P)}_{i,\mathscr{C}}) \quad \text{where} \quad \mathbf{x}^{(p)}_{i,\mathscr{C}} = \begin{cases} \mathbf{x}^{(p)}_i & \text{if } p \notin \mathscr{C} \\ \mathbf{0} & \text{otherwise} \end{cases}.$$

We define **Cutout loss** as

$$\mathscr{L}_{\text{Cutout}}(\mathbf{W}) := \frac{1}{n} \sum_{i \in [n]} \mathbb{E}_{\mathscr{C} \sim \mathscr{D}_{\mathscr{C}}} \ell(y_i f_{\mathbf{W}}(\mathbf{X}_{i,\mathscr{C}})).$$

We consider GD on Cutout loss $\mathscr{L}_{\text{Cutout}}(\mathbf{W})$ with learning rate η.

We fix $1 \leq C < P/2$. $\mathscr{D}_{\mathscr{C}}$ is a uniform distribution on $\binom{[P]}{C}$.

Training Procedure 3: CutMix

Augmented Data: For each $i, j \in [n]$ and $\mathcal{S} \subset [P]$.

$$\mathbf{X}_{i,j,\mathcal{S}} = (\mathbf{x}^{(1)}_{i,j,\mathcal{S}}, \ldots, \mathbf{x}^{(P)}_{i,j,\mathcal{S}}) \quad \text{where} \quad \mathbf{x}^{(p)}_{i,j,\mathcal{S}} = \begin{cases} \mathbf{x}_i^{(p)} & \text{if } p \in \mathcal{S} \\ \mathbf{x}_j^{(p)} & \text{otherwise} \end{cases}.$$

We define **CutMix loss** as

$$\mathcal{L}_{\text{CutMix}}(\mathbf{W}) := \frac{1}{n^2} \sum_{i,j \in [n]} \mathbb{E}_{\mathcal{S} \sim \mathcal{D}_{\mathcal{S}}} \left[\frac{|\mathcal{S}|}{P} \ell(y_i f_{\mathbf{W}}(\mathbf{X}_{i,j,\mathcal{S}})) + \left(1 - \frac{|\mathcal{S}|}{P}\right) \ell(y_j f_{\mathbf{W}}(\mathbf{X}_{i,,j,\mathcal{S}})) \right].$$

We consider GD on CutMix loss $\mathcal{L}_{\text{CutMix}}(\mathbf{W})$ with learning rate η.

$\mathcal{D}_{\mathcal{S}}$ **is a distribution such that: 1. uniformly choose size** $s \in \{0, 1, \ldots, P\}$ **and 2. uniformly choose** $\mathcal{S}$ **from** $\binom{[P]}{s}$.

Main Results - ERM

Theorem 3.1 (ERM Training)

Let $\mathbf{W}^{(t)}$ be iterates of ERM training. Then with high probability, there exists T_{ERM} such that any $T \in [T_{\mathrm{ERM}}, T^*]$ satisfies the following:

1. (Perfectly fits training set): For all $i \in [n]$, $y_i f_{\mathbf{W}^{(T)}}(\mathbf{X}_i) > 0$.

2. (Random guess on new data with rare and extremely rare features):

$$\mathbb{P}_{(\mathbf{X},y)\sim\mathscr{D}}[y f_{\mathbf{W}^{(T)}}(\mathbf{X}) > 0] \approx 1 - \frac{1}{2} \sum_{k \in \mathscr{K}_R \cup \mathscr{K}_E} \rho_k$$

Here, T^* is any large enough (polynomial in d) admissible training iterations

Main Results - Cutout

Theorem 3.2 (Cutout Training)

Let $\mathbf{W}^{(t)}$ be iterates of Cutout training. Then with high probability, there exists T_{Cutout} such that any $T \in [T_{\mathrm{Cutout}}, T^*]$ satisfies the following:

1. (Perfectly fits augmented data): For all $i \in [n]$ and $\mathscr{C} \in \binom{[P]}{C}$, $y_i f_{\mathbf{W}^{(t)}}(\mathbf{X}_{i,\mathscr{C}}) > 0$.

2. (Perfectly fits original training data): For all $i \in [n]$, $y_i f_{\mathbf{W}^{(T)}}(\mathbf{X}_i) > 0$.

3. (Random guess on new data with extremely rare features):

$$\mathbb{P}_{(\mathbf{X},y)\sim\mathscr{D}}[y f_{\mathbf{W}^{(T)}}(\mathbf{X}) > 0] \approx 1 - \frac{1}{2}\sum_{k\in\mathscr{K}_E} \rho_k$$

Main Results - CutMix

Theorem 3.3 (CutMix Training)

Let $\mathbf{W}^{(t)}$ be iterates of CutMix training. Then with high probability, there exists some $T_{\text{CutMix}} \in [0, T^*]$ that satisfies the following:

1. (Achieves a Near Stationary Point): $\left\|\nabla_{\mathbf{W}} \mathscr{L}_{\text{CutMix}}\left(\mathbf{W}^{(T_{\text{CutMix}})}\right)\right\| \approx 0$
2. (Perfectly fits original training data): For all $i \in [n]$, $y_i f_{\mathbf{W}^{(T)}}(\mathbf{X}_i) > 0$.
3. (Almost perfectly classifies test data): $\mathbb{P}_{(\mathbf{X},y)\sim\mathscr{D}}[y f_{\mathbf{W}^{(T)}}(\mathbf{X}) > 0] \approx 1$.

Main Results - Summary

Training Method \ Rarity	Common	Rare	Extremely Rare
ERM (Vanilla Training)	✓	✗	✗
Cutout	✓	✓	✗
CutMix	✓	✓	✓