Fighting Fire with Fire: Avoiding DNN Shortcuts through Priming

Chuan Wen, Jianing Qian, Jierui Lin, Jiaye Teng, Dinesh Jayaraman, Yang Gao

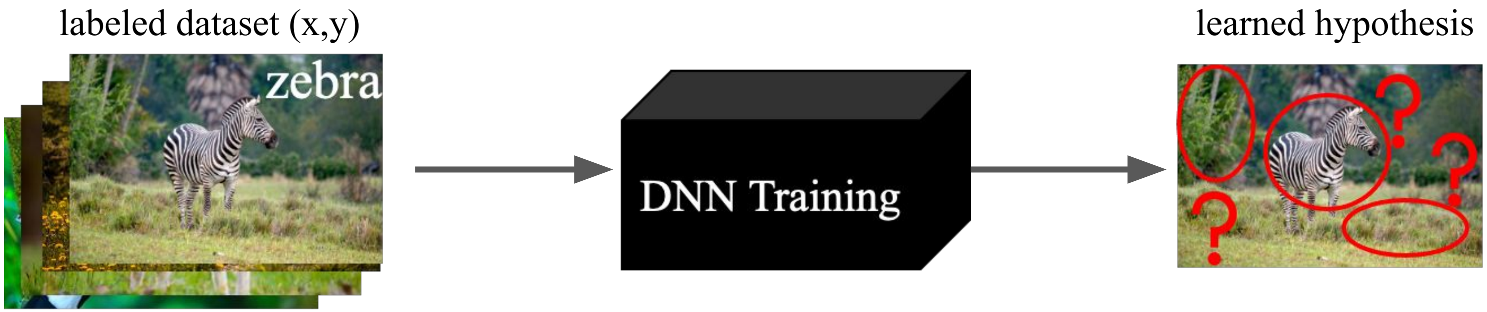
清華大學 Tsinghua University

TEXAS The University of Texas at Austin

Penn University of Pennsylvania

Shortcuts in DNNs

- DNNs tends to take the shortcut solutions rather than the intended ones.
  - Shortcuts: simple; work well in training distribution; fail in out-of-distribution region.
  - Learn the suprious correlation and hurt the generalization performance.



labeled dataset (x,y)

DNN Training

learned hypothesis

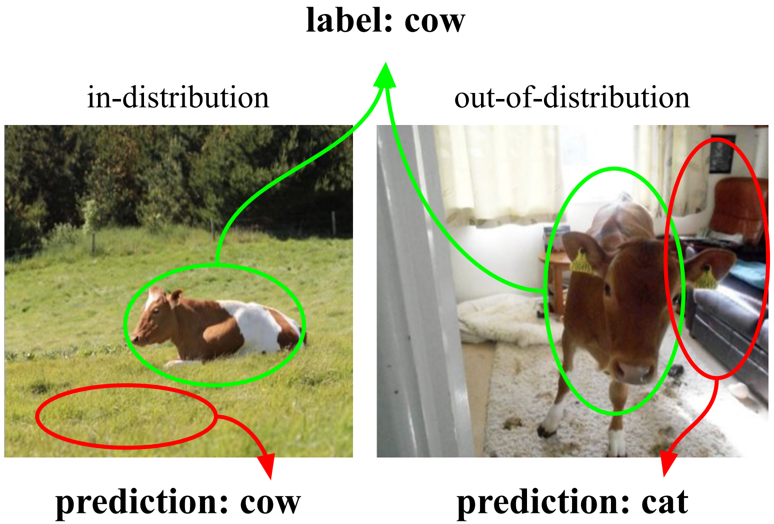
Examples: Image Classification

DNNs tend to classify the images according to the **contextual** features rather than the discriminitive **contents**.

Content: foreground object

Context: background scene

Shortcut: context → label

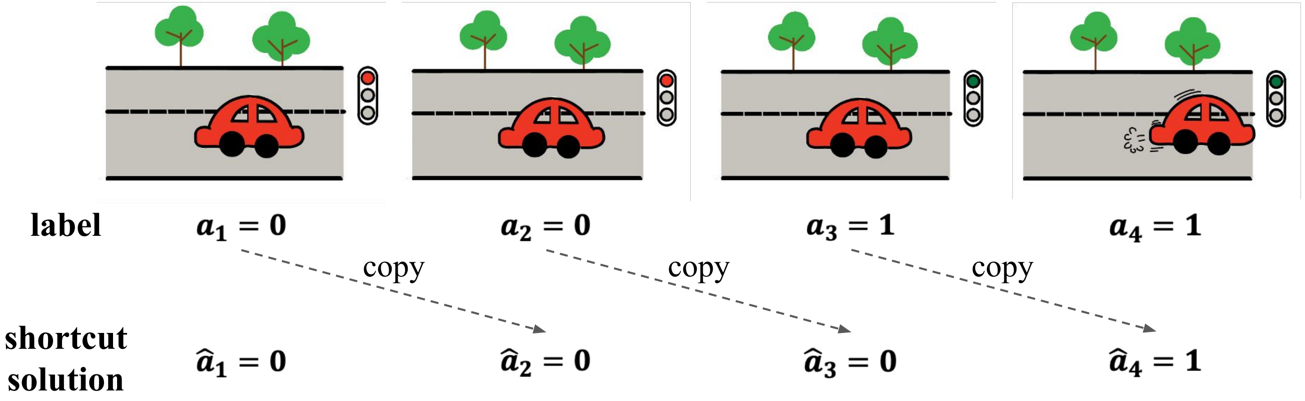


Beery, Sara, Grant Van Horn, and Pietro Perona. "Recognition in terra incognita." ECCV 2018.
Wang, Tan, et al. "Causal attention for unbiased visual recognition." ICCV 2021.

Examples: Imitation Learning

DNNs are prone to simply copy the previous action rather than learn the complex decision policy from the observations.

Shortcut: previous action → current action



label $\boldsymbol{a_1 = 0}$ $\boldsymbol{a_2 = 0}$ $\boldsymbol{a_3 = 1}$ $\boldsymbol{a_4 = 1}$

copy copy copy

shortcut solution $\boldsymbol{\hat{a}_1 = 0}$ $\boldsymbol{\hat{a}_2 = 0}$ $\boldsymbol{\hat{a}_3 = 0}$ $\boldsymbol{\hat{a}_4 = 1}$

Codevilla, Felipe, et al. "Exploring the limitations of behavior cloning for autonomous driving." ICCV 2019.
Wen, Chuan, et al. "Fighting copycat agents in behavioral cloning from observation histories." NeurIPS 2020.
Wen, Chuan, et al. "Keyframe-Focused Visual Imitation Learning." ICML 2021.

How Do Human Avoid Shortcuts?

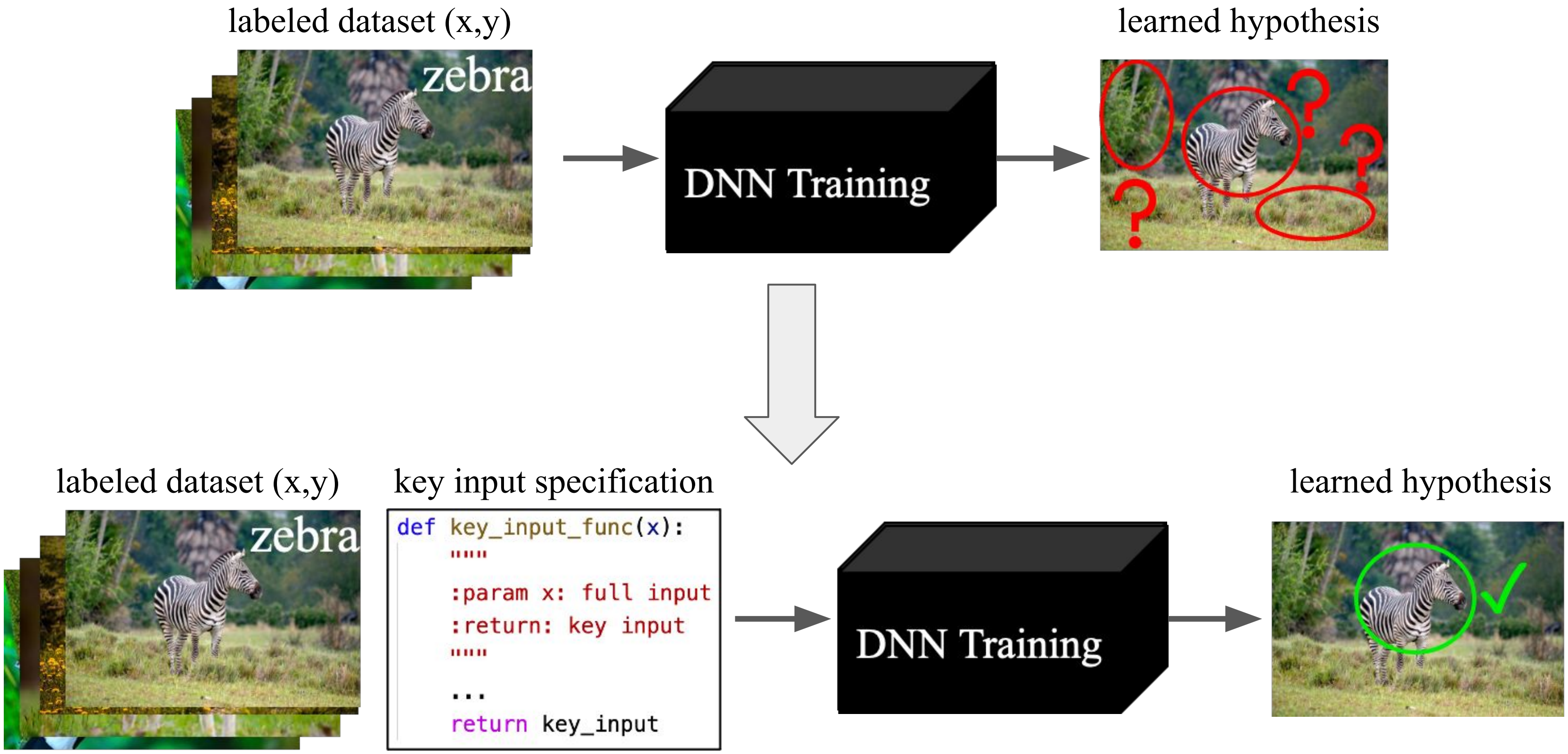
Two critical components in Human learning process:

- (1) labels
- (2) domain knowledge about which part of the input signal is key to the task

However, in supervised learning, (2) is missing.

We propose to integrate such auxiliary knowledge into DNNs, to “prime” them away from shortcuts.

PrimeNet: Prime DNNs Away from Shortcuts



labeled dataset (x,y)

DNN Training

learned hypothesis

labeled dataset (x,y)

key input specification

```
def key_input_func(x):
    """
    :param x: full input
    :return: key input
    """
    ...
    return key_input
```

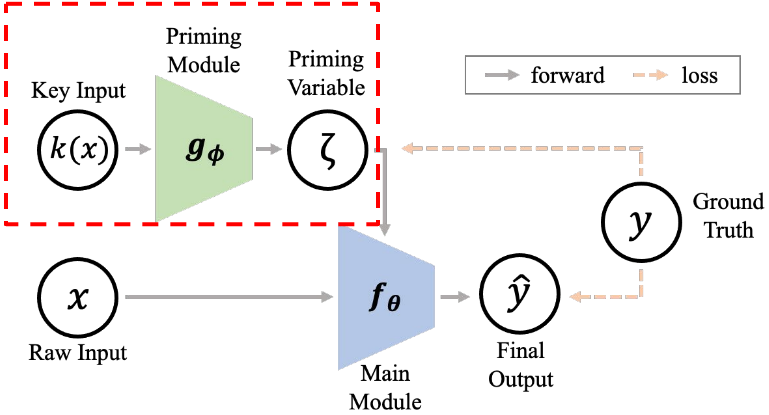
DNN Training

learned hypothesis

PrimeNet: Prime DNNs Away from Shortcuts

- Architecture of PrimeNet:
  - Priming Module
    - Key Input: $k(x)$
    - Priming Variable: $\zeta = g_\phi(k(x))$

$$\phi^* = \arg\min_\phi \frac{1}{n}\sum_{i=1}^{n} l(g_\phi(k(x_i)), y_i)$$



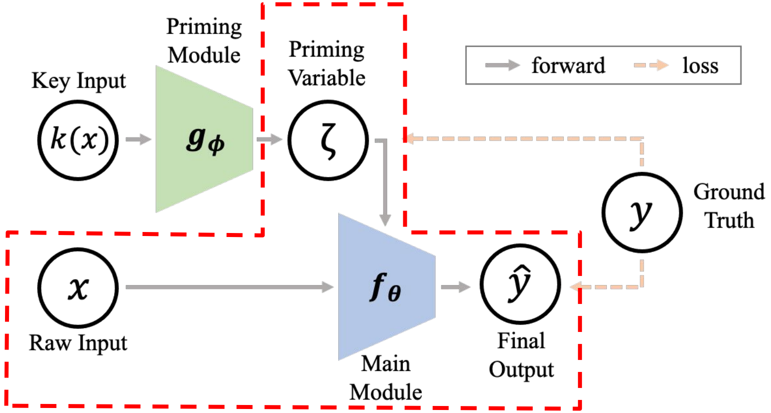
PrimeNet: Prime DNNs Away from Shortcuts

- Architecture of PrimeNet:
  - Priming Module
    - Key Input: $k(x)$
    - Priming Variable: $\zeta = g_\phi(k(x))$

$$\phi^* = \arg\min_{\phi} \frac{1}{n} \sum_{i=1}^{n} l(g_\phi(k(x_i)), y_i)$$

  - Main Module

$$\theta^* = \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} l(f_\theta(x_i, \zeta_i), y_i)$$



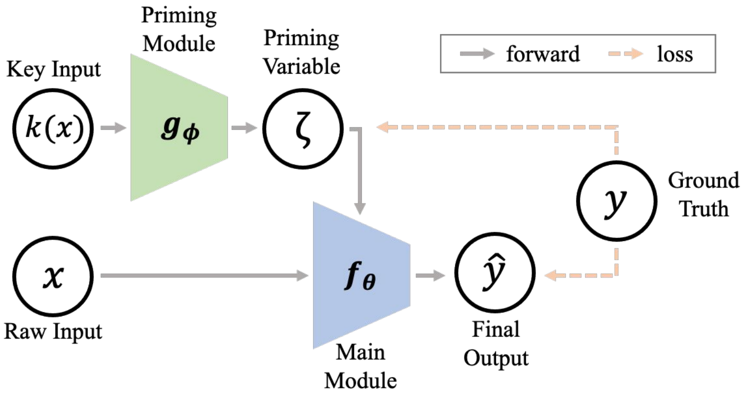
PrimeNet: Prime DNNs Away from Shortcuts

- Architecture of PrimeNet:
  - Priming Module
    - Key Input: $k(x)$
    - Priming Variable: $\zeta = g_\phi(k(x))$

$$\phi^* = \arg\min_{\phi} \frac{1}{n}\sum_{i=1}^{n} l(g_\phi(k(x_i)), y_i)$$

  - Main Module

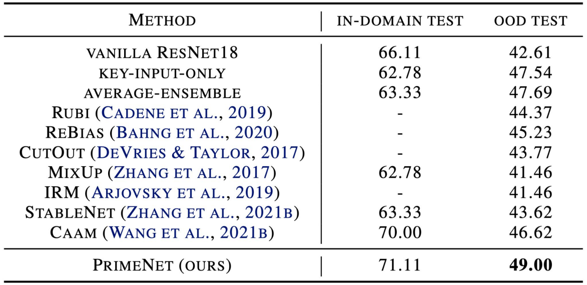
$$\theta^* = \arg\min_{\theta} \frac{1}{n}\sum_{i=1}^{n} l(f_\theta(x_i, \zeta_i), y_i)$$



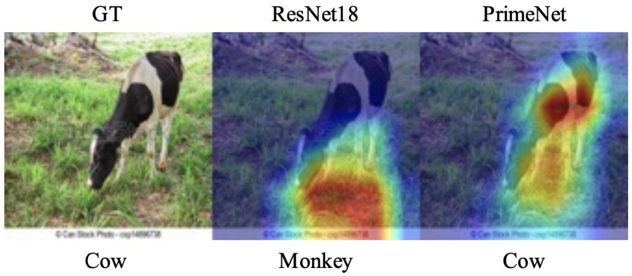
- Defination of Key Input:
  - Image Classification: image patch crop from unsupervised saliency detection.
  - Imitaiton Learning: the most recent frame.

Experimental Results

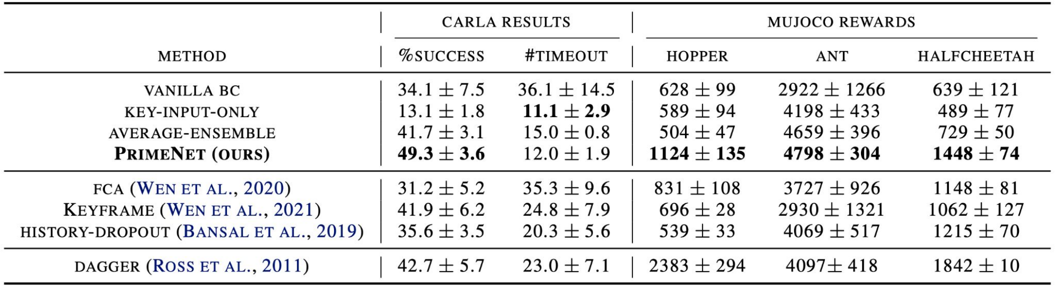
Image Classification: NICO



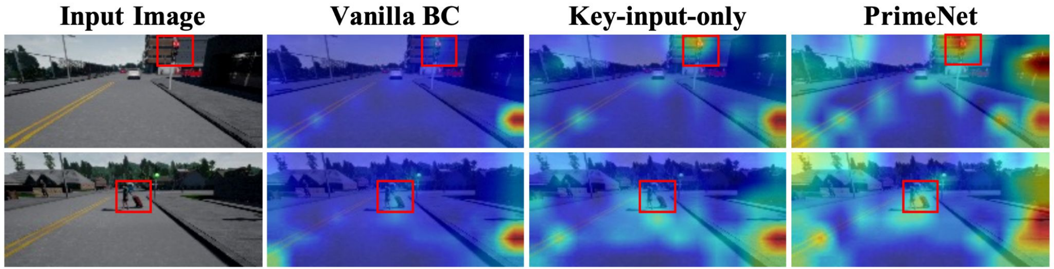
Method	In-domain test	OOD test
Vanilla ResNet18	66.11	42.61
Key-input-only	62.78	47.54
Average-ensemble	63.33	47.69
RuBi (Cadene et al., 2019)	-	44.37
ReBias (Bahng et al., 2020)	-	45.23
CutOut (DeVries & Taylor, 2017)	-	43.77
MixUp (Zhang et al., 2017)	62.78	41.46
IRM (Arjovsky et al., 2019)	-	41.46
StableNet (Zhang et al., 2021b)	63.33	43.62
CAAM (Wang et al., 2021b)	70.00	46.62
PrimeNet (ours)	71.11	**49.00**



Imitation Learning: CARLA, MuJoCo



	CARLA results		MuJoCo rewards		
Method	%Success	#Timeout	Hopper	Ant	HalfCheetah
Vanilla BC	34.1 ± 7.5	36.1 ± 14.5	628 ± 99	2922 ± 1266	639 ± 121
Key-input-only	13.1 ± 1.8	**11.1** ± **2.9**	589 ± 94	4198 ± 433	489 ± 77
Average-ensemble	41.7 ± 3.1	15.0 ± 0.8	504 ± 47	4659 ± 396	729 ± 50
PrimeNet (ours)	**49.3** ± **3.6**	12.0 ± 1.9	**1124** ± **135**	**4798** ± **304**	**1448** ± **74**
FCA (Wen et al., 2020)	31.2 ± 5.2	35.3 ± 9.6	831 ± 108	3727 ± 926	1148 ± 81
Keyframe (Wen et al., 2021)	41.9 ± 6.2	24.8 ± 7.9	696 ± 28	2930 ± 1321	1062 ± 127
History-dropout (Bansal et al., 2019)	35.6 ± 3.5	20.3 ± 5.6	539 ± 33	4069 ± 517	1215 ± 70
DAgger (Ross et al., 2011)	42.7 ± 5.7	23.0 ± 7.1	2383 ± 294	4097± 418	1842 ± 10



Experimental Results

- Visualization

Shortcut Policy

PrimeNet

copy the previous action: acceleration

Successfully stop

copy the previous action: keeping still

Successfully start up