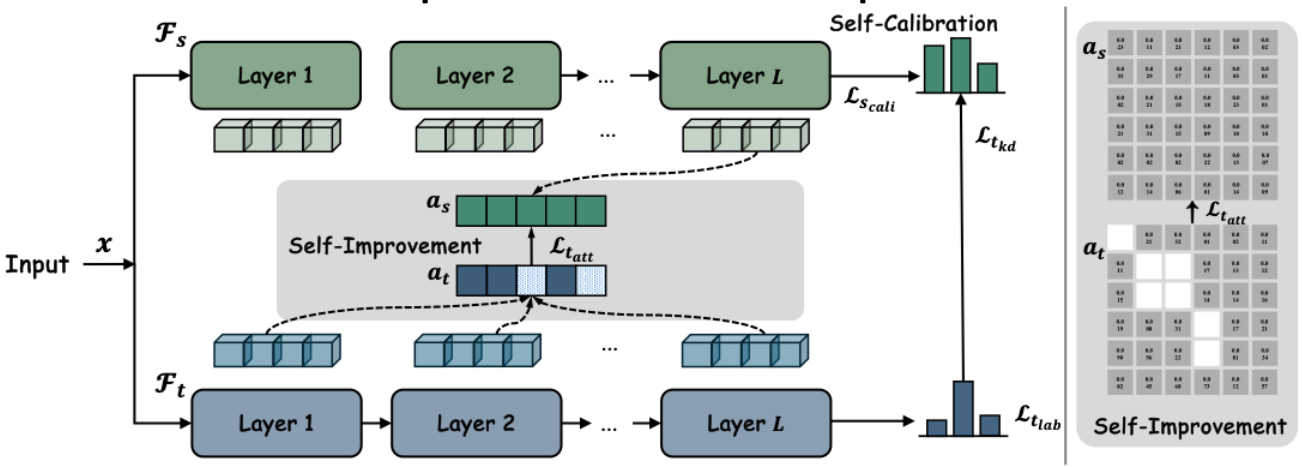
MiMu: Mitigating Multiple Shortcut Learning Behavior of Transformers

Lili ZHAO, Qi LIU, Wei CHEN, Liyi CHEN, Ruijun SUN, Min HOU, Yang WANG, Shijin WANG, Pingping REN, Jiafeng ZHOU

Frontiers of Computer Science, DOI: 10.1007/s11704-025-50448-3

Problems & Ideas

- Problems of existing methods to mitigate shortcut learning:
  - Ignore multiple shortcut learning behavior in deep learning models.
  - Existing approaches are based on prior knowledge of shortcuts in deep learning, while shortcuts a model relies on are often diverse and unknown during the training process.
- Ideas: A self-distillation model is proposed to mitigate multiple shortcut learning behaviors: it first uses a source model for calibration and then employs a target model to further address the problem of multiple shortcuts.



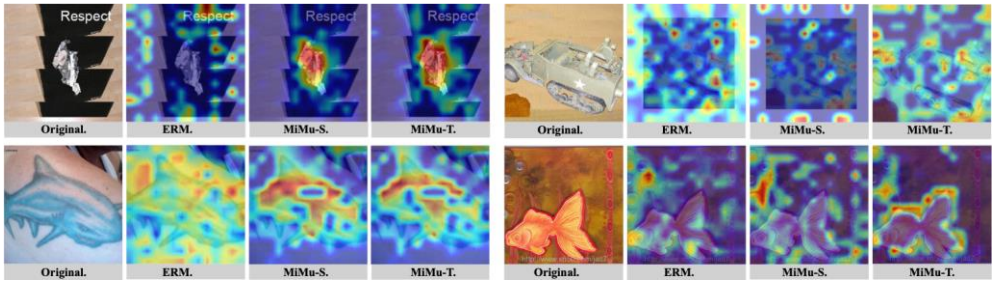
Architecture of MiMu to Mitigate Multiple shortcut learning behavior, which contains a self-calibration strategy in source model and a self-improvement strategy in target model. They jointly help mitigate multiple shortcut learning behavior and can be applied in NLP and CV domains.

Main Contributions

- Contributions:
  - In our empirical studies, we first find that models exploit shortcuts to varying degrees: they rely more heavily on strong shortcuts compared with weak ones, and their performance under multiple shortcuts often falls between the performance under a single shortcut.
  - Without prior knowledge of shortcuts, we propose a novel method MiMu, which contains self-calibration in source model and self-improvement strategy in target model following self-distillation to mitigate multiple shortcut learning behavior.
  - The proposed method-MiMu demonstrates strong versatility, and comprehensive experimental design, as it can be applied to a variety of tasks in both NLP and CV domains.

Category	Methods	Multiple	IN-9-W						
			dev	Rand-B	$\Delta^{\uparrow}$	Rand-W	$\Delta^{\uparrow}$	Rand-B+W	$\Delta^{\uparrow}$
ERM	ViT-L [66]	-	0.9796	0.7708	-	0.7357	-	0.7419	-
Augmentation & Regularization	Mixup [20]	✗	0.9789	0.7684	-0.0024	0.7345	-0.0012	0.7428	+0.0009
	Cutout [43]	✗	0.9804	0.7689	-0.0019	0.7343	-0.0014	0.7407	-0.0012
	CutMix [44]	✗	0.9753	0.7684	-0.0024	0.7307	-0.0050	0.7371	-0.0048
	AugMix [45]	✗	0.9794	0.7670	-0.0038	0.7355	-0.0002	0.7364	-0.0055
	SD [67]	✗	0.9796	0.7708	-	0.7357	-	0.7416	-
Pseudo Shortcut	LfF [21]	✗	0.9792	0.7708	-	0.7357	-	0.7409	-0.0010
	JTT [47]	✗	0.9806	0.7732	+0.0024	0.7417	+0.0060	0.7448	+0.0029
	EIIL [22]	✗	0.9811	0.7679	-0.0029	0.7300	-0.0057	0.7426	+0.0007
	DebiAN [48]	✓	0.9804	0.7710	+0.0002	0.7384	+0.0027	0.7450	+0.0031
Unknown Shortcut	COMI [8]	✗	0.9742	0.7976	+0.0268	0.7947	+0.0590	0.7795	+0.0376
Ours	**w/o s-c**	✓	0.9780	0.7968	+0.0260	0.7986	+0.0629	0.7806	+0.0387
	w/o s-i	✓	0.9751	0.7952	+0.0244	**0.8074**	+0.0717	0.7794	+0.0375
	MiMu	✓	0.9782	**0.8021**	+0.0313	0.8036	+0.0679	**0.7868**	+0.0449

The experimental results on the IN-9-W in CV.



The saliency maps in Rand-B+W (top line) and IN-R+W (bottom line). From left to right in each case are from the original image, the pre-trained ViT-Large, source model and target model in MiMu.