Conformal Prediction for Deep Classifier via Label Ranking

Jianguo Huang Huajun Xi Linjun Zhang Huaxiu Yao
Yue Qiu Hongxin Wei

ICML 2024

Paper:

Code:

Conformal Prediction

Conformal prediction is a statistical framework that generates prediction sets containing the ground-truth labels with a desired coverage guarantee, i.e.,

$$\mathbb{P}(Y \in \mathcal{C}(X)) \geq 1 - \alpha,$$

where (X, Y) is a test sample, $\mathcal{C}(X)$ represents the prediction set of test instance X and α is a significant level.

1. Any data distribution,
2. Any classifier (such as neural network, SVM, and so on).

Split Conformal Prediction

The process of Split Conformal Prediction:

1. Split a dataset into two complementary subsets, i.e., a training fold $\mathcal{D}_{tr}$ and a calibration fold $\mathcal{D}_{cal}$ whose size $|\mathcal{D}_{cal}|$ is n.
2. Train a deep learning model on $\mathcal{D}_{tr}$;
3. Define a non-conformity score function $s(\boldsymbol{x}, y)$, e.g., Adaptive Prediction sets (APS):

$$S_{aps}(\boldsymbol{x}, y, u; \hat{\pi}) := \sum_{i=1}^{o(y,\hat{\pi}(\boldsymbol{x}))-1} \hat{\pi}_{(i)}(\boldsymbol{x}) + u \cdot \hat{\pi}_{(o(y,\hat{\pi}(\boldsymbol{x})))}(\boldsymbol{x}), \quad (1)$$

   where u is an independent random variable satisfying a uniform distribution on $[0, 1]$.
4. Compute τ as the $\frac{\lceil (n+1)(1-\alpha) \rceil}{n}$ quantile of the calibration scores $\{s(\boldsymbol{x}_i, y_i) : (\boldsymbol{x}_i, y_i) \in \mathcal{D}_{cal}\}$.
5. Use the quantile to generate the uncertainty intervals for a new instance x_{test}:

$$\mathcal{C}(\boldsymbol{x}_{test}, \tau) = \{y : s(\boldsymbol{x}_{test}, y) \leq \tau\}$$

Theoretical guarantee

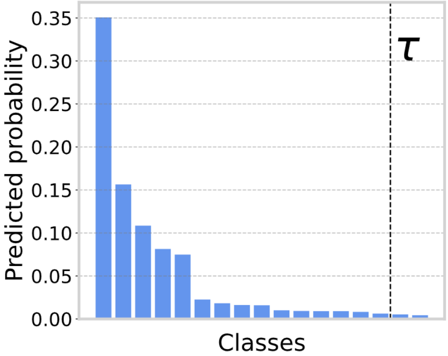
Theorem

Suppose the calibration data $(X_i, Y_i, U_i)_{i=1,\ldots,n}$ *and a test instance* $(X_{n+1}, Y_{n+1}, U_{n+1})$ *are exchangeable. Let the set-valued function* $\mathcal{C}_{1-\alpha}(\boldsymbol{x}, u; \tau)$ *satisfy the nesting property of* τ*,i.e.,* $\tau_1 \leq \tau_2 \Longrightarrow \mathcal{C}_{1-\alpha}(\boldsymbol{x}_{n+1}; \tau_1) \subseteq \mathcal{C}_{1-\alpha}(\boldsymbol{x}_{n+1}; \tau_2)$ *. For* τ*, we have the following coverage guarantee:*

$$P\left(Y_{n+1} \in \mathcal{C}_{1-\alpha}\left(X_{n+1}, U_{n+1}; \tau\right)\right) \geq 1 - \alpha.$$

Long-tailed probability

Long-tail probability distribution results in the non-conformity scores of many classes falling within the threshold τ.



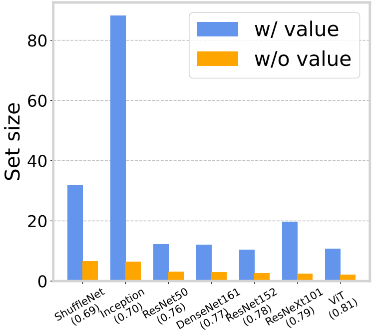
This motivates our question: **does the probability value play a critical role in conformal prediction?**

Motivation: APS without probabilities

- The score function of APS without probabilities is defined by:

$$S_{cons}(\boldsymbol{x}, y, u; \hat{\pi}) := o(y, \hat{\pi}(\boldsymbol{x})) - 1 + u.$$

- APS solely based on label ranking generates smaller prediction sets than the vanilla APS.



A theoretical analysis

Theorem

Let A_r denote the accuracy of the top r predictions for a trained model $\hat{\pi}$ on an infinite calibration set. Given a significance level α, there exists an integer k satisfying $A_k \geq 1-\alpha > A_{k-1}$. For any test instance $\boldsymbol{x} \sim \mathcal{P}_{\mathcal{X}}$ and an independent random variable $u \sim U[0,1]$, the size of the prediction set $\mathcal{C}_{1-\alpha}(\boldsymbol{x}, u)$ generated by APS without probability value can be obtained by

$$|\mathcal{C}_{1-\alpha}(\boldsymbol{x}, u)| = \begin{cases} k, & \text{if } u < \dfrac{1-\alpha-A_{k-1}}{A_k - A_{k-1}}, \\ k-1, & \text{otherwise.} \end{cases} \tag{2}$$

The expected value of the set size can be given by

$$\mathbb{E}_{u\sim[0,1]}[|\mathcal{C}_{1-\alpha}(\boldsymbol{x}, u)|] = k - 1 + \frac{1-\alpha-A_{k-1}}{A_k - A_{k-1}}. \tag{3}$$

Method: Sorted Adaptive Prediction Sets (SAPS)

Formally, the non-conformity score of SAPS for a data pair $(\boldsymbol{x}, y)$ can be calculated as

$$S_{saps}(\boldsymbol{x}, y, u; \hat{\pi}) := \begin{cases} u \cdot \hat{\pi}_{max}(\boldsymbol{x}), & \text{if} \quad o(y, \hat{\pi}(\boldsymbol{x})) = 1, \\ \hat{\pi}_{max}(\boldsymbol{x}) + (o(y, \hat{\pi}(\boldsymbol{x})) - 2 + u) \cdot \lambda, & \text{else}, \end{cases}$$

where λ is a hyperparameter representing the weight of ranking information, $\hat{\pi}_{max}(\boldsymbol{x})$ denotes the maximum softmax probability and u is a uniform random variable.

Results on various datasets

Table: Performance comparison of various methods with different error rates.

	$\alpha = 0.1$						$\alpha = 0.05$					
	Coverage			Size ↓			Coverage			Size ↓		
Datasets	APS	RAPS	SAPS	APS	RAPS	SAPS	APS	RAPS	SAPS	APS	RAPS	SAPS
ImageNet	0.899	0.900	0.900	20.95	3.29	**2.98**	0.949	0.950	0.950	44.67	8.57	**7.55**
CIFAR-100	0.899	0.900	0.899	7.88	2.99	**2.67**	0.950	0.949	0.949	13.74	6.42	**5.53**
CIFAR-10	0.899	0.900	0.898	1.97	1.79	**1.63**	0.950	0.950	0.950	2.54	2.39	**2.25**

1. SAPS generates smaller prediction sets while maintain the valid coverage.

Detailed results of ImageNet on various models

Table: The median-of-means for each column is reported over 10 different trials.

	$\alpha = 0.1$						$\alpha = 0.05$					
	Coverage			Size ↓			Coverage			Size ↓		
Datasets	APS	RAPS	SAPS	APS	RAPS	SAPS	APS	RAPS	SAPS	APS	RAPS	SAPS
ResNeXt101	0.899	0.902	0.901	19.49	2.01	**1.82**	0.950	0.951	0.950	46.58	4.24	**3.83**
ResNet152	0.900	0.900	0.900	10.51	2.10	**1.92**	0.950	0.950	0.950	22.65	4.39	**4.07**
ResNet101	0.898	0.900	0.900	10.83	2.24	**2.07**	0.948	0.949	0.950	23.20	4.78	**4.34**
ResNet50	0.899	0.900	0.900	12.29	2.51	**2.31**	0.948	0.950	0.950	25.99	5.57	**5.25**
ResNet18	0.899	0.900	0.900	16.10	4.43	**4.00**	0.949	0.950	0.950	32.89	11.75	**10.47**
DenseNet161	0.900	0.900	0.900	12.03	2.27	**2.08**	0.949	0.950	0.951	28.06	5.11	**4.61**
VGG16	0.897	0.901	0.900	14.00	3.59	**3.25**	0.948	0.950	0.949	27.55	8.80	**7.84**
Inception	0.900	0.902	0.902	87.93	5.32	**4.58**	0.949	0.951	0.950	167.98	18.71	**14.43**
ShuffleNet	0.900	0.899	0.900	31.77	5.04	**4.54**	0.949	0.950	0.950	69.39	16.13	**14.05**
ViT	0.900	0.898	0.900	10.55	1.70	**1.61**	0.950	0.949	0.950	31.75	3.91	**3.21**
DeiT	0.901	0.900	0.900	8.51	1.48	**1.41**	0.950	0.949	0.949	24.88	2.69	**2.49**
CLIP	0.899	0.900	0.900	17.45	6.81	**6.23**	0.951	0.949	0.949	35.09	16.79	**16.07**
average	0.899	0.900	0.900	20.95	3.29	**2.98**	0.949	0.950	0.950	44.67	8.57	**7.55**

Results on conditional coverage

- SAPS acquires lower ESCV, which is defined as in

$$\mathrm{ESCV}(\mathcal{C}, K) = \sup_j \max(0, 1 - \alpha - \frac{\left|\{i \in \mathcal{J}_j : y_i \in \mathcal{C}(\boldsymbol{x}_i)\}\right|}{\left|\mathcal{J}_j\right|}),$$

where $\mathcal{J}_j = \{i : |\mathcal{C}(\boldsymbol{x}_i)| = j\}$ and $j \in \{1, \ldots, K\}$.

Each-Size Coverage Violation (ESCV), which is given by

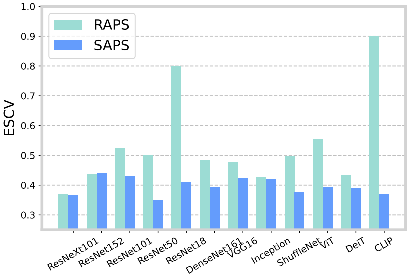


Figure: ESCV on ImageNet.

Results on different distribution

- SAPS generates smaller prediction sets when the training distribution is different from the calibration distribution.

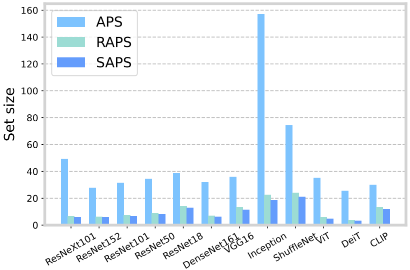
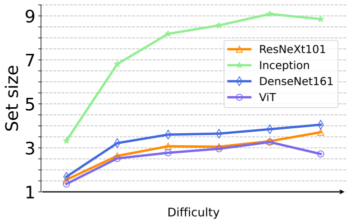


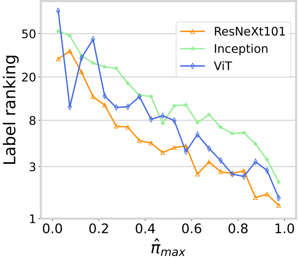
Figure: Average Set size on ImageNet-V2.

Discussion on maximum softmax probabilities

- Maximum softmax probabilities can exhibit sample difficulty, which allows SAPS to communicate instance-wise uncertainty.



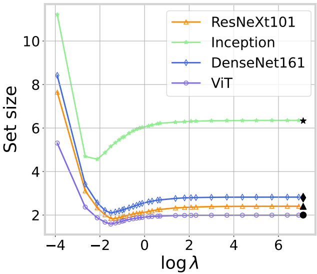
(a) Set size of various difficulties



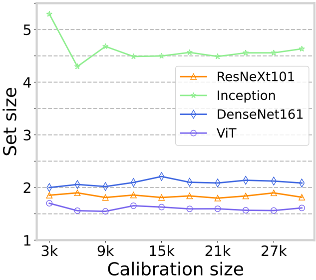
(b) The average ground-truth label ranking under different maximum softmax probabilities.

Ablation Study

- The set size of SAPS is not sensitive to variations in λ and the calibration size.



(a) Set size of various λ



(b) Set the size based on different numbers of the calibration set.

Conclusion

- **Problem**: Previous conformal prediction methods utilize unreliable softmax probabilities, which leads to suboptimal performance.
- **Analysis**: We compare the performance of APS with and without softmax probabilities. We show that APS solely based on label ranking generates smaller prediction sets than the vanilla APS.
- **Method**: We present SAPS, a simple and effective conformal prediction score function that discards almost all the probability values except for the maximum softmax probability.

ArXiv: `https://arxiv.org/abs/2310.06430`
Code: `https://github.com/ml-stat-Sustech/conformal_prediction_via_label_ranking`