ICML
International Conference
On Machine Learning

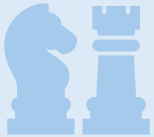
Bar-Ilan University
אוניברסיטת בר־אילן

On Temperarure Scaling and Conformal Prediction of Deep Classifiers

Lahav Dabah, Tom Tirer

ICML 2025

Uncertainty Quantification



Quantifying molel's uncertainty is critical, especially in high stakes applications.

2

Two main used methods:

1. **Calibration**
2. **Conformal Prediction**

Temperature Scaling Calibration

Key Idea

Adjusts confidence scores to better match actual correctness probabilities.

How?

- Divide the model's logits by a scalar named **temperature**
- Optimize the temperature to improve calibration

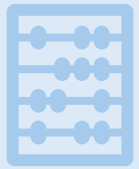
Conformal Prediction

Key Properties

- Works with any model (**black-box access**)
- Acts as a **post-processing** step

Key Idea

Outputs a **set of possible classes** that is guaranteed to contain the true label with a user-defined confidence level.

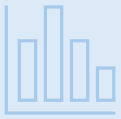


Evaluating Conformal Prediction?

1. AvgSize – average size of prediction sets
2. TopCovGap – worst-case gap in **coverage** across classes.

**Note the for both metrics – the lower the better*

Temperature Scaling **Calibration** Before Conformal Prediction?



We applied **Temperature Scaling Calibration** before running Conformal Prediction.

Table 1. **Prediction Set Size.** AvgSize metric along with T^* and accuracy for dataset-model pairs using LAC, APS, and RAPS algorithms with $\alpha = 0.1$, CP set size 10%, pre- and post-TS calibration.

		Accuracy(%)		**AvgSize**			**AvgSize after TS**		
Dataset-Model	T^*	Top-1	Top-5	LAC	APS	RAPS	LAC	APS	RAPS
ImageNet, ViT-B/16	1.180	83.9	97.0	2.22	10.10	1.93	2.23	19.27	2.34
CIFAR-100, ResNet50	1.524	80.9	95.4	1.62	5.31	2.88	1.57	9.14	4.96

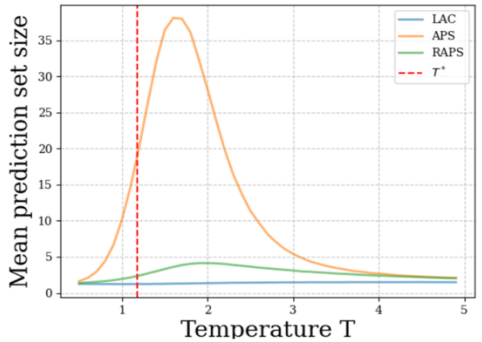
Table 2. **Coverage Metrics.** MarCovGap and TopCovGap metrics for dataset-model pairs using LAC, APS, and RAPS algorithms with $\alpha = 0.1$, CP set size 10%, pre- and post-TS calibration.

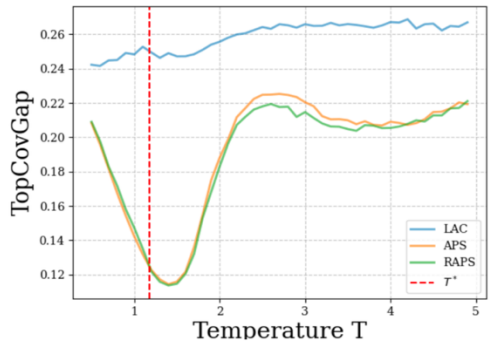
	MarCovGap(%)			**MarCovGap TS(%)**			**TopCovGap(%)**			**TopCovGap TS(%)**		
Dataset-Model	LAC	APS	RAPS	LAC	APS	RAPS	LAC	APS	RAPS	LAC	APS	RAPS
ImageNet, ViT-B/16	0	0	0	0.1	0.1	0	24.8	14.2	14.7	24.9	12.2	12.5
CIFAR-100, ResNet50	0.1	0	0	0	0.1	0	13.9	12.6	11.7	12.9	9.0	7.9

Temperature Scaling **Before** Conformal Prediction?

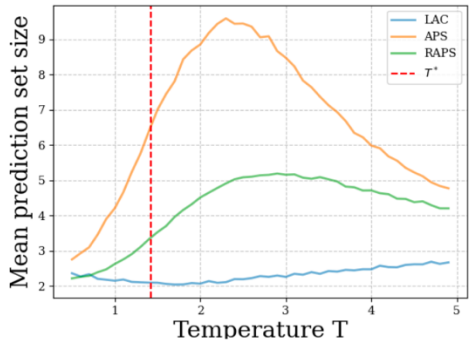
We applied **Temperature Scaling** with a range of temperatures **before** running Conformal Prediction.

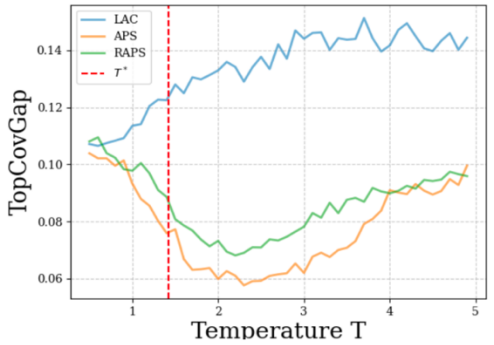
ImageNet, ViT





CIFAR-100, DenseNet121



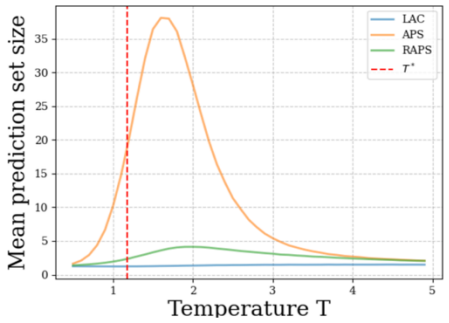


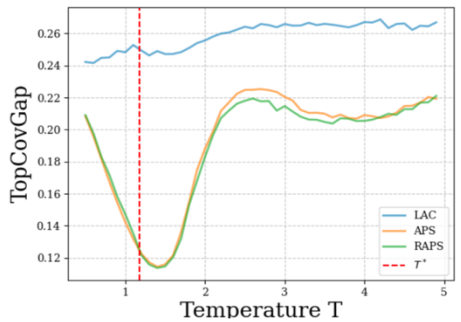
Temperature Scaling **Before** Conformal Prediction?

We applied **Temperature Scaling** with a range of temperatures **before** running Conformal Prediction.

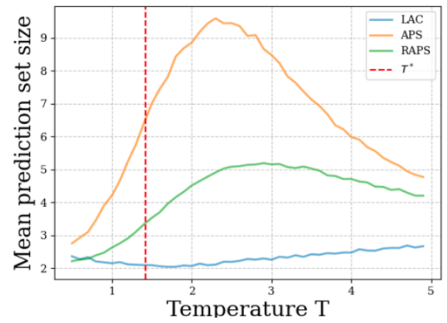
- Adaptive CP methods show similar patterns across all datasets:
  - **AvgSize** rises, peaks, then declines.
  - **TopCovGap** drops, reaches a minimum, then increases.
- **trade-off** between set size and conditional coverage — tunable via T
- We developed a **mathematical theory** that explains this non-monotonic effect.

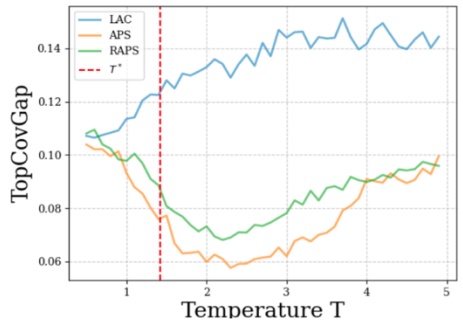
ImageNet, ViT





CIFAR-100, DenseNet121





Practical Guidelines to Practitioners

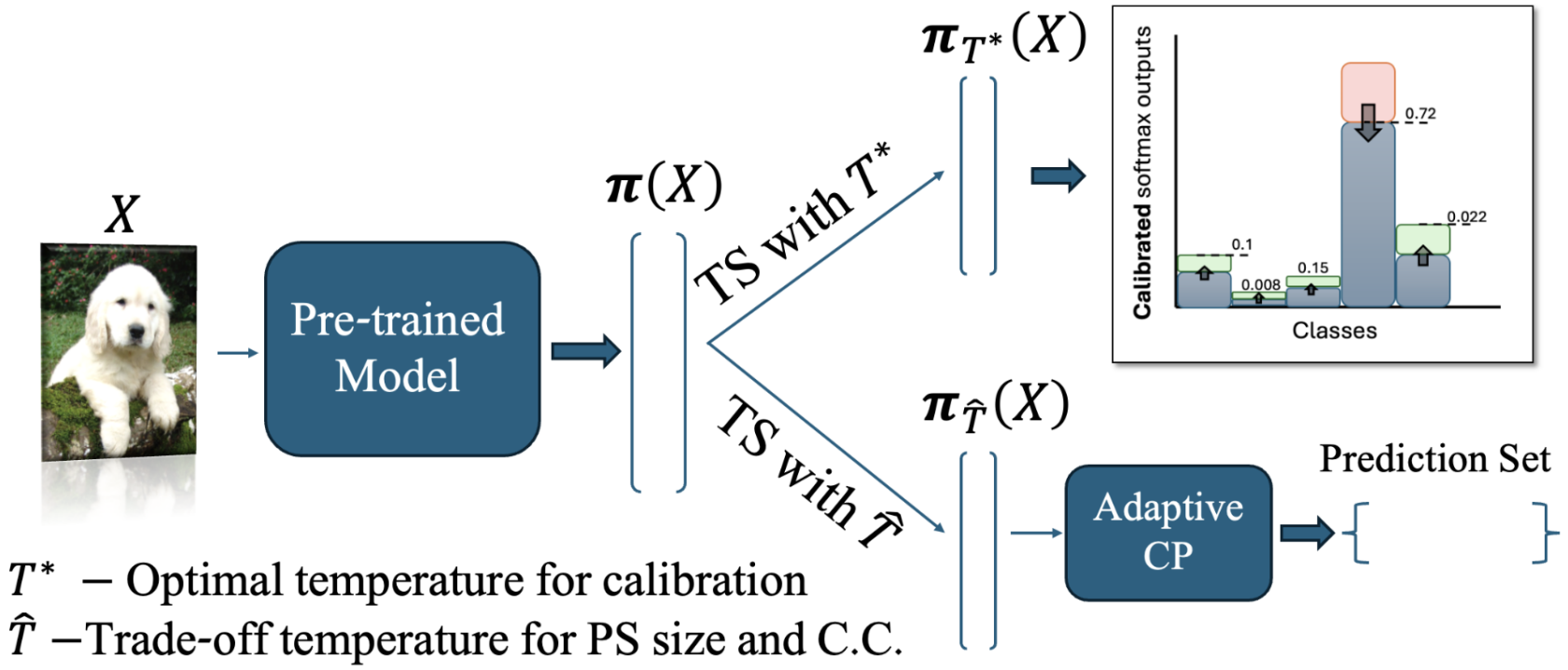
Use TS with Two temperature values

- T^* for calibration
- $\hat{T}$ for controlloing CP trade-off

How to calculate $\hat{T}$?

- Previous AvgSize/TopCovGap curves used large evaluation sets and 100 trials — **not practical** in real-world use.
- We suggest using evaluation set for choosing $\hat{T}$, without violating exchangeability.
- We empirically show that using $n_{eval} \approx n_{cal}$ in a single trial leads to good approximation of $\hat{T}$.

Practical Guidelines to Practitioners



Prioritize prediction set sizes: use $\hat{T} \to 0$

Prioritize conditional coverage: use $\hat{T} \to T_c$

Thank you for your attention!

For more details and experiments, check out our paper and code:

Paper

Code

lahavdabah@gmail.com