(Not so) Technical Summary of Conformal Semantic Image Segmentation

Summary by Marcelo Jiménez – May, 2025

May 9, 2025

1 Summary

This paper introduces a method for quantifying predictive uncertainty in semantic image segmentation using Conformal Prediction (CP). You want to know how confident your model is about its predictions in an image segmentation task? This is the paper for you.

1.1 So what did these guys do?

- Took a model that performs semantic image segmentation (you know, identifying which pixels of the photo of a cow belong to the cow, and which ones are grass, clouds, sky, and so on).
- Implemented an improved version of conformal prediction that uses a **loss threshold** instead of a simple certainty threshold.
- The improved conformal prediction outputs a **set of classes for every pixel** (more classes if it's uncertain, fewer if it's confident).
- Visualized this with **varisco heat maps**: pixels with more classes look ”hotter”, and pixels with fewer classes look ”cooler” — really useful to quickly spot where the model is confident and where it's not.

2 Conformal Prediction

In a nutshell, conformal prediction is a way to evaluate the uncertainty of a model in its classification. Instead of just having one class as output, it gives you a set of classes as output and assures you that (on average) the correct label belongs to one of those classes with a user-defined level of confidence, let's say, 90% confidence.

Why not use only the softmax probabilities of the last layer of the classification? Like, my model says this pixel belongs to a cow with 95% probability, is that not enough? No! Because softmax tends to be overconfident, even when it is making a mistake in the classification.

So, conformal prediction works like this:

- You define a confidence threshold, let's say 90%.

- You take a calibration dataset (a small group of labeled data not used during training) and make predictions on it.
- You collect the softmax probabilities of the predicted classes.
- Sort these probabilities from lower to higher.
- Take the quantile corresponding to the 90% (if you have 100 ordered probabilities, it would be the 90th value in the list).
- Call that number the "Certainty threshold".

Then, you go to the outside world and make inferences about unseen data. The model will output a final softmax layer with probabilities for each class. If a class has a probability higher than the certainty threshold, then that class is added to the output set. And that's it! That's how you get an output set that is statistically guaranteed to contain the correct class 90% of the time.

2.1 Note:

If you want a lot of certainty, like 99%, then the output sets will have a lot of classes. On the other hand, if you are fine with low certainty, like 60%, the output set will probably have one or two classes (it really depends on your model's performance).

3 Improved Conformal Prediction: Conformal Risk Control (CRC)

Standard conformal prediction is great if you just want to know whether the true label is in the prediction set with a given probability (like 90%). But in semantic segmentation, not all errors are the same! Imagine your model messes up identifying a pedestrian versus a patch of grass in an autonomous driving vehicle, obviously, one mistake is way worse than the other.

That's where CRC comes in. Instead of just saying "I'm 90% sure the right label is in there," it goes a step further. CRC says, "I want to make sure my mistakes aren't too big on average." Basically, it controls the **empirical risk**, meaning it keeps the average error below a level you set (let's say 0.1).

3.1 Mathematical Foundation and Procedure

- **Step 1: Define LAC (Least Ambiguous Set-Valued Classifier)** First, they use the **LAC function** to decide whether a class should be included in the prediction set. It's like a binary filter:

  $$T_\lambda(p) = \begin{cases} 1 & \text{if } p \geq 1 - \lambda \\ 0 & \text{otherwise} \end{cases}$$

  So basically, if the **softmax score** for a class is **high enough**, it gets included. Otherwise, it's left out.

- **Step 2: Multi-Labeled Mask Creation** Then, they generate a **multi-labeled mask** using the LAC function. For each pixel, the mask is defined as:

  $$C_\lambda(X)_{ij} = \{k : f_{ijk}(X) \geq 1 - \lambda\}$$

  This means that for every pixel (i, j), the mask contains **all classes whose scores exceed the threshold**.

- **Step 3: Calibrate the Risk** Now they calculate the **empirical risk** from the calibration data:

  $$\hat{R}_n(\lambda) = \frac{1}{n} \sum_{i=1}^{n} L_i(\lambda)$$

  The idea is to find the **smallest** λ such that the **average loss** is within the acceptable risk level α:

  $$\frac{n}{n+1}\hat{R}_n(\lambda) + \frac{B}{n+1} \leq \alpha$$

  Here:

  - $L_i(\lambda)$ = Loss function for the i-th sample (measuring how much the prediction set misses the true label).
  - B = Upper bound of the loss function.

- **Step 4: Dichotomic Search for Optimal** λ To find the right λ, they use a **dichotomic search** (basically a binary search) because the **empirical risk decreases monotonically** as λ increases. This way, they efficiently find the **smallest threshold** that meets the risk requirement.

- **Step 5: Predict with the Calibrated Threshold** Once they have the optimal λ, they make predictions on new data. For each pixel, they create a prediction set including **all classes whose softmax score** is **greater than** $1 - \lambda$.

- **Step 6: Visualize with Varisco Heatmaps** Finally, they generate **Varisco heatmaps** to visually represent the **number of classes per pixel**. Pixels with **more classes** look "hotter" (more uncertain), while **confident pixels** look "cooler" (fewer classes).

4 Datasets Used

- Cityscapes: Urban scenes with fine-grained annotations.
- ADE20K: General-purpose semantic segmentation.
- LoveDA: Aerial images for land cover segmentation.

5 Results

The method achieves robust uncertainty quantification, with the empirical risk closely matching the desired confidence levels. The Varisco heatmaps highlight areas of high uncertainty, particularly around object boundaries.

6 Limitations

The quality of the prediction sets heavily depends on the accuracy of the initial model predictions. High-confidence errors can still result in incorrect calibration.

7 References

Mossina, L., Dalmau, J., & Andéol, L. (2024). Conformal Semantic Image Segmentation: Post-hoc Quantification of Predictive Uncertainty. arXiv:2405.05145 [cs.CV].