Claims:

- Board Achieved claims
    - A-Our approach aims to add interpretability to the “dark” knowledge transferred from the teacher to the shallower student model.
    - B-Global Explanation loss forces the student prototypes to be close to teacher prototypes
        - Irrespective of the input, the model can tell which parts/regions it may focus on to make decisions.
            - Can we feed it input where this does not work?
        - Based on the activations of prototypes, the model recognizes the image
            - Show activations?
            - Difficult (8/10)
    - C-Patch-Prototype Correspondence loss enforces the local representations of the student to be similar to that of the teacher
        - It mimics the local representations of the teacher for which prototypes become active.
            - Does this happen?
            - Difficult (5/10)
        - Unlike Fitnets, which mimics the entire feature map of a teacher for knowledge transfer, we propose to mimic local representations of the teacher that activate prototypes.
            - Is it really different from Fitnets?
            - Very difficult (9/10)

- Achieved claims
    - D-Our student is more faithful to the teacher in retaining similar prototypes to make decisions compared to baseline student. (Figure 1)
        - D-This is evident for a given test image, with visualization of top-k prototypes. (Teacher, baseline student, paper student) (Subjective)
            - Is Baseline valid?
            - Is Paper student valid?
    - E(B & C)-Reason for performance is “Global Explanation loss” **and** “Patch-Prototype Correspondence loss” to achieve the objective of transferring the interpretability of the teacher to the student ((B | C) or (B & C)?)
        - Can we achieve similar results with just 1?
- Experimental results:
    - F-Table 2. Results of Proto2Proto student
    - G-Figure 2, Effect on Active Patches by varying τ using L2 as a metric
    - H-The proposed model is interpretable without forgoing accuracy.
    - Table 3, PShare results & Table 4
    - In terms of accuracy, minor improvement over RKD but on interpretability scores, significantly better, Table 5.
    - Figure 4.
- Open questions:
    - Use different dataset in inference (**transfer learning**???) (difficulty 3/10)

        - Extremely well doable and interesting extension.
    - Add transparency metrics from Yang et al (1/10)
    - Apply the proposed evaluation metrics in a ensemble setup (7/10).