Knowledge Distillation Changes Where Models Look

By Thomas Day - Supervised by Dr. Alina Bialkowski

The University Of Queensland
AUSTRALIA

Motivation

- Deep learning powers critical vision tasks (collision avoidance, medical imaging).
- Models are ballooning in size, with state of the art vision models containing billions of parameters.
- Compression is necessary, but we must be able to trust not only what compressed models predict, but why.

Key Definitions

- Compression:
  - Techniques to shrink models while preserving accuracy.
- Knowledge Distillation (KD):
  - Teacher → student training. Student learns from the teacher's output, transferring "dark knowledge".
- Explainable AI (XAI)
  - Activation maps (Saliency, Integrated Gradients) that show which pixels/regions influences a prediction.

Research Questions

RQ1: How does compression through knowledge distillation comparatively affect standard explainability methods in computer vision models?

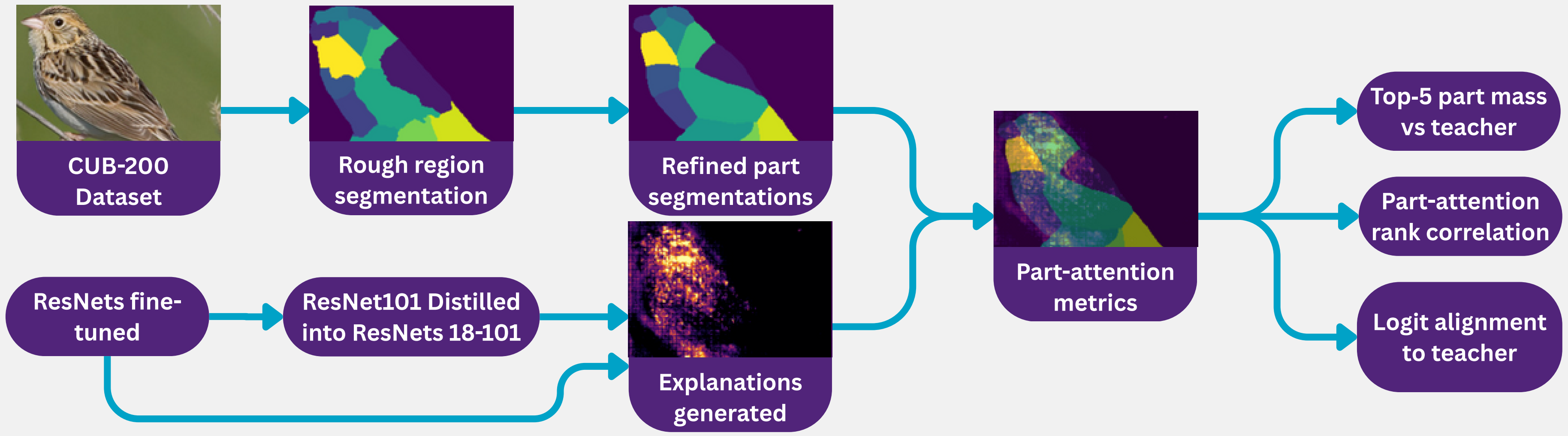
RQ2: How does the trade-off between compression ratio, model accuracy, and explainability vary across distilling different models of different sizes?

RQ3: How can the change in a model's decision-making process be quantified?

Key Findings

- Stronger knowledge distillation (lower α) improves student accuracy for smaller students.
- Stronger knowledge distillation increases student-teacher explanation alignment.
- Knowledge distillation concentrates student focus on the teacher's most important parts.
- Model outputs and explanations co-vary under KD.

Experimental Setup



Training Results

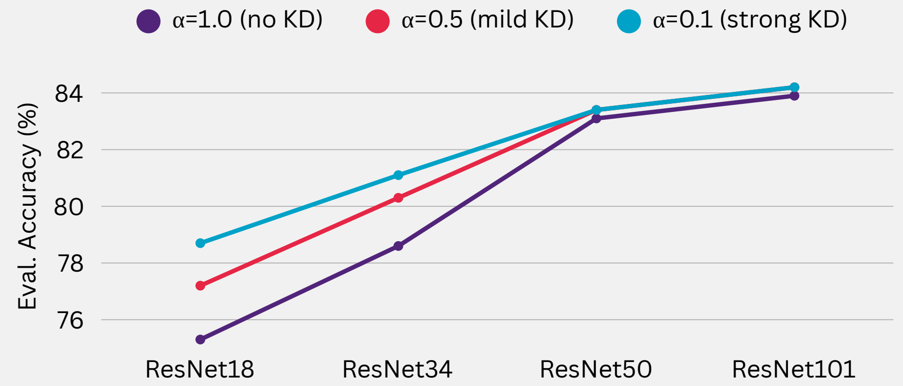


Figure 1: Evaluation accuracy by architecture under training regimes of no KD, mild KD, and strong KD.

Focus on Top-5 Teacher Parts

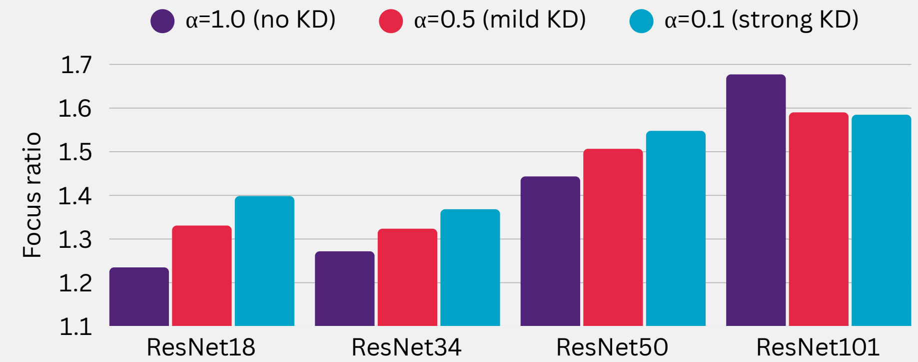


Figure 3: Focus on the Teacher's (ResNet101 α=1.0) top-5 parts, measured as the mean ratio of attribution mass in the teacher's top-5 parts vs outside.

Teacher Attention Rank Alignment

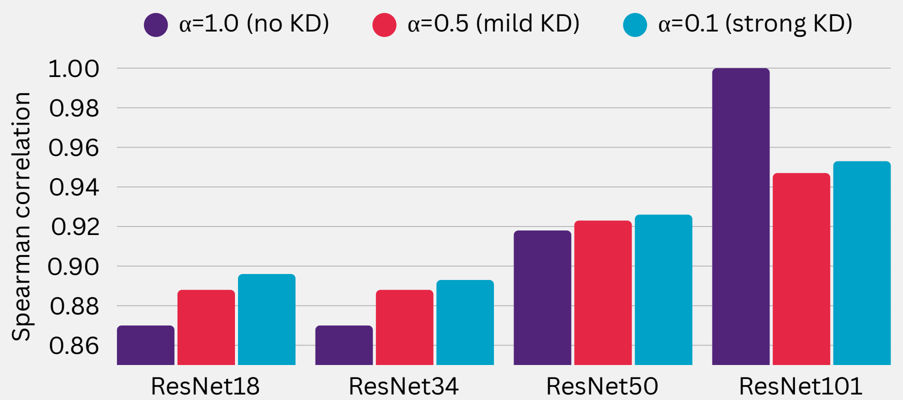


Figure 2: Attention alignment to teacher (ResNet101 α=1.0) measured as the Spearman correlation between student and teacher per-image part-importance vectors

Teacher Logit Alignment

	ResNet101	ResNet50	ResNet34	ResNet18
α=1.0 (no KD)	1.00	0.52	0.28	0.26
α=0.5 (mild kd)	0.76	0.70	0.48	0.38
α=0.1 (strong KD)	0.81	0.76	0.56	0.50

Figure 4: Output logit similarity between student and teacher (ResNet101 α=1.0). Measured as mean Spearman correlation between model output vectors.