Deliverable 01

This report presents an analysis comparing the internal neuron representations learned by ResNet18 models trained on **ImageNet** (object classification) and **Places365** (scene classification). We have used neuron-description annotations to uncover what concepts neurons in different layers respond to and how model training data impacts learned representations.

ImageNet learned 374 total unique concepts and Places365 learned 427 total unique concepts. As shown in Figure 20, the total unique concepts learned increases from layer 2 to layer 4.

- ImageNet: 68 → 123 → 279
- Places365: 65 → 135 → 313

This also shows that Places365 encodes more diverse concepts overall and develops richer representations in deeper layers.

As shown in **Figure 17 and 18** (Top 15 Concepts Bar Charts), the concepts learned by most neurons across both models are texture-based patterns such as: dotted (most common in both models), checker, textile, stripes, grid, stripe, and green.

These concepts appear early and are shared across both object and scene recognition tasks, highlighting their low-level importance.

The Venn diagram in Figure 19 reveals **180 shared concepts** between models. However, the **Jaccard Similarity** of **28.99%** – shows relatively low overlap, indicating that the models focus on different features. **ImageNet** (focused on objects) leans slightly more on fine-grained textures and object-part cues like cat, aircraft, grille whereas **Places365**, tuned for scenes, includes **scene-level semantics** such as bathroom, kitchen, lobby, bedrooms (Figure 3 & Figure 10 – respectively).

Some additional analysis includes Figure 21 (Boxplots by Layer) shows that in both models, **concept similarity scores increase with depth**, but Places365 neurons in layer4 show significantly higher confidence (average similarity ~0.32 vs. ImageNet's ~0.25).

Both models share foundational textures but diverge significantly in deeper layers. In ImageNet, the progression is:

- layer2: basic patterns (e.g., checker)
- layer3: object parts (fur, mesh)
- layer4: full objects or environments (restaurant, kitchen, grille)

In Places365:

- layer2: texture primitives (dotted, grid)
- layer3: mix of texture and layout
- layer4: full scenes or places (attic, lobby, locker)

This confirms that **ImageNet representations remain more object-focused**, while **Places365 representations transition toward spatial semantics**.

These findings confirm that **the choice of training dataset strongly influences the nature and structure of internal representations** in neural networks, even when the architecture remains identical.