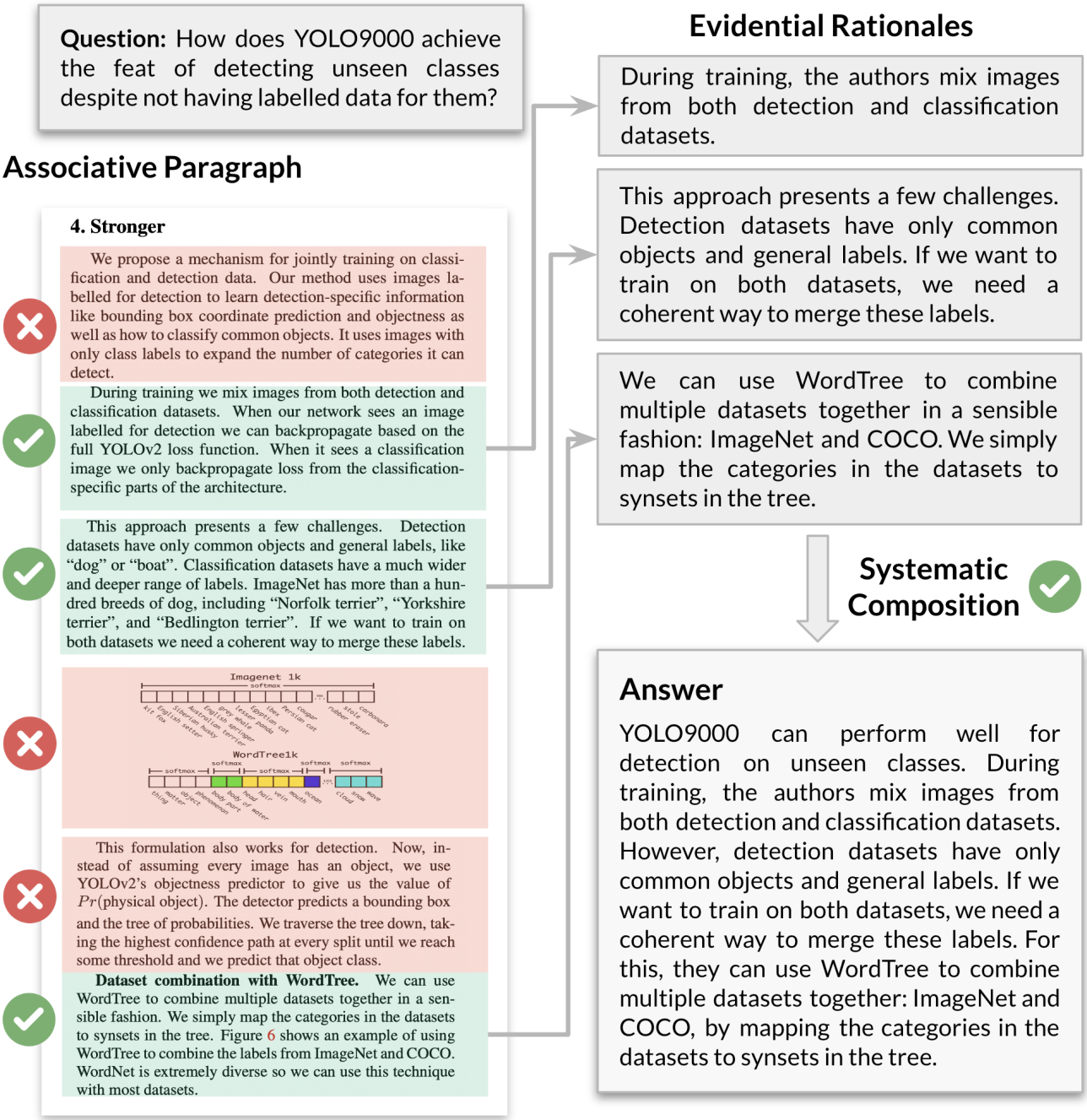

Question: How does YOLO9000 achieve the feat of detecting unseen classes despite not having labelled data for them?
Associative Paragraph
4. Stronger
We propose a mechanism for jointly training on classification and detection data. Our method uses images labelled for detection to learn detection-specific information like bounding box coordinate prediction and objectness as well as how to classify common objects. It uses images with only class labels to expand the number of categories it can detect.
During training we mix images from both detection and classification datasets. When our network sees an image labelled for detection we can backpropagate based on the full YOLOv2 loss function. When it sees a classification image we only backpropagate loss from the classification-specific parts of the architecture.
This approach presents a few challenges. Detection datasets have only common objects and general labels, like "dog" or "boat". Classification datasets have a much wider and deeper range of labels. ImageNet has more than a hundred breeds of dog, including "Norfolk terrier", "Yorkshire terrier", and "Bedlington terrier". If we want to train on both datasets we need a coherent way to merge these labels.
Imagenet 1k
softmax
WordTree1k
softmax
This formulation also works for detection. Now, instead of assuming every image has an object, we use YOLOv2's objectness predictor to give us the value of Pr(physical object). The detector predicts a bounding box and the tree of probabilities. We traverse the tree down, taking the highest confidence path at every split until we reach some threshold and we predict that object class.
Dataset combination with WordTree. We can use WordTree to combine multiple datasets together in a sensible fashion. We simply map the categories in the datasets to synsets in the tree. Figure 6 shows an example of using WordTree to combine the labels from ImageNet and COCO. WordNet is extremely diverse so we can use this technique with most datasets.
Evidential Rationales
During training, the authors mix images from both detection and classification datasets.
This approach presents a few challenges. Detection datasets have only common objects and general labels. If we want to train on both datasets, we need a coherent way to merge these labels.
We can use WordTree to combine multiple datasets together in a sensible fashion: ImageNet and COCO. We simply map the categories in the datasets to synsets in the tree.
Systematic Composition
Answer
YOLO9000 can perform well for detection on unseen classes. During training, the authors mix images from both detection and classification datasets. However, detection datasets have only common objects and general labels. If we want to train on both datasets, we need a coherent way to merge these labels. For this, they can use WordTree to combine multiple datasets together: ImageNet and COCO, by mapping the categories in the datasets to synsets in the tree.