Multi-task learning of objects and parts – Initial Problem Description, Solution and Plan

Hailin Weng

Introduction

This project's primary objective is to develop and evaluate a computational vision model, specifically a Convolutional Neural Network (CNN), capable of simultaneous training on object name labels and property labels derived from CSLB property norms.

Problem Description

In the field of computer vision, the goal is to enable machines to emulate human vision, making informed decisions based on visual data. While deep learning techniques have made significant strides in tasks like image classification, they have predominantly focused on recognizing objects based on predefined labels, such as "EAGLE" or "CAR." However, a new dimension arises when considering object properties, referred to as property norms.

Stakeholders in this problem include researchers, developers, and businesses seeking a more comprehensive understanding of image content beyond conventional object labels. The core issue lies in the limitations of existing models, like AlexNet [1], VGG [2] and ResNet [3], which excel at image categorization but fail to capture the intrinsic attributes or parts of objects. This narrow, object-centric perspective lacks the depth required to identify shared properties among different object categories.

This problem manifests itself in applications where understanding object properties is crucial, such as advanced image search engines, augmented reality, or robotics. Historically, the emphasis has been on label-centric training, exemplified by AlexNet, which, groundbreaking as it was, followed this paradigm. The challenge arises when machines need to identify shared properties across objects, such as recognizing that both "EAGLES" and "WOLVES" have "EYES." Existing models lack the granularity to achieve this.

Existing solutions, like the influential AlexNet, prioritize raw categorization performance and may struggle to recognize objects in unfamiliar contexts where shared properties are more indicative than holistic labels.

Who has the problem?

Researchers, developers, and businesses trying to further understand image content beyond just object labels.

What is the problem?

While models like AlexNet excel at categorizing images using predefined labels, they don't capture the intrinsic attributes or parts of objects. This object-centric perspective lacks a deeper, attribute-centric view that considers shared properties across different object categories.

Where does the problem manifest itself?

In applications where understanding the properties of objects is crucial, such as advanced image search engines, augmented reality, or robotics.

Historically, the emphasis has been majorly on label-centric training. AlexNet, for instance, was groundbreaking for its time, yet focused on this paradigm [1]. The problem emerges when we need machines to identify shared properties between objects, like recognizing both EAGLES and WOLVES have "EYES". Existing models lack this granularity.

Why other solutions do not work?

Previous solutions, like the influential AlexNet, focused on raw categorization performance. Such models might misclassify or fail to recognize objects in unfamiliar contexts where shared properties might be more indicative than holistic labels.

How might your work be innovative or helpful?

This project seeks to pivot from the traditional label-centric training to property norms. By training a model on these property norms using the CSLB specification [4], we intend to provide a richer, more nuanced understanding of images. It becomes imperative to ask: Can a model trained this way outperform or complement traditional models? And, can such a model identify parts of an object category it hasn't seen?

Research Questions:

1. How does a property norm-trained model's performance compare with one trained solely on object labels?
2. How do internal representations differ in both models?
3. Can a property norm-trained model correctly identify object parts it hasn't been trained on?

Goals and requirements

- Acquire, align, and integrate the CSLB property norms dataset and THINGS dataset [5], considering concepts mapping between them.
- Become familiar with the Pytorch framework for building, training and evaluating neural networks.
- Become familiar with influential deep convolutional neural network models for vision (e.g. AlexNet; VGG, ResNet).
- Extend and evaluate a deep convolutional neural network model for vision to predict semantic properties of objects, as well as labels.
- Analyse the internal representations of the model and compare them to the internal representations in a model trained on the labelling task alone.

Success Criteria

1. Implementation: Implement AlexNet using PyTorch with comparable accuracy on the ImageNet dataset.
2. Property Norms Extension: Adapt AlexNet to predict properties based on the CSLB norms, achieving a predefined accuracy on a validation set.
3. Comparative Analysis: Compare performance metrics (e.g., top-1 accuracy) between the traditional and property norms-based models.
4. Representation Insight: Analyze internal representations of both models to understand differences in feature learning.
5. Generalization: Demonstrate the property norms model's ability to identify untrained object parts.
6. Documentation: Offer clear documentation for model replication and evaluation.
7. Robustness: Validate models against varied real-world scenarios and datasets.

CSLB Property Norms Infilling

In addition to the above, we propose a method called "property norm infilling" to enhance the accuracy of property norms datasets like CSLB. Property norm datasets often lack exhaustive feature information, as participants tend to provide only what they consider informative. For instance, only six animals in the property norms dataset may be labeled with the feature "does breathe," although this property applies to all animal concepts.

To address this limitation, we leverage ChatGPT (GPT4)'s API to programmatically acquire "yes" or "no" responses for all features in the property norms dataset. This approach would allow us to create a more extensive set of true features.

In addition, there are 638 concepts and 2725 features in the property norms. This gives ~ 2 million possible combinations of concept and feature pair. It is not feasible to test all 2 million combinations. However, we can target features that are true of one concept and not true of a **similar** concept. For example, giraffe has the property “has a neck” in the property norms, whilst elephant does not. Giraffe and elephant will be similar in terms of their existing CSLB properties (has legs, is a mammal, etc).

We can measure the similarity of concepts based on the cosine similarity of their feature vectors.

The algorithm involves:

Starting with binarized CSLB property norms (P0) with 1 for feature present and 0 for feature absent.

Adding all features of P0 to a new dataset (P1).

Computing cosine similarity for every pair of concepts in P0.

For the N most similar concept pairs:

a. Identifying pairs of concepts (c1 and c2).

b. Selecting a feature from c1 that is absent in c2:

i. Testing the feature with ChatGPT and adding it to c2 in P1 if the answer is yes.

c. Selecting a feature from c2 that is absent in c1:

i. Testing the feature with ChatGPT and adding it to c1 in P1 if the answer is yes.

This iterative process can be repeated for P1, expanding the dataset further.

This approach enhances the completeness and accuracy of property norms datasets, contributing to a more robust computational vision model.

Mapping Between CSLB Property Norms and THINGS Dataset

Many of the object concepts present in the THINGS dataset exhibit significant overlap with those found in the CSLB property norms. To effectively train our computer vision model, it is imperative that we possess both images and feature vectors for these shared object concepts. While some instances align seamlessly, with concept names matching between the two datasets, challenges arise when the same concept appears under different names. For instance, "courgette" in the CSLB norms may correspond to "zucchini" in the THINGS database.

To seamlessly integrate these datasets, we meticulously crafted a mapping process that establishes a robust correspondence between CSLB names and THINGS names. This mapping is pivotal for ensuring that we can accurately associate the correct images and features with each concept during

the model training process. The mapping results in a structured dictionary known as "namemap," characterized by entries in the form of namemap["cslbname"] = "THINGS name."

It is noteworthy that the initial mapping process, while comprehensive, contained inaccuracies and mismatches. To address this, we undertook a refinement process using WordNet and ConceptNet. However, we encountered challenges and inconsistencies in these efforts. Subsequently, we leveraged the power of GPT-4's API to filter and optimize the mapping, resulting in an improved mapping dictionary stored as "updated_namemap.json."

Through this meticulous mapping effort, we have successfully identified and corrected erroneous concept pairs, ensuring accurate mappings between CSLB Property Norms and the THINGS dataset. These meticulously curated concept pairs, along with their associated images and feature vectors, will serve as the foundational building blocks for training our feature prediction model. This robust dataset empowers our model to make precise predictions about object attributes, thereby enhancing the success and accuracy of our research endeavors.

Expected Project Development Plan

CSC3002: Multi-task learning of objects and parts

Student: Hailin Weng

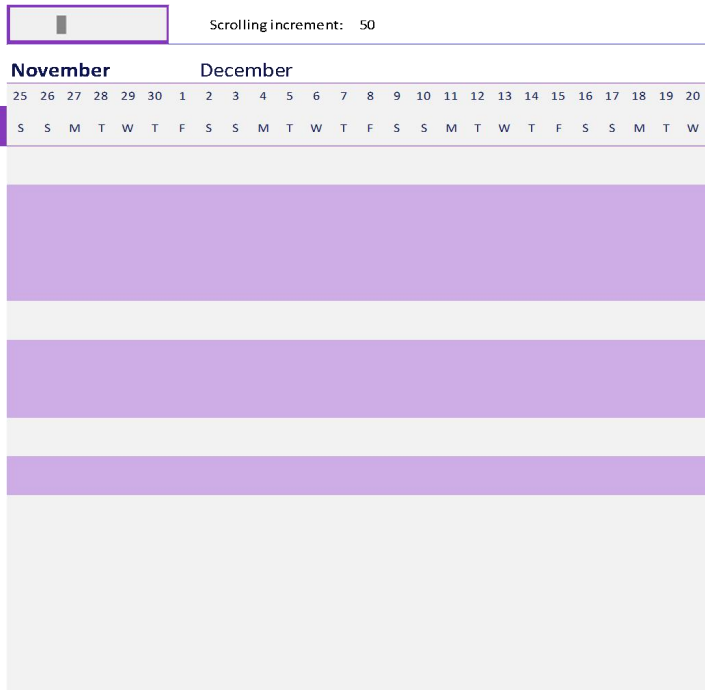
Supervisor: Barry Devereux

Project start date: 06/10/2023

Milestone marker: 1

Scrolling increment: 50

Milestone description	Assigned to	Progress	Start	Days
Foundation				
Machine Learning		70%	06/10/2023	100
Deep Learning		40%	09/10/2023	97
Computer Vision		30%	11/10/2023	95
Implementation				
AlexNet (Pytorch)		100%	13/10/2023	90
Property Norms		80%	23/10/2023	83
Develop (Solution)				
My own CNN (Project)		10%	09/11/2023	66
Project Description, Solution Approach and		100%	01/01/2024	13
Demo			12/01/2024	1
Dissertation				
Final Dissertation			1/13/2024	91
Demonstration			13/04/2024	1



References

[1] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). *ImageNet Classification with Deep Convolutional Neural Networks*. In *Advances in Neural Information Processing Systems 25 (NIPS 2012)*. Available at: https://papers.nips.cc/paper_files/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html.

[2] Simonyan, K., & Zisserman, A. (2014). *Very Deep Convolutional Networks for Large-Scale Image Recognition*. *arXiv preprint arXiv:1409.1556* [cs.CV]. Available at: https://doi.org/10.48550/arXiv.1409.1556.

[3] He, K., Zhang, X., Ren, S., and Sun, J. (2015). *Deep Residual Learning for Image Recognition*. *Tech report*, *arXiv:1512.03385* [cs.CV]. Available at: https://doi.org/10.48550/arXiv.1512.03385.

[4] Devereux, B.J., Tyler, L.K., Geertzen, J. et al. (2014). *The Centre for Speech, Language and the Brain (CSLB) concept property norms*. *Behavior Research*, 46, 1119–1127. Available at: https://doi.org/10.3758/s13428-013-0420-4.

[5] Hebart, M. N., Dickter, A. H., Kidder, A., Kwok, W. Y., Corriveau, A., Van Wicklin, C., & Baker, C. I. (2019). *THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images*. *PLOS ONE*. Available at: https://doi.org/10.1371/journal.pone.0223792.