Research and Implementation of Multi-Dataset Training for Image Classification with Discrepant Taxonomies

Master Thesis Presentation

Björn Buschkämper

Technical Faculty, Bielefeld University

September 8, 2025

Outline

The Challenge: Limited Scope of Traditional Models

- Traditional image classification models are trained on specific datasets
- Each model recognizes only a predefined set of categories
- Multiple models needed for different domains = inefficient storage and deployment

Current Approaches:

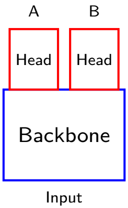
- **Transfer Learning**: Adapt pre-trained models to new tasks
- **Multi-head Architecture**: Shared backbone + task-specific heads
- **Multi-task Learning**: Train on multiple tasks simultaneously

Problem: Still requires separate models or heads for each domain

Our Idea: Universal Model vs. Multi-Head

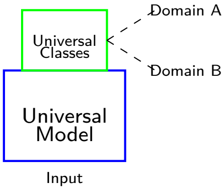
Multi-Head Approach

- Shared backbone
- Task-specific heads
- Automatic feature distillation
- Domain alignment challenges



Universal Model Approach

- Single shared model
- Universal output layer
- Predefined concept mapping
- Static domain conversion



Key Insight: Build explicit concept-to-domain mappings to train a single universal model

Our Method in 3 Steps

Step 1: Cross-Domain Prediction → **Step 2:** Relationship Building → **Step 3:** Universal Model Training

1. **Cross-Domain Prediction**: Train domain-specific models, run inference on other domains
2. **Relationship Building**: Extract meaningful relationships from cross-domain predictions and create universal taxonomy
3. **Universal Model Training**: Create and train a single model using universal taxonomy

Cross-Domain Prediction Process

Goal: Discover relationships between classes from different datasets

Process:

1. Train domain-specific models
2. Run each model on images from *all other* domains
3. Create probability matrices $P_{ab}(i,j) =$ probability of classifying class c_i^a as class c_j^b

Example: CIFAR-100 model predicting on Caltech-101 images

- CIFAR-100 "automobile" $\rightarrow$ Caltech-101 "car_side" (high probability)
- CIFAR-100 "dog" $\rightarrow$ Caltech-101 "dalmatian" (moderate probability)
- CIFAR-100 "airplane" $\rightarrow$ Caltech-101 "butterfly" (low probability)

$$P_{ab}(i,j) = \frac{M_{ab}(i,j)}{\sum_{k=1}^{|C_a|} M_{ab}(i,k)} \tag{1}$$

Challenge: Selecting Relevant Relationships

Problems with raw probability matrices:

- Noisy predictions from imperfect models
- Unknown number of true relationships
- Different datasets have different scales of similarity

Solution: Develop & evaluate multiple relationship selection methods

Methods Evaluated:

1. Naive Thresholding
2. Most Common Foreign Prediction (MCFP)
3. Density Thresholding
4. Relationship Hypothesis

Evaluation Metrics:

- Edge Difference Ratio (EDR)
- Precision & Recall
- F1 Score

Relationship Selection Methods Explained

1. **Naive Thresholding**:

$$\text{select\_relationships}(P_{ab}) = \{(i,j) \mid P_{ab}(i,j) \geq t\} \tag{2}$$

2. **Most Common Foreign Prediction (MCFP)**:

$$\text{select\_relationships}(P_{ab}) = \{(i,j) \mid j = \text{argmax}_{j'} P_{ab}(i,j')\} \tag{3}$$

3. **Density Thresholding**: Select minimum relationships covering $p\%$ of probability mass
4. **Relationship Hypothesis**: Assumes relationships based on shared concepts should have equal probabilities. For each class, find optimal k relationships by minimizing:

$$\sum_{j=1}^{k} \left| X_i(j) - \frac{1}{k} \right| + \sum_{j=k+1}^{|C_b|} X_i(j) \tag{4}$$

where $X_i(j)$ are sorted probabilities in descending order.

The Need for Controlled Ground Truth

Problem: Real datasets lack clear inter-dataset relationships

- **WordNet**: Text-based relationships don't match visual similarities
- **Open Images**: Multi-label, automatically generated
- **Existing taxonomies**: Too domain-specific or hierarchical

Solution: Generate synthetic datasets with controlled relationships

1. Define **atomic concepts** $\mathcal{U} = \{1, 2, \ldots, n\}$
2. Create synthetic classes as subsets: $c_j^i \subseteq \mathcal{U}$
3. Generate multiple domains by sampling concepts
4. Calculate relationships based on concept overlap

$$P_{i,j} = \frac{|c_i^A \cap c_j^B|}{|c_i^A|} + \frac{1 - \frac{|c_i^A \cap c_j^B|}{|c_i^A|}}{|C_B|} \tag{5}$$

Domain-Shifted Synthetic Datasets

Problem: Original synthetic variants too easy (same underlying images)

Solution: Apply domain transformations to create realistic challenges

Transformations Applied:

- **Domain A**: Noisy grayscale
- **Domain B**: Rotation + blur
- **Domain C**: Random erasing + color jitter + perspective shifts

Example images from domain-shifted variants

Results:

- Model accuracy drops by 10%
- More realistic cross-domain predictions
- Better evaluation of relationship selection methods

How do we measure the quality of selected relationships?

1. **Edge Difference Ratio (EDR)**: Measures difference in edge weights between predicted and ground truth graphs

$$\mathrm{EDR}(G_1, G_2) = \frac{\sum_{i,j} |A_1(i,j) - A_2(i,j)|}{\sum_{i,j} \max(A_1(i,j), A_2(i,j))} \tag{6}$$

   Range: [0,1], where $0 =$ identical graphs, $1 =$ no common edges

2. **Precision & Recall**: Binary evaluation of relationship presence
   - Convert adjacency matrices to binary: $B(i,j) = 1$ if $A(i,j) > 0$
   - **Precision** $= \frac{TP}{TP+FP}$ (correctness of selected relationships)
   - **Recall** $= \frac{TP}{TP+FN}$ (coverage of true relationships)
   - **F1 Score** = Harmonic mean of precision and recall

Relationship Selection Results: Domain-Shifted Evaluation

Evaluation on domain-shifted synthetic datasets (more realistic)

Method	Parameter	EDR	Precision	Recall	F1 Score
Relationship Hypothesis	5	**0.759**	0.390	0.543	0.444
Naive Thresholding	0.10	0.761	0.418	0.519	**0.450**
Density Thresholding	0.60	0.766	0.349	**0.582**	0.426
MCFP	N/A	0.842	**0.490**	0.226	0.305

Table: Performance on domain-shifted synthetic datasets with globally optimal parameters

Key Observations:

- **Relationship Hypothesis** achieves best EDR 0.759
- **MCFP** maintains highest precision but lowest recall
- Maybe still not realistic enough, but suitable for getting method parameters

Universal Taxonomy Building Rules

How do we convert relationship graphs into universal taxonomies?

1. **Isolated Node Rule**: Classes with no relationships
   - Create new universal class for standalone domain classes
   - Ensures all classes are represented in universal taxonomy
2. **Bidirectional Relationship Rule**: Classes with mutual relationships ($A \leftrightarrow B$)
   - Create single universal class C with relationships $A \rightarrow C$, $B \rightarrow C$
   - Indicates classes likely represent the same concept
3. **Transitive Cycle Rule**: Prevent invalid cycles ($A \rightarrow B \rightarrow C$ where A, C same domain)
   - Remove relationship with lower probability
   - Classes within same domain should be disjoint
4. **Unilateral Relationship Rule**: Handle subset relationships ($A \rightarrow B$)
   - Create universal class for shared concepts ($A \cup B$)
   - Create universal class for unique concepts (B only)

Universal Model Architecture

Key Differences from Baseline Models:

Baseline Model:

- 6-layer FC classifier
- Dropout regularization
- One-hot targets
- Cross-entropy loss
- Domain-specific outputs

Universal Model:

- 2-layer FC classifier
- No dropout
- Probability distribution targets
- Cross-entropy for discrete distributions
- Universal class outputs

Target Generation: Convert domain labels to universal class distributions

$$\mathbf{t} = \hat{M}_i[j,:] \quad \text{where } \hat{M}_i(j,u) = \frac{M_i(j,u)}{\sum_{u'} M_i(j,u')} \tag{7}$$

Loss Function:

$$\mathcal{L} = -\sum_{u=1}^{|U|} \mathbf{t}(u) \log(\mathbf{p}(u)) \tag{8}$$

Multi-Domain Training Process

Training Procedure:

1. Combine multiple datasets while preserving domain identity
2. Each sample: $(\text{image}, (\text{domain\_id}, \text{label})) \rightarrow (\text{image}, \text{universal\_target})$
3. Train single model on unified dataset
4. Use domain-specific mapping matrices for target generation

Inference:

$$\mathbf{d}_i = M_i^T \mathbf{p} \tag{9}$$

$$\hat{c}_i = \text{argmax}(\mathbf{d}_i) \tag{10}$$

where $\mathbf{p}$ are universal class predictions and $\mathbf{d}_i$ are domain-specific predictions.

Key Challenge: Validation loss increases while accuracy improves

- Solution: Monitor validation accuracy for checkpointing
- Caused by label smoothing effects and multi-target distributions

Universal Model Performance

Datasets: Caltech-101, Caltech-256, CIFAR-100

Model Type	**Caltech-101**	**Caltech-256**	**CIFAR-100**
Baseline (Single Domain)	0.828	0.798	0.734
Universal (MCFP)	**0.845**	**0.812**	0.728
Universal (Density)	0.834	0.809	**0.741**
Universal (Hypothesis)	0.831	0.801	0.735

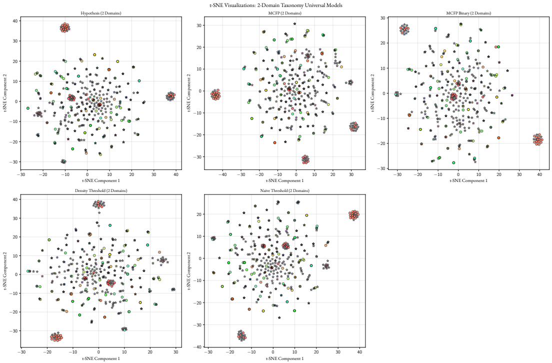
Table: Test accuracy comparison (2-domain: Caltech-101 + Caltech-256)

Key Findings:

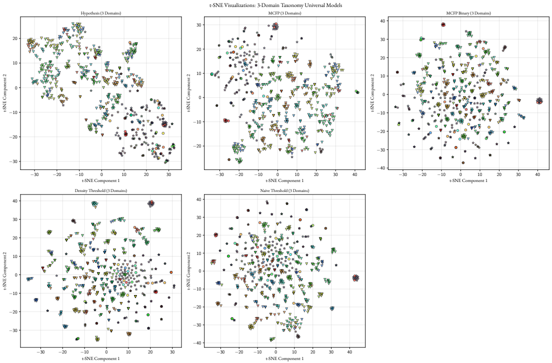
- Universal models **outperform** single-domain baselines
- Different relationship selection methods excel on different datasets
- No single method consistently optimal across all scenarios

Feature Visualization with t-SNE

2-Domain (Caltech-101 + 256)



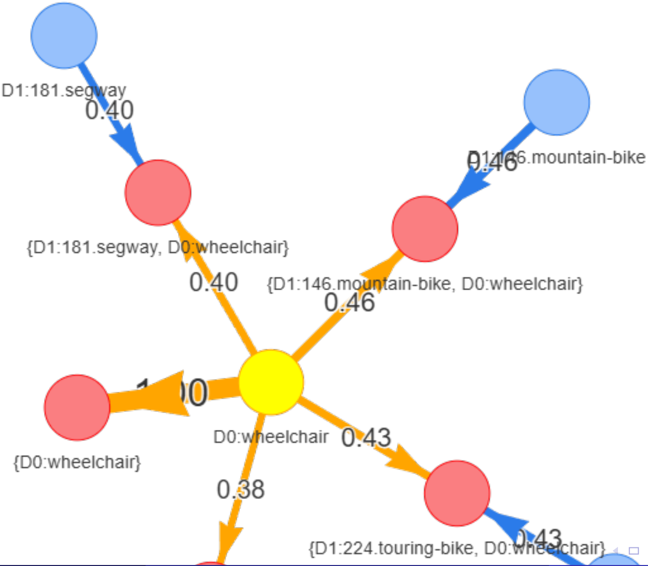
3-Domain (+ CIFAR-100)



Observations:

- Mixed clusters indicate successful cross-domain feature learning
- Some domain-specific clusters for unique classes

Taxonomy Visualization Example



Challenges and Limitations

Key Contributions

1. **Novel Universal Model Approach**: Single model for multiple domains without task-specific heads
2. **Comprehensive Relationship Selection Methods**: Four different approaches with systematic evaluation
3. **Synthetic Ground Truth Framework**: Controlled evaluation environment with domain-shifted variants
4. **Cross-Domain Prediction Pipeline**: Complete methodology from raw datasets to universal models
5. **Performance Validation**: Universal models outperform single-domain baselines

Main Findings

Successful Aspects:

- Universal models achieve better accuracy than single-domain models
- Cross-domain prediction effectively captures semantic relationships
- Synthetic datasets provide valuable controlled evaluation environment
- Feature visualizations show meaningful cross-domain clustering

Challenges Identified:

- No single relationship selection method works optimally for all cases
- Gap between evaluation metrics and actual model performance
- Parameter selection requires careful tuning for each dataset combination
- Some false positive relationships in generated taxonomies

Future Work & Improvements

Immediate Improvements:

- Grid search for optimal relationship selection parameters
- Develop better correlation between evaluation metrics and model performance
- Explore different universal model architectures

Extensions:

- Scale to larger datasets (ImageNet, COCO)
- Apply to other vision tasks (object detection, segmentation)
- Investigate dynamic relationship adjustment during training
- Explore unsupervised relationship discovery methods

Real-World Applications:

- Multi-domain medical imaging
- Cross-dataset autonomous driving
- Unified content moderation systems

Summary

Problem: Traditional image classification limited to single domains
Solution: Universal model trained on cross-domain taxonomy
Method: Cross-domain prediction $\rightarrow$ Relationship selection $\rightarrow$ Universal training
Results: Universal models outperform single-domain baselines

Thank you for your attention!

Questions?