Log

Logo Classifier

4.14

Presented our final project proposal on logo classification. We are planning to use datasets from Kaggle (Dataset 1, Dataset 2) and integrate in WikiData API calls to cross validate and label images by the respective GCIS identification. However, we found that this approach may not be the best way, especially because WikiData leads to ambiguous results that have no standardization.

We are pivoting to using a different dataset

4.24

At this day we PLANNED

4.27

We got the Data!

- dataset_code

Useful paper: https://github.com/jcalz23/logo_detection_w281

- Inspired by the preprocessing approach used in the logo detection GitHub project, which applied CLAHE, random 3D rotations, color inversions, and class balancing, we adapted a milder augmentation pipeline for EfficientNetB0 sector classification. Since our task uses cropped logo images and sector-level labels, we applied small rotations, zooms, translations, and contrast adjustments while avoiding color inversion and horizontal flipping to preserve logo identity and text readability.

Used this to help me understand how to prepare and load data for efficientnetB0

- Parameters
- Casting and normalization

NOTE

tf.cast(img, tf.float32) / 255.0

- We cannot do this because EfficientNet actually has a Rescaling layer that does this already so this will cause the image data to be zeroed out

First epoch run on Kaggle! WOOHOO

```
H = model.fit(train_data, validation_data=val_data, epochs=20, callbacks=callbacks)

Epoch 1/20
2650/5127 ━━━━━━━━━━━━━━━━━━━━ 2:48 68ms/step - accuracy: 0.1116 - loss: 3.0168 - top_3_accuracy: 0.2653
```

+ Code + Markdown

Logo Classifier

4.28

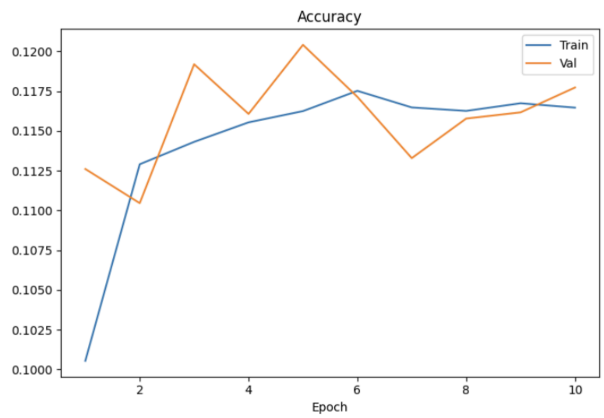
5125/5127 ━━━━━━━━━━━━━━━━━━━━ **0s** 33ms/step - accuracy: 0.1219 - loss: 3.0061 - top_3_accuracy: 0.2797

Best accuracy recorded when running the model for the first time. I think we will go ahead and try to do some data augmentation – influenced by the github paper.

- Randomly rotate, translate, zoom, and add contrast to help model learn more effectively

[Official Keras paper]

- Recommend keeping BatchNormalization layers frozen during fine-tuning and unfreezing only the **top** part of the model.

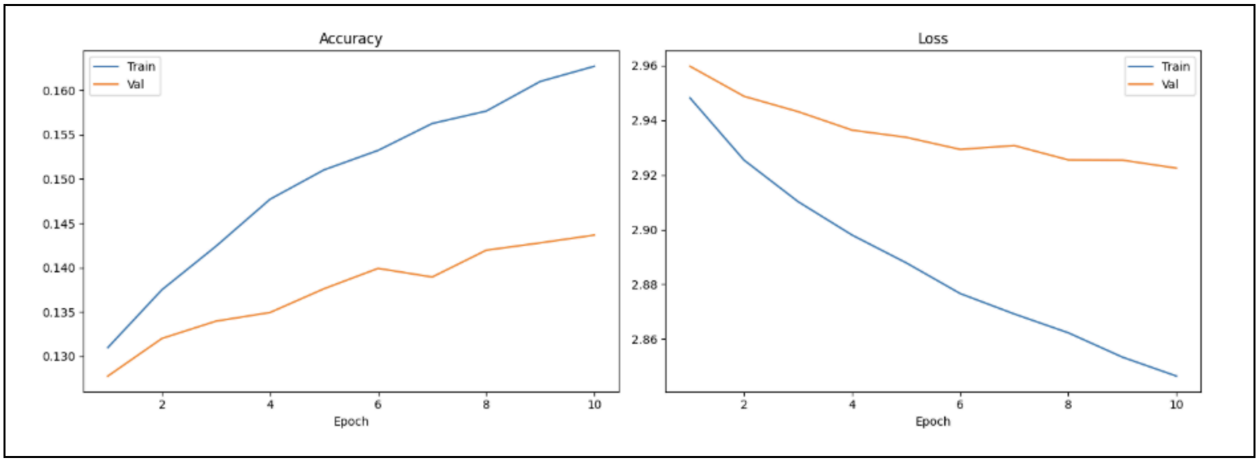


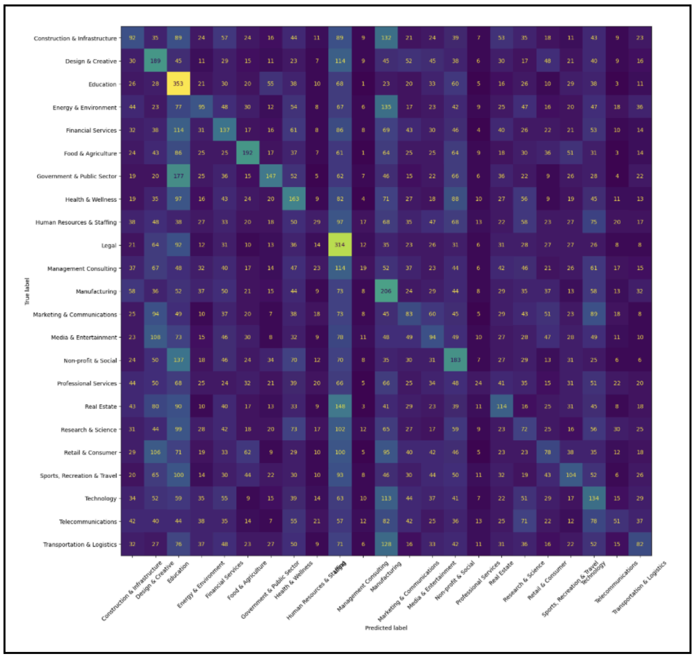
4.29

We have the weights of the baseline model (output layer changed to 23 classifications) as our initial benchmark. From here, we are going to try many different approaches to improve accuracy. Because we are facing a very difficult task of logo categorization and classification, we are going to focus on the top 3 sparse accuracy.

20 layers frozen (batch normalization layers kept frozen, epochs = 10, lr = 1e-5)

Logo Classifier





`accuracy: 0.1594 - loss: 2.8555 - top_3_SparseTopKCategoricalAccuracy: 0.3414`

So obviously looking at just unfreezing the top 20 layers and changing learning rate to 1e-5, we have some promising improvement in validation accuracy. However, we want to do better. I think that this may come down to one or more of three reasons:

1. Dataset issues such as labeling or dataset quality
2. Impossibility of task
3. Many frozen layers of the base model are not doing anything to help learn these images

PIVOT

We are going to go ahead and try to freeze more layers and adjust how many epochs + the learning rate to account for possible forgetting.

Logo Classifier

- `ReduceLROnPlateau(patience=3)` will be used to ensure that the learning rate halves if no learning is observed during training
- epochs = 30, lr = 1e-6

Furthermore, I found a potential optimization for compute time. Because we are using 2 T4 GPUs, we can make use of the fact that the tensor core hardware is designed to multiply float16 matrices more efficiently and faster than float32. Implement this before any model loading code; however, if I ever want to retrain the original baseline model, I added casting of the output layer back to float32.

tf.keras.mixed_precision.set_global_policy('mixed_float16')

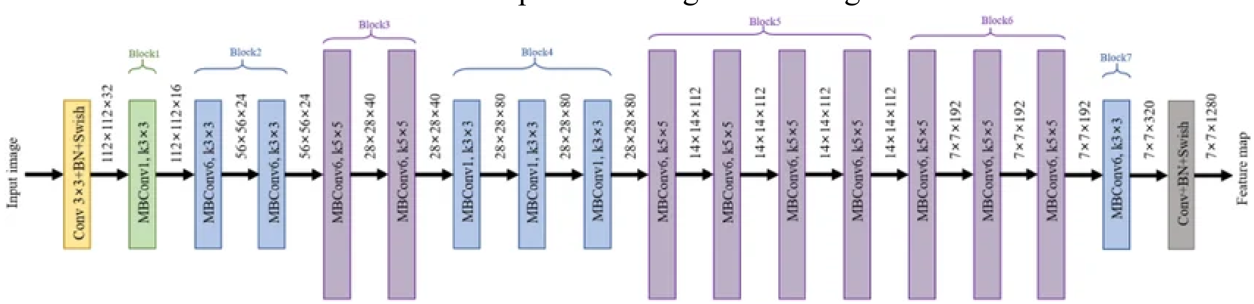
We can also make use of the fact that we have two T4 GPUs by using mirrored strategy to create replicas of training instances on both GPUs. This should theoretically cut down training time. I prompted Claude to help me figure out how to make use of these optimizations to reduce training time since we have time constraints in our Kaggle notebooks.

Claude Output:

- When you double the batch size, each gradient update is computed from twice as many samples, making the gradient estimate smoother and larger in magnitude. To keep the same effective learning step, you scale LR by the same factor:

[EfficientNetB0 Paper]

- ***IMPORTANT**
- I looked more into the architecture of the model and it looks like it has <u>blocks</u> that group layers.
    - We can use this for more precise tuning / unfreezing

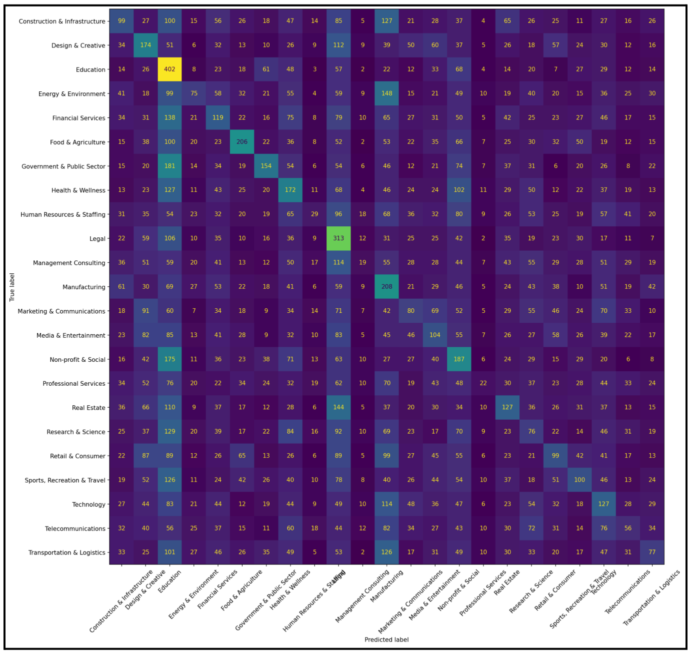


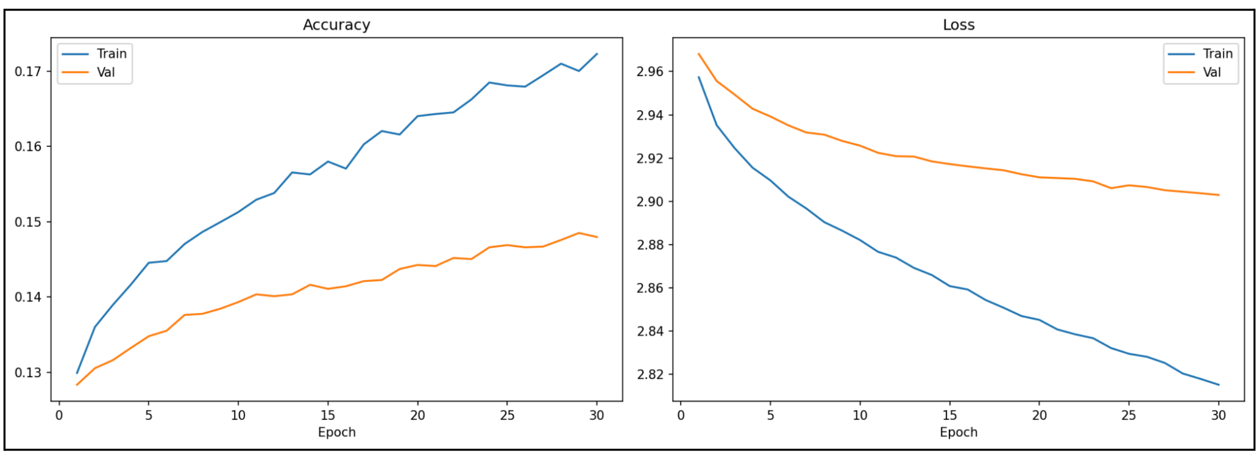
Because of this, we will now be changing how we unfreeze our layers. Now we will refer to the blocks of the model rather than individual layers. We will **<u>still</u>** ensure that we do not unfreeze batch normalization layers to preserve accuracy per Keras.

Logo Classifier

4.30

After unfreezing Block 5 → top, I think the biggest culprit of our model's bad performance is the number of classifications.





So I think we need to do a couple of things:

1. Re-evaluate dataset by either lowering the number of classifications, doing more data preprocessing (whitening background, etc), re-assess why for arman's code it is overfitting to one category (my runs seemed to overfit on education I think? idk)
2. Expand from top 3 to top 4 categories accuracy and figure out a way to report metrics on how much of each category the model predicted an image to belong to
3. Continue unfreezing layers and hope ts improves

Logo Classifier

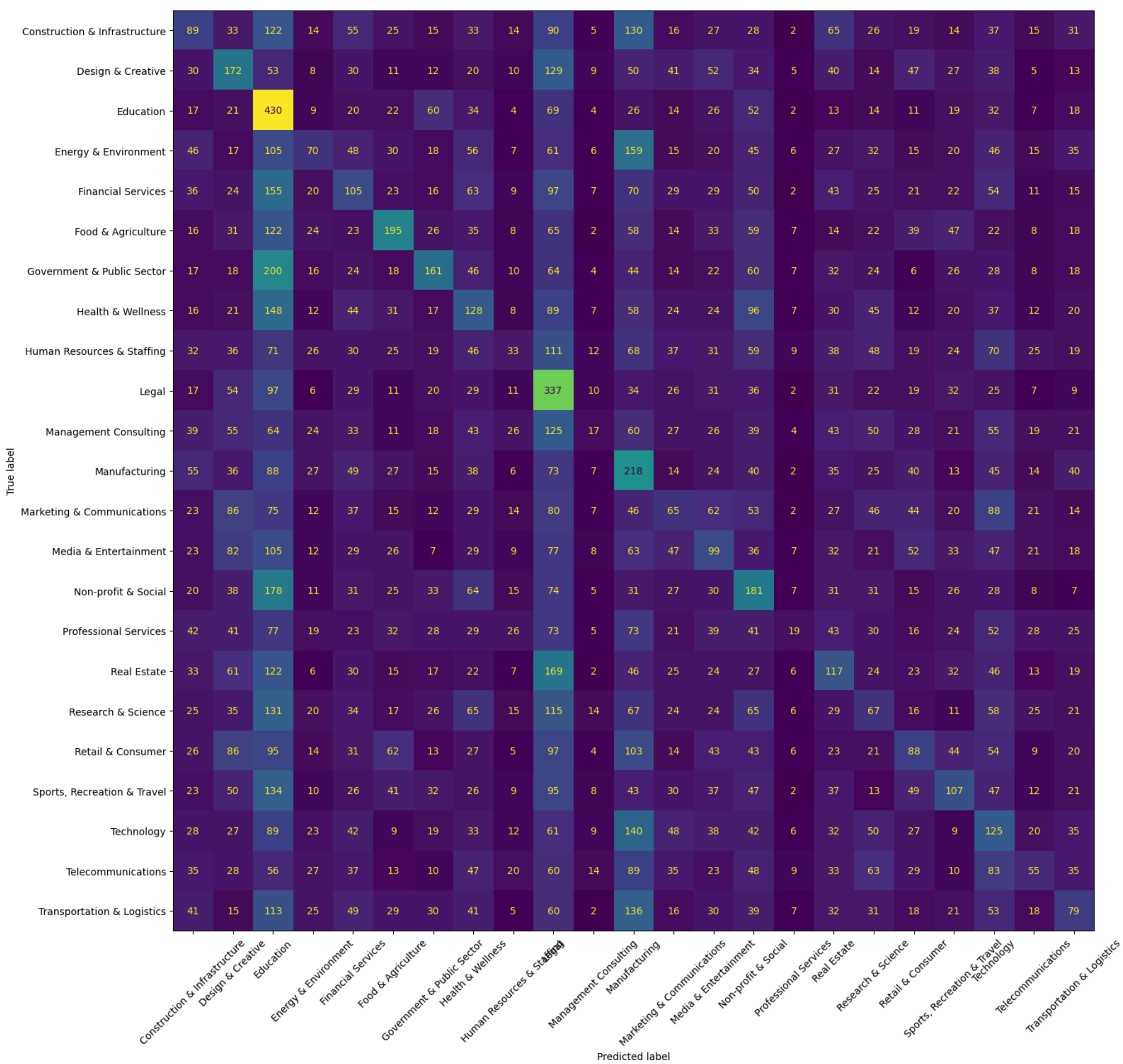
5.1

Today I tested a few things. One, I tested unfreezing more blocks at once (i.,e block3 to top), but it didn't make a significant difference. Furthermore, I tried testing changing the background color of all of the logos to white, however that ended up performing worse than the none changed background color icons. Finally, I am testing to see if we need less categories (14 instead of 23).

Block 3 to Top Frozen w/ standard background:

```
2564/2564 ━━━━━━━━━━━━━━━━━━━━ 791s 308ms/step - accuracy: 0.1548 - loss: 2.8690 - top_4_SparseTopKCategoricalAccuracy: 0.4050 - val_accuracy: 0.1430 - val_loss: 2.9142 - val_top_4_SparseTopKCategoricalAccuracy: 0.3843 - learning_rate: 1.0000e-06
Restoring model weights from the end of the best epoch: 30.
```

Logo Classifier

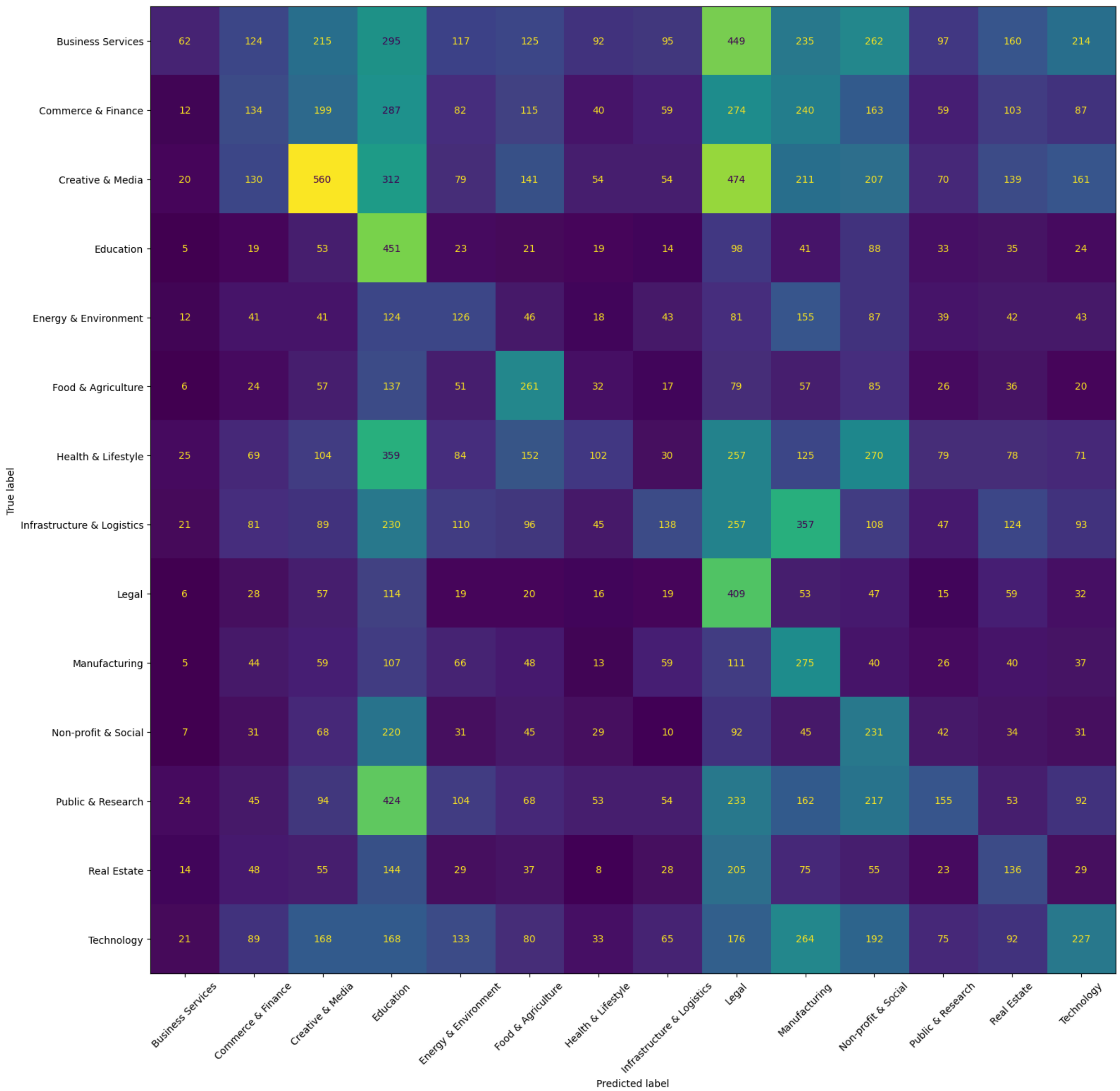


5.4

Today, we focused on minimizing the number of classifications (choosing 14 instead of the 23) to see how it affects the accuracy of the model. Now, we are retraining the final output layer, and then seeing how unfreezing certain other layers affect the accuracy of the model. Furthermore, I am testing to see if changing the background color affects the accuracy of the model, although I may have to start over and retrain the whole model for this perhaps.

This is what 14 classifications with background image color unfreeznig blocks 3 to top looked like:

Logo Classifier



I’m going to try unfreezing the first few blocks and see if anything changes.

Education	12725
Health & Wellness	12190
Technology	11718
Manufacturing	9303
Retail & Consumer	9274
Financial Services	9260
Non-profit & Social	9159
Construction & Infrastructure	9050
Sports, Recreation & Travel	8988

Logo Classifier

Energy & Environment	8985
Transportation & Logistics	8905
Food & Agriculture	8882
Human Resources & Staffing	8881
Real Estate	8859
Media & Entertainment	8837
Marketing & Communications	8785
Government & Public Sector	8675
Telecommunications	8593
Design & Creative	8500
Management Consulting	8479
Professional Services	8062

Notes:

Clean dataset (take out blurry images, see if u can determine country/language to model)

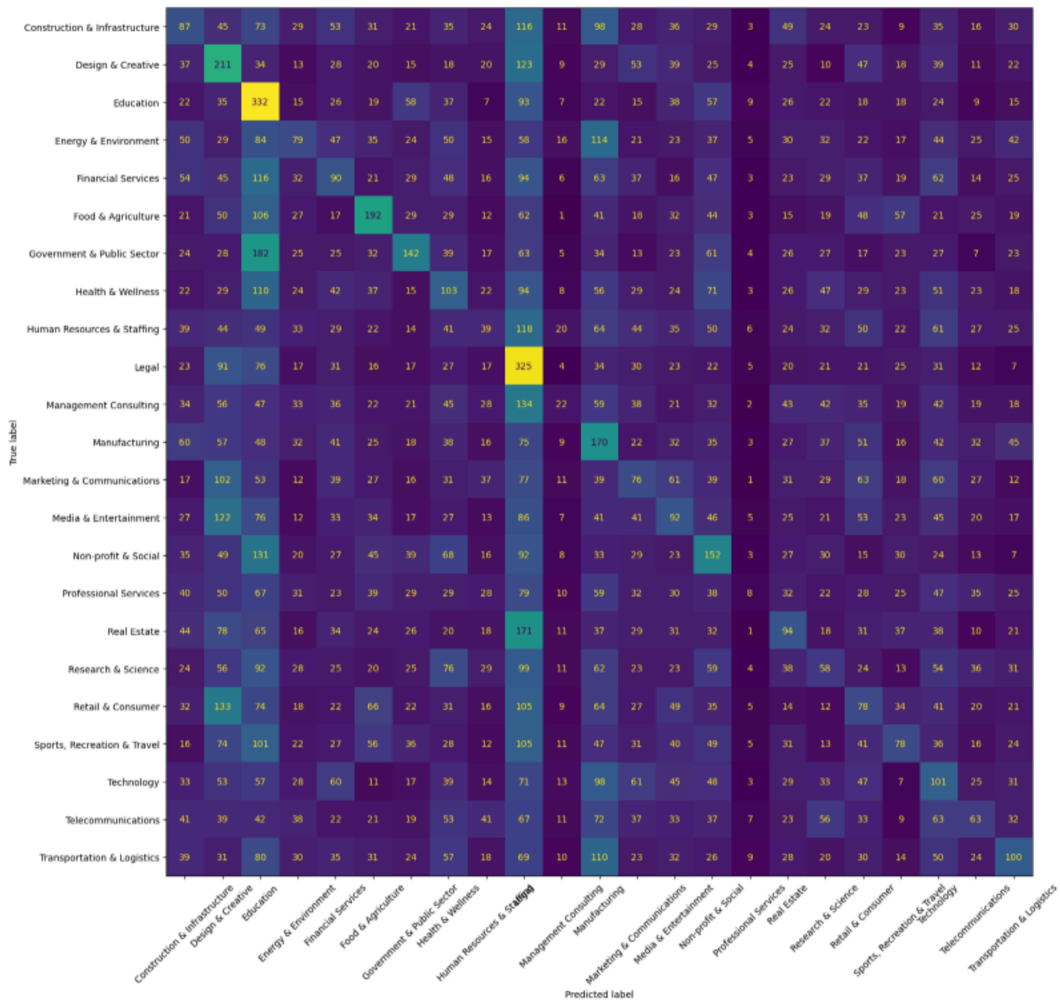
Trying a different baseline (regular cnn)
Justin: Testing relabeling (taking out legal, merging some categories)
Tevin: optimizing original accuracy
Arman: testing metadata to help with logo classification
Shadab: testing taking out blurry images.

Hard to take out blurry images specifically, code was taking up more “minimalist” images or images without as sharp edges so it didn’t make a significant difference.

5.5

- Batch size = 64, epochs = 5, higher training learning rate (1e-4)
- accuracy: 0.1164 - loss: 2.9889 - top3: 0.2751 - top5: 0.3969 - val_accuracy: 0.1150 - val_loss: 2.9876 - val_top3: 0.2781 - val_top5: 0.3980
- accuracy: 0.1322 - loss: 2.9498 - top3: 0.2980 - top5: 0.4192

Logo Classifier



Best performance on 21 classifications:

```
187ms/step - accuracy: 0.1945 - loss: 2.6844 -
top_5_SparseTopKCategoricalAccuracy: 0.5363 - val_accuracy: 0.1655 -
val_loss: 2.8092 - val_top_5_SparseTopKCategoricalAccuracy: 0.4834 -
learning_rate: 1.0000e-05
```

Maybe let us try out label smoothing?

- https://arxiv.org/abs/1512.00567

Logo Classifier

dataset code

```
# dataset labeling code
INDUSTRY_RENAMES: dict[str, str] = {
    "nonprofit organization management": "non-profit organization management",
    "airlines/aviation": "aviation & aerospace",
    "law practice": "legal services",
}

INDUSTRY_TO_SECTOR: dict[str, str] = {
    # Technology
    "information technology and services": "Technology",
    "computer software": "Technology",
    "internet": "Technology",
    "computer hardware": "Technology",
    "computer networking": "Technology",
    "computer & network security": "Technology",
    "wireless": "Technology",
    "information services": "Technology",
    "semiconductors": "Technology",
    "computer games": "Technology",
    # Marketing & Communications
    "marketing and advertising": "Marketing & Communications",
    "public relations and communications": "Marketing & Communications",
    "market research": "Marketing & Communications",
    # Construction & Infrastructure
    "construction": "Construction & Infrastructure",
    "civil engineering": "Construction & Infrastructure",
    "building materials": "Construction & Infrastructure",
    "facilities services": "Construction & Infrastructure",
    # Management Consulting
    "management consulting": "Management Consulting",
    # Real Estate
    "real estate": "Real Estate",
    "commercial real estate": "Real Estate",
    # Financial Services  (accounting folded in)
    "financial services": "Financial Services",
    "banking": "Financial Services",
    "investment management": "Financial Services",
    "investment banking": "Financial Services",
    "capital markets": "Financial Services",
    "venture capital & private equity": "Financial Services",
    "insurance": "Financial Services",
    "accounting": "Financial Services",
    # Health & Wellness
    "health, wellness and fitness": "Health & Wellness",
```

```
"hospital & health care": "Health & Wellness",
"medical practice": "Health & Wellness",
"mental health care": "Health & Wellness",
"alternative medicine": "Health & Wellness",
"veterinary": "Health & Wellness",
"pharmaceuticals": "Health & Wellness",
# Design & Creative
"design": "Design & Creative",
"graphic design": "Design & Creative",
"photography": "Design & Creative",
"fine art": "Design & Creative",
"arts and crafts": "Design & Creative",
"architecture & planning": "Design & Creative",
"animation": "Design & Creative",
# Education
"education management": "Education",
"higher education": "Education",
"primary/secondary education": "Education",
"e-learning": "Education",
"professional training & coaching": "Education",
"libraries": "Education",
"program development": "Education",
# Non-profit & Social
"non-profit organization management": "Non-profit & Social",
"civic & social organization": "Non-profit & Social",
"philanthropy": "Non-profit & Social",
"religious institutions": "Non-profit & Social",
"fund-raising": "Non-profit & Social",
"individual & family services": "Non-profit & Social",
"museums and institutions": "Non-profit & Social",
# Human Resources & Staffing
"human resources": "Human Resources & Staffing",
"staffing and recruiting": "Human Resources & Staffing",
"outsourcing/offshoring": "Human Resources & Staffing",
# Government & Public Sector
"government administration": "Government & Public Sector",
"government relations": "Government & Public Sector",
"public policy": "Government & Public Sector",
"political organization": "Government & Public Sector",
"international affairs": "Government & Public Sector",
"legislative office": "Government & Public Sector",
"military": "Government & Public Sector",
"public safety": "Government & Public Sector",
"law enforcement": "Government & Public Sector",
```

```
"judiciary": "Government & Public Sector",
"defense & space": "Government & Public Sector",
"think tanks": "Government & Public Sector",
"international trade and development": "Government & Public Sector",
# Legal
"legal services": "Legal",
"alternative dispute resolution": "Legal",
# Energy & Environment
"oil & energy": "Energy & Environment",
"renewables & environment": "Energy & Environment",
"environmental services": "Energy & Environment",
"utilities": "Energy & Environment",
"mining & metals": "Energy & Environment",
# Food & Agriculture
"food & beverages": "Food & Agriculture",
"food production": "Food & Agriculture",
"restaurants": "Food & Agriculture",
"farming": "Food & Agriculture",
"dairy": "Food & Agriculture",
"fishery": "Food & Agriculture",
"ranching": "Food & Agriculture",
"supermarkets": "Food & Agriculture",
"wine and spirits": "Food & Agriculture",
# Transportation & Logistics
"transportation/trucking/railroad": "Transportation & Logistics",
"logistics and supply chain": "Transportation & Logistics",
"maritime": "Transportation & Logistics",
"warehousing": "Transportation & Logistics",
"package/freight delivery": "Transportation & Logistics",
"aviation & aerospace": "Transportation & Logistics",
"import and export": "Transportation & Logistics",
# Manufacturing
"mechanical or industrial engineering": "Manufacturing",
"electrical/electronic manufacturing": "Manufacturing",
"machinery": "Manufacturing",
"industrial automation": "Manufacturing",
"chemicals": "Manufacturing",
"plastics": "Manufacturing",
"packaging and containers": "Manufacturing",
"paper & forest products": "Manufacturing",
"glass, ceramics & concrete": "Manufacturing",
"textiles": "Manufacturing",
"printing": "Manufacturing",
"shipbuilding": "Manufacturing",
```

```
"railroad manufacture": "Manufacturing",
"automotive": "Manufacturing",
# Retail & Consumer
"retail": "Retail & Consumer",
"apparel & fashion": "Retail & Consumer",
"cosmetics": "Retail & Consumer",
"luxury goods & jewelry": "Retail & Consumer",
"furniture": "Retail & Consumer",
"consumer goods": "Retail & Consumer",
"consumer electronics": "Retail & Consumer",
"sporting goods": "Retail & Consumer",
"wholesale": "Retail & Consumer",
"business supplies and equipment": "Retail & Consumer",
"tobacco": "Retail & Consumer",
# Media & Entertainment
"media production": "Media & Entertainment",
"entertainment": "Media & Entertainment",
"broadcast media": "Media & Entertainment",
"music": "Media & Entertainment",
"publishing": "Media & Entertainment",
"online media": "Media & Entertainment",
"newspapers": "Media & Entertainment",
"motion pictures and film": "Media & Entertainment",
"performing arts": "Media & Entertainment",
"writing and editing": "Media & Entertainment",
# Sports, Recreation & Travel
"sports": "Sports, Recreation & Travel",
"recreational facilities and services": "Sports, Recreation & Travel",
"leisure, travel & tourism": "Sports, Recreation & Travel",
"hospitality": "Sports, Recreation & Travel",
"gambling & casinos": "Sports, Recreation & Travel",
"events services": "Sports, Recreation & Travel",
# Telecommunications
"telecommunications": "Telecommunications",
# Research & Science
"research": "Research & Science",
"nanotechnology": "Research & Science",
"biotechnology": "Research & Science",
"medical devices": "Research & Science",
# Professional Services
"consumer services": "Professional Services",
"security and investigations": "Professional Services",
"translation and localization": "Professional Services",
"executive office": "Professional Services",
```

```
    "non-profit organization management": "Non-profit & Social",  # keep alias safe
}

def normalize_data(df, industry_col):
    df = df.copy()
    df[industry_col] = df[industry_col].replace(INDUSTRY_RENAMES)
    df["sector"] = df[industry_col].map(INDUSTRY_TO_SECTOR).fillna("NaN")

    return df


raw_ds =
"/kaggle/input/datasets/peopledatalabssf/free-7-million-company-dataset/companies_sorted.csv"
logo_ds = "/kaggle/input/datasets/shadabsharif/logoss/academic"

images = glob.glob("/kaggle/input/datasets/shadabsharif/logoss/academic/*.png")
logos = pd.DataFrame({
    "domain": [os.path.splitext(os.path.basename(f))[0] for f in images],
    "filepath": images,
})

raw = pd.read_csv(
    raw_ds,
    usecols=["name", "domain", "industry", "size range"],
    low_memory=False,
)
raw = raw[raw["domain"].notna() & (raw["size range"] != "1 - 10") & raw["industry"].notna()]
raw = normalize_data(raw, "industry")

labeled = logos.merge(raw[["name", "domain", "sector"]], on="domain", how="inner")
labeled["sector"].value_counts()

labeled = labeled[“name”, “domain”, “sector”, “filepath”]
labeled.to_csv("/kaggle/working/labels.csv", index=False)
```

Article

Introduction and Problem Statement - Tevin

- Clearly state the purpose and objectives of the project.
- Provide context and background information on the topic.
- Explain the problem or research question the project aims to address.

Methodology & Data - Tevin

- Touch upon theory if it is necessary to understand your work.
- Describe the project design, analysis techniques, pre-processing and the Neural Network(s) you expect to use.
- List any data, resources, etc. needed to carry out the project.

Results - Shadab

- Outline the results and findings of the project.
- Prepare plots and figures to show the results - ensure they are clear, concise, and relevant.

Heatmap paper

Visualization on Keras tutorial

Discussion - Arman

- Discuss the context of your results, how the results relate to each other and how to interpret them.
- Revisit challenges you encountered and how you mitigated or solved them.
- If applicable, put your project in relation to other work in the same field.

Conclusion - Arman

- Summarize the main results and implications of your project.

Figures, Diagrams, etc. - Arman

- Use them to report on results and explain complex parts of your work.
- A well-made figure with an explanatory caption can save a lot of text.

Introduction

One of the best ways to change the trajectory of a company is through a rebrand and a new logo. Around 79% of S&P 100 companies have rebranded within their first seven years, showing how important it is for companies trying to grow. However, changing a logo comes with major risks for the company.

Gap Fails to Rebrand

In 2010, Gap tried to rebrand itself after its sales started to decline. They got rid of their well-established and recognizable logo and introduced a more modern logo. However, as soon as it was announced, it was heavily criticized for being too plain and unrepresentative of Gap's clothing brand, forcing Gap to change it back within 6 days. It proved to be a costly lesson for Gap as the redesign, marketing, and beginning stages of physical store changes cost them around $100 million.

What We Are Trying to Solve (Problem Statement)

We want to make it easier for companies to see if their new logo still aligns with the industry they represent.

To do this, we tried to develop a predictive model that can classify what industry a logo belongs to based on visual features.

Data

We had a hard time searching for data. There were many incomplete datasets: some had company logos but no industry labels, while others had industry information but no company names or logos. There were also many datasets that were badly labeled.

After looking through several datasets, we decided to use this Kaggle dataset (with hyperlink) that provides information on over 7 million companies. We then contacted logo.dev for logo images (in 224x224 size) for hundreds of thousands of companies. They kindly agreed to provide all of it for research purposes in exchange for attribution.

Preprocessing

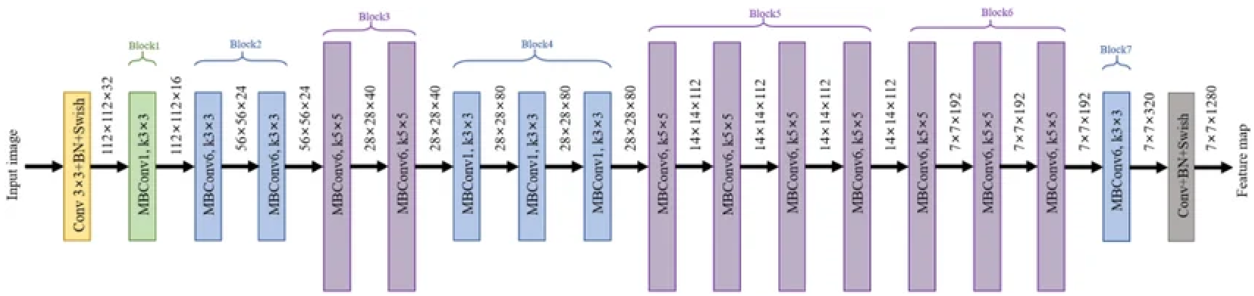
From the 7 million rows in the original dataset, we first removed any entries that were missing an industry, domain, or company name. We then grouped the 145 original industries into 23 broader sectors, such as technology, marketing, and communications, to make the categories easier for the model to learn from.

After cleaning and grouping the data, we were left with around 1.4 million usable entries. From there, we selected 230,000 companies by taking the top 10,000 from each sector based on company size, since those were the logos that logo.dev was most likely to have available.

Logo.dev was able to provide about 210,000 logos, which became our final dataset for training and testing the model.

We then split the dataset into 3 sets: 80% training (for training), 10% validation (for monitoring performance during training), and 10% testing (for evaluating model afterwards).

Architecture - Why CNN and EfficientNetB0?



Since logos are visual, we needed a model architecture that could learn patterns directly from images. A Convolutional Neural Network, or CNN, was a good fit because it is designed to recognize visual features such as shapes, colors, edges, typography structures, etc.

We chose EfficientNetB0 because it allowed us to start with a model that already understands general visual patterns from image data. This gave our model a stronger starting point instead of having to learn everything from scratch.

We specifically used the B0 version because it offers a good balance between performance and size. It is powerful enough for image classification, but not so large that it becomes too slow or difficult to train. We tested it both as a feature extractor and with fine-tuning to see which approach gave us better results.

Furthermore, EfficientNetB0 divides the layers into blocks. This lets us more easily control how much of the model we are modifying.

Iteration and testing

After building the first version of the model, we tested several different ways to improve its performance. One of the main changes we tried was unfreezing more layers of EfficientNetB0 so the model could adjust more of its visual understanding to our specific logo dataset. We first unfroze the later layers, then tested unfreezing blocks 5 and 4, which led to a small improvement.

We first unfroze only block 6, then blocks 5 and 6, then blocks 4, 5, and 6. Our results got better but unfreezing more than that led to no more improvements.

We also experimented with changes to the dataset and training process. This included removing certain categories, reorganizing some sectors in a second pass, adding company metadata (language and date founded), training for more epochs, changing the learning rate, increasing the batch size, and applying data augmentation. Additionally, we opted for label smoothing regularization because we found that our model over-confidently grouped logos into certain classifications in a disproportionate scale—for example, 99.99% in education and 0.001 in design and creative. Using this technique required that we cast our labels to be one-hot encoded, promoting the model to stay a little uncertain about its predictions.

Most of these changes only led to minor improvements, while adding metadata caused the model to overfit.

To compare against a simpler approach, we also tested a basic four-layer CNN. However, this model underfit the data, meaning it was not complex enough to learn the visual differences between industries. Overall, EfficientNetB0 gave us the best balance between accuracy, flexibility, and training performance.

Analysis Technique

We focused on the training and validation accuracy. We aimed for the validation accuracy match or even surpass the training accuracy. We also aimed to have both the validation loss and training loss to be similar and as small as possible.

Furthermore, many industries are similar in nature, which means their logos often share similar branding styles and visual cues. For example, technology and telecommunications companies may use similar colors or shapes, making them harder for the model to separate into one exact category. Because of this, we decided to focus more on top-3 and top-5 accuracy rather than the top-1 accuracy. This helped us better understand whether the correct industry was still included among the model's strongest predictions, even if it was not always ranked first.

Results ish

- Some of them helped a little, but nothing meaningful. Increased top 5 accuracy from 39% to 49% w/ 21 classifications + reorganization, unfreeze block 3 -> top, label smoothing, early stopping, dropout, some minor augmentation, 40 epochs, 1e-5 with learning rate plateau

- Need Confusion Matrix, Training Loss VS Validation Loss Graph, per-class F1 bar chart, and a sample of the images, their predicted labels, and their actual labels. For the results, I can just explain each of the graphs.

Our best model achieved a top-5 accuracy of 49.3% and a top-1 accuracy of 17.4% on a test set across 21 industry sectors. The model was trained in two phases: first fine-tuning only the output layer on top of a frozen EfficientNetB0 backbone, then progressively unfreezing blocks 3 through the top of the network, which was approximately 90-100 layers, for further fine-tuning. The following figures illustrate model performance across training, per-industry accuracy, and prediction examples:

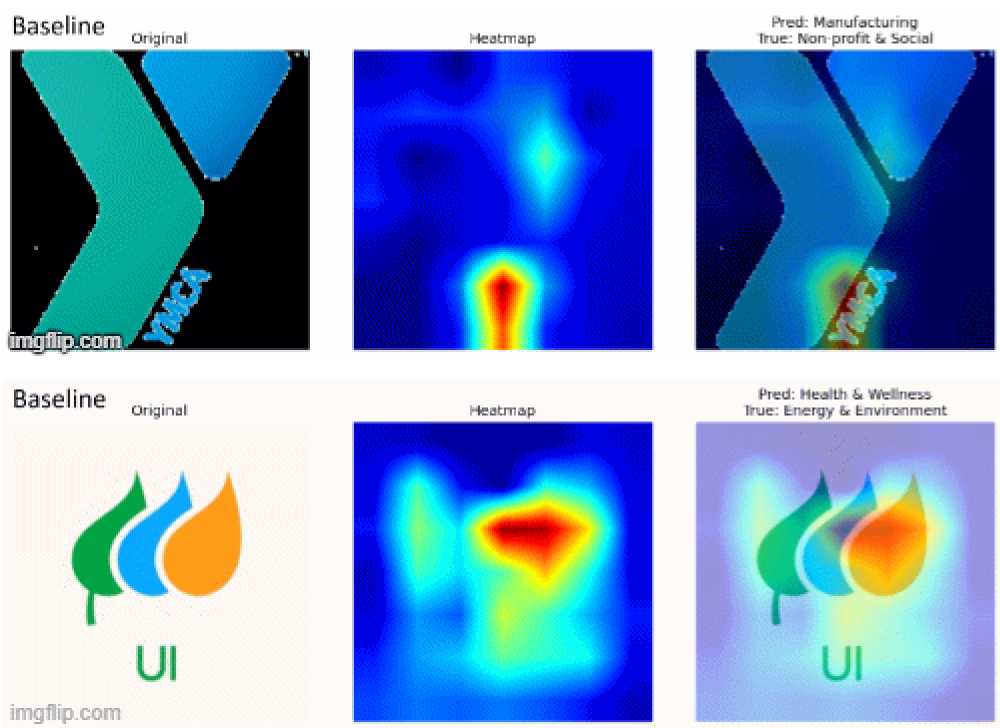


Figure X.y: These are two random samples from our dataset whose visualizations reveal which regions of each logo most influenced the model's predictions. The gifs starts at how the model recognized the logo based off of just unfreezing the output layer, and then each frame shows how the model reacts after we unfroze blocks 3-7 as unfroze one more block at each frame starting from block 7 (ex. Baseline, block 7, then blocks 6-7, then blocks 5-7, etc). Here is an example of an incorrect output, where the model is focusing on the curves and edges of the logo, and predicting it comes from an Energy and Environment company when it is in fact Non Profit and Social. The second sample is the model focusing on the top portion of the logo, and correctly predicting that the logo belongs to an Energy and Environment company.

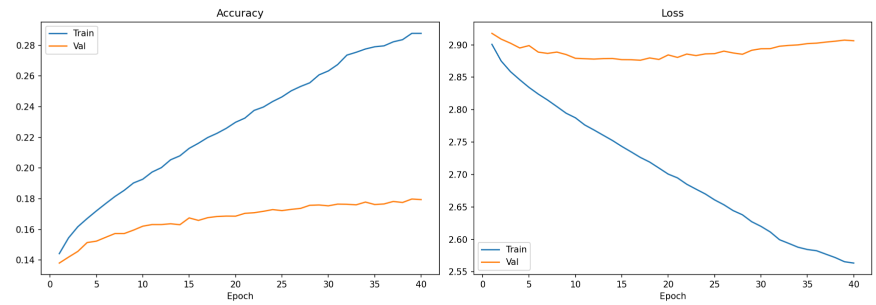


Figure x.y+1: The training and validation accuracy curves had upward trends for the 40 epochs, with training accuracy reaching approximately 28% and validation accuracy plateauing around 16-17%. The divergence between the two curves indicates mild overfitting as the model continues learning patterns in the training data that do not fully generalize to unseen logos. This is further reflected in the loss curves, where training loss decreases steadily while validation loss initially drops before leveling off and slightly increasing in later epochs. The gap between training and validation performance suggests that while progressive unfreezing of EfficientNetB0 layers did improve feature extraction, the lack of distinction between visually similar sectors was a challenge for the model.

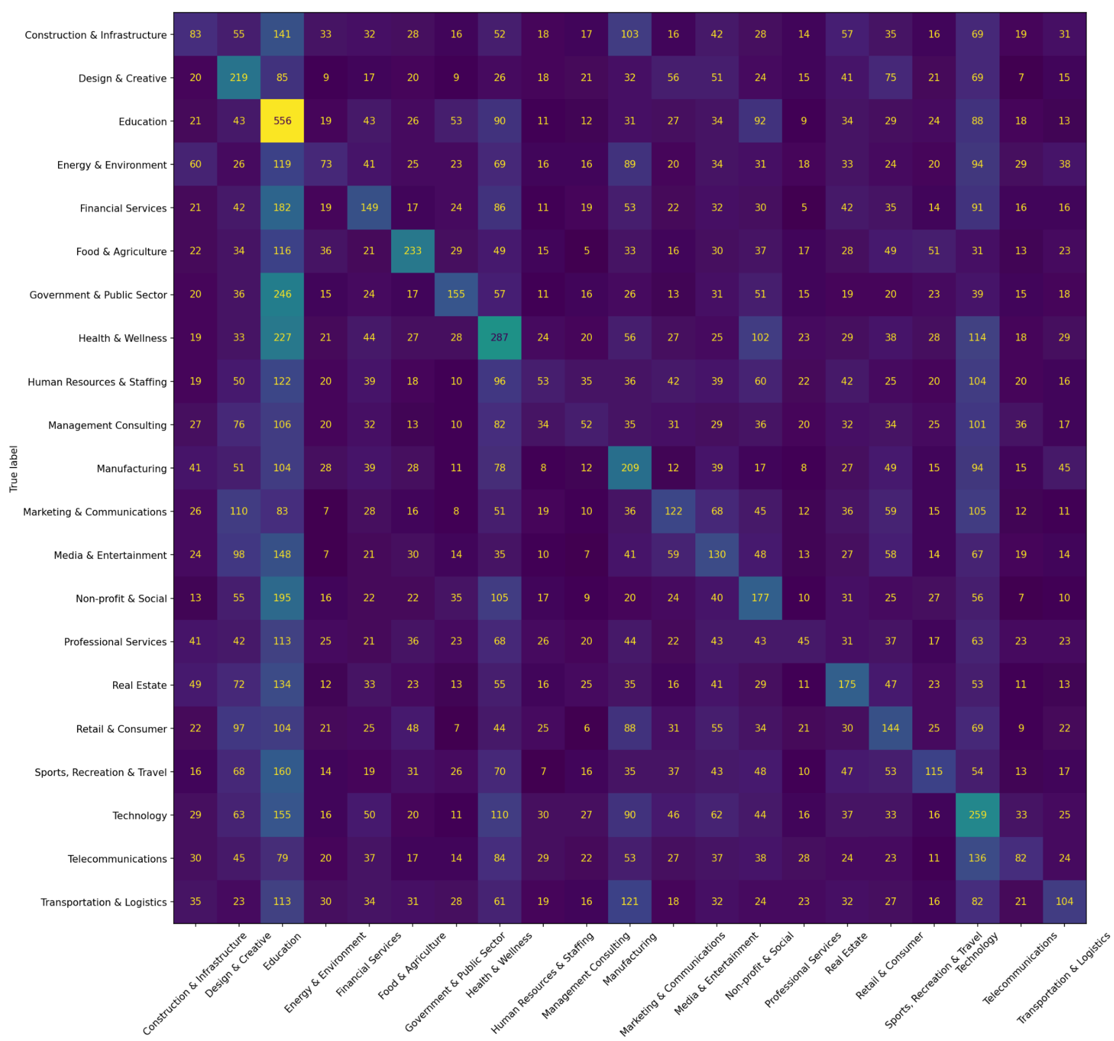


Figure x.y+2: The confusion matrix reveals the per-class prediction behavior of the model across all 21 sectors. The diagonal entries represent correct classifications, with Education (556), and Health and Wellness (287) showing the strongest performance. The majority of predictions are scattered away from the diagonal, which shows a widespread misclassification across sectors. A noticeable issue is that the model is incorrectly predicting many different logos into Education, Health & Wellness, and Technology suggesting the model is struggling for visually ambiguous logos and just puts them under those three categories. Sectors with weaker diagonal counts such as Management Consulting, Professional Services, and Human Resources & Staffing are almost indistinguishable to the model, which may be due to logos in business-oriented sectors sharing similar minimalist typographic styles with little individual visual signal. These results suggest that many industries do not embed their sector identity visually in a consistent or learnable way.

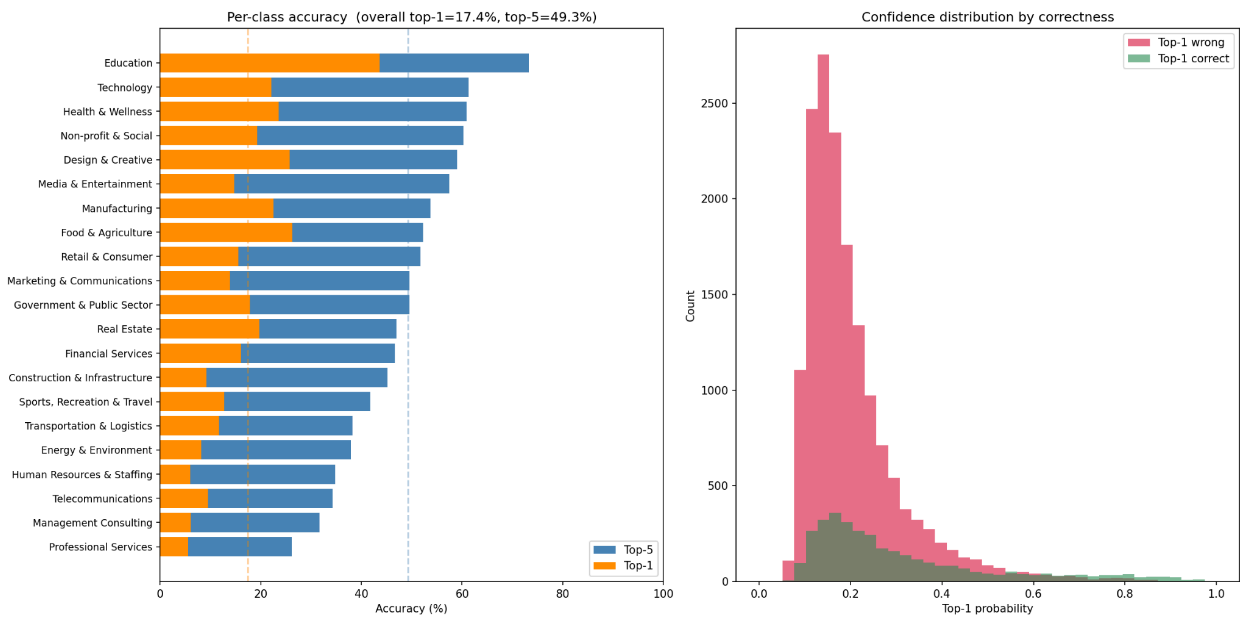


Figure X.y+3: The per-class accuracy chart shows top-1 and top-5 accuracy per sector, with an overall top-1 of 17.4% and top-5 of 49.3%. Education and Technology have the highest top-1 accuracy, while Professional Services and Management Consulting have the worst. The gap between top-1 and top-5 bars is notably large across all classes, indicating the model frequently ranks the correct answer within its top 5 predictions even when it fails to select it as the top-1. This suggests the model has learned some meaningful sector signal but lacks the precision to consistently rank the correct class first.

The confidence distribution chart further supports this as the incorrect predictions (red) are heavily concentrated at low probabilities around 0.1-0.2, meaning the model is generally uncertain when it gets things wrong rather than confidently wrong. Correct predictions (green) show a broader spread toward higher confidence values. This means that the model tends to be more confident when it is right and less confident when it is wrong.

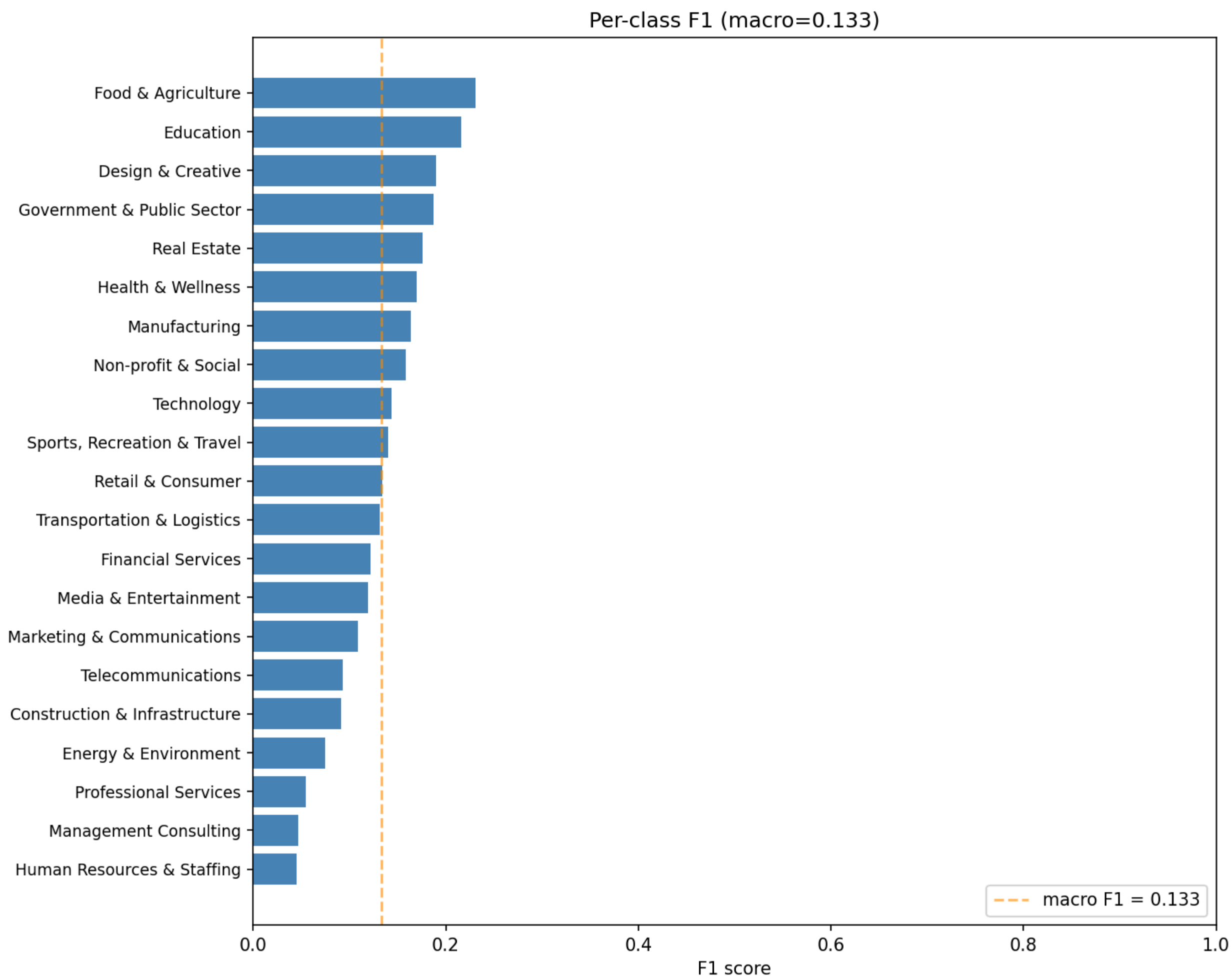


Figure X.y+4: The per industry F1 score shows the precision and recall for each sector, with a macro average F1 score of 0.133. Food & Agriculture and Education lead with the highest F1 scores, while Human Resources & Staffing, Management Consulting, and Professional Services are at the bottom, reinforcing the pattern seen in the confusion matrix where business oriented sectors are the hardest to distinguish. The dashed orange line representing the average highlights how most classes cluster below 0.2, showing a weak but non-trivial performance across all of the sectors. Since generally most sectors struggled, this points to a potential fundamental data challenge rather than just a model architecture issue.

In general, the model did not perform as well as we had originally expected. It consistently overfit on certain classes, to the point where we had to remove Legal and the problems remained, and underfit on others. There is, however, a clear diagonal along the confusion matrix indicating that the model did learn certain features of each class. It is also worth noting that classification accuracy was ~3.6x higher than random chance (4.8% vs. 17.4%) and top-5 accuracy was ~2.07x higher than random chance (23.8% vs 49.3%).

Limitations and Future Research

There were two significant limitations to our research.

First, since logos often come in different sizes and we required the data to be consistently sized to 224x224, the quality of images was highly variable. Some logos were pixelated while others had evidently been downsized and thus were very high fidelity. Additionally, some logos had solid colour backgrounds while others were white/transparent, a factor which could have influenced the model's projects.

Second, we were limited by time and compute. We had up to 1.4 million potential logos from the original Kaggle dataset, but cut it to 230,000 to reduce training time. Having more logos would have increased the number of entries for each sector and even enabled us to train a model to classify the 145 industries available in the original dataset rather than the 21 or 23 categories we distilled them down to. A task this complex benefits heavily from more data by giving the model more context into broad patterns. We also could have benefitted from using more complex and/or heavyweight models

Further research could attempt to classify logos within a particular sector, for example within technology, categorizing telecommunications, consumer software, hardware, semiconductors, etc. Additionally, attempting to classify not by sector or industry, but by era, company size, country, or region could yield very interesting results about the impact of those factors on design. Lastly, incorporating real design briefs into the training data somehow could lead to much more nuanced analysis and potentially allow the model to look at logos more like a human rather than a machine.

Conclusion

Logos are far more complex than we had anticipated. There are strong confounding variables such as company size, era, country that can be as significant or more significant indicators than sector. However, when we attempted to include data such as year founded, company size, and country in the training data, the model overfit heavily and was unable to pick up wider patterns in the logos.

Some industries tend towards very simple monochromatic wordmark logos, such as legal and HR, causing the model to overfit on them. This is likely because these logos maintain the basic features of a logo without committing strongly to an identity.

Furthermore, today's minimalistic era of design makes our project even harder. Many big companies like Apple and Nike have very similar design languages despite having completely different focuses.

Our results notwithstanding, algorithmic analysis of logos remains important given both the importance of logos in creating brand image and the newfound accessibility to logo creation with tools like NanoBanana. We look forward to seeing future research in this field.

Mappings

Mappings

Technology

1. information technology and services
2. computer software
3. internet
4. computer hardware
5. computer networking
6. computer & network security
7. wireless
8. information services
9. semiconductors
10. computer games
11. Nanotechnology
12. biotechnology

Marketing & Communications

11. marketing and advertising
12. public relations and communications
13. market research

Construction & Infrastructure (4)

14. construction
15. civil engineering
16. building materials
17. facilities services

Management Consulting (1)

18. management consulting

Real Estate (2)

19. real estate
20. commercial real estate

Financial Services (8)

21. financial services
22. banking
23. investment management
24. investment banking
25. capital markets
26. venture capital & private equity
27. insurance
28. accounting

Health & Wellness (7)

29. health, wellness and fitness
30. hospital & health care
31. medical practice
32. mental health care
33. alternative medicine

34. veterinary

35. Pharmaceuticals
36. medical devices

Design & Creative (7)

36. design

37. graphic design

38. photography

39. fine art

40. arts and crafts

41. architecture & planning

42. animation

Education (7)

43. education management

44. higher education

45. primary/secondary education

46. e-learning

47. professional training & coaching

48. libraries

49. program development
50. research

Non-profit & Social (7)

50. non-profit organization management
51. civic & social organization
52. philanthropy
53. religious institutions
54. fund-raising
55. individual & family services
56. museums and institutions

Human Resources & Staffing (3)

57. human resources
58. staffing and recruiting
59. outsourcing/offshoring

Government & Public Sector (13)

60. government administration
61. government relations
62. public policy
63. political organization
64. international affairs

65. legislative office
66. military
67. public safety
68. law enforcement
69. judiciary
70. defense & space
71. think tanks
72. international trade and development

Energy & Environment (5)

75. oil & energy
76. renewables & environment
77. environmental services
78. utilities
79. mining & metals

Food & Agriculture (9)

80. food & beverages
81. food production
82. restaurants

83. farming

84. dairy

85. fishery

86. ranching

87. supermarkets

88. wine and spirits

Transportation & Logistics (7)

89. transportation/trucking/railroad

90. logistics and supply chain

91. maritime

92. warehousing

93. package/freight delivery

94. aviation & aerospace

95. import and export

Manufacturing (14)

96. mechanical or industrial engineering

97. electrical/electronic manufacturing

98. machinery

99. industrial automation
100. chemicals
101. plastics
102. packaging and containers
103. paper & forest products
104. glass, ceramics & concrete
105. textiles
106. printing
107. shipbuilding
108. railroad manufacture
109. automotive

Retail & Consumer (11)

110. retail
111. apparel & fashion
112. cosmetics
113. luxury goods & jewelry
114. furniture
115. consumer goods

116. consumer electronics
117. sporting goods
118. wholesale
119. business supplies and equipment
120. tobacco

Media & Entertainment (10)

121. media production
122. entertainment
123. broadcast media
124. music
125. publishing
126. online media
127. newspapers
128. motion pictures and film
129. performing arts
130. writing and editing

Sports

131. Sports

Recreation & Travel (6)

132. recreational facilities and services
133. leisure, travel & tourism
134. hospitality
135. gambling & casinos
136. events services

Telecommunications (1)

137. telecommunications

Professional Services (4)

142. consumer services
143. security and investigations
144. translation and localization
145. executive office

Tab 5

Introduction

- Logo classification by sector
- Easy to generate logos but taste/perception is hard
- Gap thing that tevin brought up

Data Collection

When beginning to search for data, we ran into a very specific problem. We found datasets with logo but not industry, industry but not name, name and industry but not logo. We also found logo repositories that required domain, but didn't allow you to batch fetch.ALTERNATIVES In the end, we decided on using this (hyperlink here) Kaggle dataset with over seven million companies in conjunction with logo.dev, who kindl
y agreed to provide a few hundred thousand logos in size 224x224 for research purposes in exchange for attribution.

Preprocessing

From the seven million rows the dataset provided, we removed entries with no value for industry, domain, or name, then grouped the 145 industries into 23 broader-scope sectors like 'technology' and 'marketing and communications.' We were left with 1.4 million entries, of which we took 230,000 (top 10,000 per sector by company size given those were the logos that logo.dev was most likely to have) and retrieved the corresponding logos from logo.dev. They were able to source ~210,000, which became our final dataset.
Anything else?

Architecture

- Efficient net b0 for efficiency given so many photos, image net weights bc already learned basic things like colour and shape
- Tried as feature extractor and w/ fine tuning

Iteration and Testing

- Unfreezing layers, then block 5 and block 4 (minor improvement)
- Tried removing categories, reorganizing categories on second pass (minor improvement), adding metadata (overfitted), more epochs (generally stopped with early stopping), changing learning rate (not much changed), increasing batch size (minor improvement), data augmentation (minor improvement), etc.
- Also tried basic four layer cnn, but underfitted