Combining the Predictions of Out-of-Domain Classifiers Using Etcetera Abduction

Andrew S. Gordon
Institute for Creative Technologies
University of Southern California
Los Angeles, USA
gordon@ict.usc.edu

Andrew Feng
Institute for Creative Technologies
University of Southern California
Los Angeles, USA
feng@ict.usc.edu

***Abstract*—Research in machine learning on Domain Adaptation has led to numerous methods for re-purposing high-performance pre-trained models for novel tasks, e.g., via fine-tuning a model with out-of-domain training data. When model weights are unavailable or otherwise fixed, there are fewer options available for exploiting its predictive power. In this paper we investigate whether the predictions of ensembles of fixed, pre-trained, out-of-domain image classification models can be used to improve the performance of an in-domain classifier, or replace it outright with comparable performance. Our approach involves computing the conditional probabilities from the confusion matrixes of out-of-domain predictions for in-domain training samples, then combining this information with prior probabilities and classification confidence using probability-ordered logical abduction, Etcetera Abduction, to select the most likely label for an in-domain test sample. We evaluate this approach using four image classification models in highly disparate domains. Results indicate that this method may be well-suited to applications where insufficient training data is available to train an accurate model on a novel task.**

***Index Terms*—I.2.6.g Machine learning I.2.3.d Inference engines I.2.4 Knowledge Representation Formalisms and Methods**

I. Introduction

The wide availability of large annotated datasets and pre-trained models has been instrumental to the rapid advancement of contemporary machine learning using deep neural networks. However, the commercial industry that has grown around these technologies is increasingly protective of both the datasets used to train their models and the resulting model weights, and sometimes offer only a paid API service to use their top-performing models. This shift toward closed and opaque models creates new hurdles for traditional methods of domain-adaptation, such as fine-tuning, which requires pre-trained model weights in order to tune a model to a novel, out-of-domain task. This raises the question, How best can we exploit the power of pre-trained models for out-of-domain tasks in application contexts where these models are both fixed and opaque?

In this paper, we investigate this question for the task of image classification, and evaluate a novel approach to the re-purposing of ensembles of pre-trained image classifiers for out-of-domain tasks. In this work, we consider the case where several pre-trained classifiers for various tasks are available for inference, but their training data and model weights are otherwise opaque. Our approach is to use any available training data for the novel image classification task to determine how each pre-trained model responds to out-of-domain input, computing confusion matrices between out-of-domain predictions and the actual labels in the available training data for the novel task. When assigning labels to test data, these confusion matrices are then used to translate the combined predictions of the pre-trained models into the most likely labels of the novel task. Our method uses a logic-based approach called Etcetera Abduction, a probability-ordered first-order logic abductive reasoning algorithm, which manages the combinatorial search for the most likely label given the evidence from classifiers from disparate domains.

After explaining our method, we describe a set of experiments involving four image classifiers trained using standard (available) datasets. We look specifically at differences between test accuracy when using a classifier trained on the base model's available training data, when incorporating evidence from ensembles of out-of-domain classifiers, and when using only these out-of-domain predictions to select a label for the novel task. Results indicate that our method may be well-suited to applications where insufficient training data is available to train an accurate model on a novel task, or where such a model for the novel task is unavailable.

The project or effort depicted was or is sponsored by the U.S. Army Research Laboratory (ARL) under contract number W911NF-14-D-0005, and that the content of the information does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

II. Related Work

Adapting pre-trained models to novel domains is a standard practice in contemporary machine learning using deep neural networks. Typically, a model is first pre-trained on a large out-of-domain dataset, then trained using the available in-domain data for the novel task. In this second step, training for the novel task can be done by "fine-tuning", where all model weights are updated via gradient descent, or alternatively by linear probing, where only the output layers of the model are tuned [1] [2] [3].

Our approach, where ensembles of out-of-domain pre-trained models are used without fine-tuning or linear probing, is most similar to the work of Li et al. [4]. In their work, the

output layer of each model in the ensemble is used as an input to an additional network that learns the translation between label spaces, while a specialized dispatcher layer weights the contribution of each ensemble model conditioned on an embedding of the input sample. Our approach differs from this previous work in that we do not train an additional network to aggregate the predictions of ensemble models. Instead, we compute the pairwise confusion between the predictions of pre-trained models and the training data of the novel task, and aggregate ensemble evidence using probabilistic reasoning. The practical impact of this difference is that, in our approach, the composition of the ensemble can be changed without requiring any additional retraining of an aggregating model.

III. Etcetera Abduction

Logical abduction, distinct from deductive or inductive reasoning, is a reasoning method that answers the question: Given a knowledge base of axioms and a set of input observations, what are the sets of assumptions would logically entail the observations, if they were true? Beginning with the work of Hobbs et al. [5], logical abduction has been employed in the fields of natural language understanding and commonsense reasoning as a means of identifying higher-level interpretations of sentences and other forms of discourse. In various implementations, logical abduction is conceived as a form of combinatorial search among the possible explanations for each observation, identified by back-chaining on knowledge base axioms to a defined depth d, unifying assumptions where possible, and ranking solutions according to various criteria. A popular implementation of logical abduction is Etcetera Abduction [6], which provides a probabilistic foundation for encoding defeasible knowledge base axioms and ranking possible solutions. Here, both the prior and conditional probabilities of and among assumptions are encoded as so-called etcetera literals, unique to a single knowledge base axiom, and used to (naively) compute the joint probability of a given set of entailing assumptions. To manage the size of the combinatorial search, Incremental Etcetera Abduction [7] adds a sliding context window w for input observations and a beam b of the most probable partial solutions that are incrementally expanded in the search for the most probable solution.

Although logical abduction is more typically associated with commonsense reasoning tasks involving brittle knowledge bases of hand-crafted axioms, the provisions of Etcetera Abduction for managing combinatorial search make it a useful tool in various other reasoning tasks where prior and conditional probabilities can be estimated from data. In our previous work [8], we applied Etcetera Abduction to the problem of assigning labels to multiple objects in images from the COCO dataset [9] and to multiple player actions in videos in the Volleyball dataset [10]. In these two tasks, we first estimated the prior and conditional probabilities for label assignments from the available training data, e.g., the conditional probability that an object is a "carrot" given that another object in the image has the label "carrot". Each of these estimates were (automatically) encoded as knowledge base axioms, with the probabilities reified as etcetera literals in first-order definite clauses. Then we applied a trained classifier to each entity in a given input context (image or video), and encoded the top-four most-confident class predictions for each entity as a separate input observation. Given these observations and the knowledge base of probability axioms, Etcetera Abduction was used to find the most probable set of assumptions (class label assignments for each entity), that would logically entail the observations. By incorporating co-occurrence statistics into the search for the most probable combination of label assignments, we demonstrated significant gains in accuracy over simply selecting the most confident class.

The results seen in this previous application of Etcetera Abduction in computer vision tasks sparked a new question: Could Etcetera Abduction also be used to aggregate predictions from out-of-domain classifiers and improve accuracy in novel image classification tasks?

IV. Method

To investigate the application of Etcetera Abduction in aggregating evidence from out-of-domain classifiers, we conducted four experiments involving four different classifiers, where each experiment used one classifier as the base and the remaining three as out-of-domain ensemble models, using the following method.

A. Base and Ensemble Models

We began by selecting four standard image classification datasets for use in our experiments, each with unique characteristics with respect to size, scope of labels, and number of classes, as follows:

CIFAR-100: A large dataset with 100 broad-coverage classes, e.g., *mouse, bicycle, telephone* [11]

Flowers-102: A small dataset with 102 narrow-coverage classes, e.g., *fire lily, corn poppy, siam tulip* [12]

Food-101: A large dataset with 101 narrow-coverage classes, e.g., *beignets, lasagna, waffles* [13]

Fashion-MNIST: A large dataset with 10 narrow-coverage classes, e.g., *trouser, sneaker, ankle boot* [14]

Using the standard training data split for each dataset, we trained our own image classification model using the Torch machine learning library. As a model architecture, we selected ResNet-18 in each case, consisting of 17 convolutional layers, a fully-connected layer, and an additional softmax layer to perform the classification task. All models were trained with 10 epochs, using cross entropy loss as the criterion, and an SGD optimizer with learning rate of 0.001 and momentum of 0.9.

Table I describes each dataset and the accuracy of the resulting model on its own test data. As expected, the observed accuracy of each model greatly depends on the size of the training data split and the number of classes. While none of these classifiers achieves the performance of top-performing models seen in previous research using these datasets, their

accuracy is representative of a generic application of the ResNet-18 architecture and training regime.

TABLE I
FOUR TRAINED RESNET-18 CLASSIFIERS USED IN THIS RESEARCH.

Dataset	Classes	Training	Test	Test accuracy
CIFAR-100 [11]	100	50000	10000	0.3921
Flowers-102 [12]	102	1020	6149	0.1057
Food-101 [13]	101	75750	25250	0.3569
Fashion-MNIST [14]	10	60000	10000	0.9118

B. Knowledge Base Generation

Our experimental design for each of the four tasks was to select one of datasets as the base domain, and use the three remaining classifiers as an out-of-domain ensemble, evaluating the degree to which the ensemble could augment or replace the accuracy of the base classifier for the dataset using its test data split. In each task, we generated a distinct knowledge base of axioms (first-order definite clauses) to encode prior and conditional probabilities relevant to the interpretation of ensemble output. In each case, these probabilities were computed empirically from only the available training data of the base domain. In this section, we describe how this knowledge was generated as exemplified by using the Flowers-102 dataset as the base domain with an out-of-domain ensemble of CIFAR-100, Food-101, and Fashion-MNIST classifiers.

We computed prior probabilities based on the number of examples for each class in the training data split. In the Flowers-102 dataset, and the others in our study as well, the training data is distributed equally among the class labels ($1/102 = 0.0098$). With these numbers, we generated prior probabilities for base domain class labels, as in this example:

$$(\forall s)\ (Etc_{pinkprimrose}(0.0098, s) \rightarrow Class(s, Flowers, \text{“pink primrose”})) \quad (1)$$

Next, we computed the conditional probabilities of base domain class labels given the predictions of each model in the ensemble. This was done by generating a confusion matrix for each out-of-domain classifier when applied to the training data samples in the base domain. Figure 1 shows a traditional confusion matrix as it is commonly seen, in this case showing the predictions of our trained CIFAR-100 ResNet18 model on the test data split of the CIFAR-100 dataset. Note that this confusion matrix exhibits the distinctive diagonal line that is indicative of a well-trained, accurate model. In our approach, however, we instead generate a confusion matrix for the ensemble classifier when applied to the *training* data of the base domain. For example, Figure 2 shows the confusion matrix for predictions of the CIFAR-100 ResNet18 model when presented with images in the Flowers-102 training data split. This figure exhibits distinct vertical bands, indicating that our CIFAR-100 ResNet18 model is biased toward selecting among only a handful of classes when given an image from the Flowers-102 training data. For example, the second dark band around the “predicted” label number 40 indicates that our model will often mistake a given flower for a “keyboard”.

We treat the values in these confusion matrixes as conditional probabilities, e.g., the probability of predicting a given CIFAR-100 label given that the actual label is a given Flowers-102 label. These probabilities are encoded as knowledge base axioms as in the following example:

$$(\forall s)\ Class(s, Flowers, \text{“pink primrose”}) \wedge Etc_{keyboard|pinkprimrose}(0.2, s) \rightarrow Class(s, Cifar, \text{“keyboard”})) \quad (2)$$

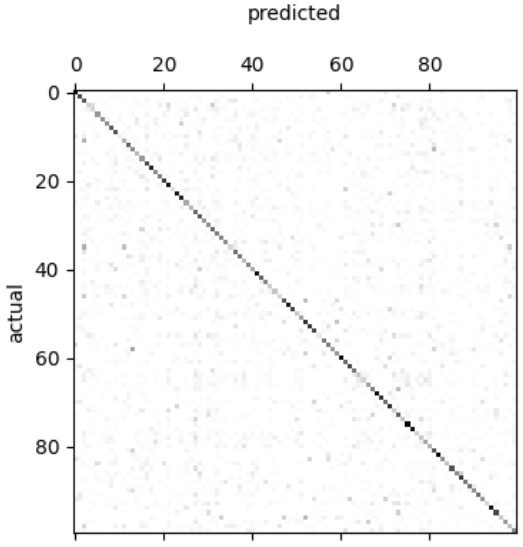

Fig. 1. CIFAR-100 predictions for CIFAR-100 test data.

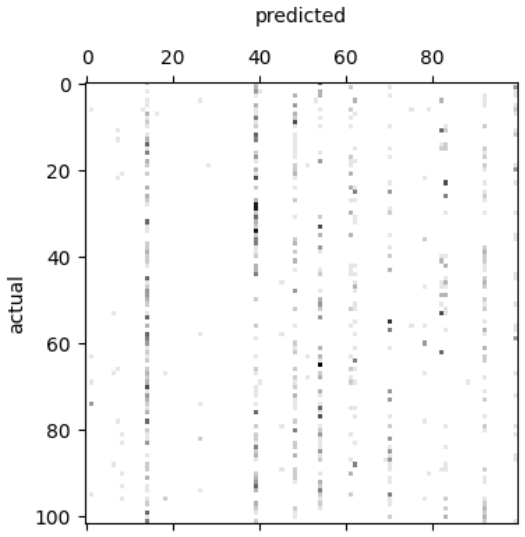

Fig. 2. CIFAR-100 predictions for Flowers-102 training data.

With one axiom generated for each pairwise combination of base domain label and each out-of-domain label for each ensemble classifier, the number of generated axioms in this work is quite large, e.g. 21,624 generated axioms for the Flowers-102 task. To prevent Etcetera Abduction from considering solutions that have zero joint probability, we remove axioms that encode a zero conditional probability. For the Flowers-102 task, the remaining knowledge base consists of only 1,290 generated axioms.

C. Base and Ensemble Predictions

Etcetera Abduction is used to find the most probable base domain label for a given test sample, given the observed output of the ensemble classifiers. In our experiments, we evaluate the case where the predictions of a base-domain classifier (trained on the available training data) is included as well, alongside ensemble classifiers. When included, the aim is to see whether the predictions of the base classifier can be improved by including evidence from the ensemble classifiers. When excluded, the aim is to see how well the ensemble can replace the base classifier if it were not available.

In each task, we present each sample in the test data split to each model, and encode its top four most-confident label predictions as a first-order logical literal, as in the following example for Flowers-102 sample number 121 given as input to the CIFAR-100 ResNet-18 model:

$$Top4(S121, Cifar, \text{“worm”}, 0.2265, \text{“plate”}, 0.1399, \text{“keyboard”}, 0.1190, \text{“aquarium fish”}, 0.0524) \quad (3)$$

As done in our previous work [8], we treat confidence values as likelihoods, and factor these values into the probability of a given solution. In order to direct Etcetera Abduction to consider only one of the top four most-confident label assignments, we include four special axioms in each knowledge base to force the label selection and include the corresponding likelihood when computing a solution's probability. For example, the following axiom selects the third-most probable label from a classifier's top four, with its likelihood equal to its confidence (encoded in the first argument of the axiom's etcetera literal).

$$(\forall\ s, c, c1, p1, c2, p2, c3, p3, c4, p4) \quad Class(s, c, c3)\ \wedge Etc_3(p3, s, c, c3) \rightarrow Top4(s, c, c1, p1, c2, p2, c3, p3, c4, p4) \quad (4)$$

D. Search for Most-probable Label

To identify the most likely base domain label in our tasks, the $Top4$ literals for each ensemble classifier are passed as input to Etcetera Abduction[1], along with the knowledge base for the task. The search then commences by back-chaining from these input observations through knowledge base axioms until solutions are found that only contain etcetera literals. The joint probability of each of these solutions is then naively estimated as the product of the etcetera literals, and the base domain class label entailed by the most-likely solution is selected as the predicted label.

For example, Etcetera Abduction would unify the observed literal (3) with the consequent of axiom (4), leading it to consider two new assumptions in its search:

$$Class(S121, CIFAR, \text{“keyboard”}) \quad (5)$$

$$Etc_3(0.1190, Cifar, S121, \text{“keyboard”}) \quad (6)$$

The likelihood (0.1190) of the etcetera literal (6) would be included in the probability calculation for any solution involving this axiom, while the $Class$ literal (5) would further back-chain via axiom (2), above, which would further add two additional assumptions in the search:

$$Class(S121, Flowers, \text{“pink primrose”}) \quad (7)$$

$$Etc_{keyboard|pinkprimrose}(0.2, S121) \quad (8)$$

Again, the etcetera literal (8) would factor into the probability of the solutions, while the $Class$ literal would further backchain on axiom (1), above, such that the prior probability the base domain class label is included in solutions:

$$Etc_{pinkprimrose}(0.0098, S121) \quad (9)$$

Each of the other most-confident labels is similarly considered, as are each of the possibilities predicted by the other classifiers in the ensemble. When a particular combination of label predictions back-chain to the same base domain class label, they are unified as common factors in the solution, increasing its overall probability. When the most-likely solution is identified, the entailed class label is selected as the predicted label for the test sample.

V. Results

Table II shows the results of our four experiments, with the accuracy of the task's base model, the accuracy of our approach when the base model is included in the observations alongside the ensemble predictions, and the accuracy of the ensemble predictions without inclusion of the base model.

For the ensembles that included the base classifier in its observations, we see that the inclusion of the ensemble had nearly no impact on the resulting accuracy, except in one task. When compared to the accuracy of a ResNet-18 model trained using the Flowers-102 training data split, the ensemble improved accuracy by 4.36%, which represents a 41% gain over the poor performance of the base classifier. With 102 classes and only 10 training examples per class, the Flowers-102 dataset is indeed much too small to train an accurate model on its own, and the additional evidence provided by the ensemble made a substantial impact on performance. In contrast, this additional evidence had no benefit in the other three tasks.

For the ensembles that did not include the base classifier in its observations, we see markedly lower performance compared to the base classifier, except again in the Flowers-102 task. For the Flowers-102 task, the ensemble shows only a slight drop in accuracy compared with the poor-performing

[1]In this research, we used a Python implementation of Etcetera Abduction available at: https://github.com/asgordon/EtcAbductionPy

TABLE II
ACCURACY OF BASE CLASSIFIER, ENSEMBLE WITH BASE, AND ENSEMBLE WITHOUT BASE.

Base	Ensemble	Base accuracy	Ensemble w/base	Ensemble w/o base
CIFAR-100	Flowers-102, Food-101, MNIST-Fashion	**0.3921**	0.3712	0.0549
Flowers-102	CIFAR-100, Food-101, MNIST-Fashion	0.1057	**0.1493**	0.0872
Food-101	CIFAR-100, Flowers-102, MNIST-Fashion	**0.3569**	0.3529	0.0515
Fashion-MNIST	CIFAR-100, Flowers-102, Food-101	**0.9118**	0.9001	0.3381

base model. In contrast, while the ensembles each perform substantially above a majority baseline in each of the other tasks, each performs markedly worse than the base models.

VI. DISCUSSION

When interpreting the results of our experiments, we first consider the question, Why did we think this would work in the first place? The accuracy of each ensemble model on an out-of-domain input is always zero, e.g., a ResNet-18 model trained on Fashion-MNIST data will never guess "pink primrose" when presented with an image of a flower. However, we do not expect (or observe) that its predictions are going to be uniformly random over its out-of-domain labels. Intuitively, there must be some visual features in a picture of a pink primrose flower that push the model toward guessing "ankle boot" over "sneaker". Insomuch as these output labels are correlated with discriminate features in the Flowers-102 domain, then these out-of-domain predictions provide some information that should be helpful in selecting the most likely in-domain class label. While these correlations may be somewhat weak, we expected that combining such information from an ensemble of pre-trained classifiers would improve the performance of an in-domain base classifier.

Instead, our results suggest that the outputs of our ensemble are providing no additional information to help discriminate among classes in the base domain, except in the case where a base classifier is itself exceedingly inadequate. Where sufficient training data is available, the discriminating features that are learned by a base model seem to already include all of the discriminating information that might be gleaned from correlations with out-of-domain predictions.

Our results indicate that training a base classifier on the available in-domain training data is preferable to our method in most contexts. Still, out method may have some application when a training dataset is very small, as in the case with the Flowers-102 dataset, within some limits. As our method estimates conditional probabilities by presenting training data to ensemble classifiers, these estimates will degrade as the available training data decreases, leading to lower accuracy in ensemble predictions. With exceedingly tiny training datasets, a zero-shot multi-modal foundational model like CLIP [15] is likely to outperform our approach for most image classification tasks.

While the results in our experiments are largely negative, we see several improvements to our approach that could be explored in future work. First, we expect that much larger ensembles of pre-trained classifiers would improve accuracy, particularly when the domains of ensemble classifiers include those that more closely aligned with the base domain. While the inclusion of radically out-of-domain models did not have a substantial negative impact on performance in our experiments, we expect that the addition of several close-domain models would be highly beneficial. Likewise, the inclusion of only high-performing models in the ensemble might yield better results, allowing for more confidence in the predicted ensemble labels. Finally, we expect that obtaining better confidence estimates from our ensemble classifiers would improve the probabilistic ranking of candidate solutions. Well-calibrated models, where confidence of predicted labels corresponds to their likelihood, is critical when applying our method, as extreme overconfidence in an incorrect label is hard to overcome. The effect of calibration techniques such as temperature scaling [16] is an important consideration in future experiments.

One additional contribution of our research is to demonstrate a new application area for Etcetera Abduction, the software tool used in this research. Whereas the use of first-order logical abduction is highly unusual in contemporary computer vision research, we found Etcetera Abduction to be well-suited to tasks that involve combinatorial search and probabilistic reasoning. Our large automatically-generated knowledge bases of thousands of axioms are certainly among the largest ever used in Etcetera Abduction research, which has historically used smallish knowledge bases of hand-crafted commonsense axioms with fabricated probability estimates. We are encouraged by the performance of this tool given the large knowledge base and search space, and we will continue to look for opportunities to apply it in other computer vision tasks involving the aggregation and interpretation of evidence.

REFERENCES

[1] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International Conference on Machine Learning (ICML)*, 2020, p. 1597– 1607.

[2] A. Kumar, A. Raghunathan, R. Jones, T. Ma, and P. Liang, "Fine-tuning can distort pretrained features and underperform out-of-distribution," in *Tenth International Conference on Learning Representations (ICLR 2022) (Virtual)*, 2022.

[3] Z. Li, K. Ren, X. Jiang, Y. Shen, H. Zhang, and D. Li, "SIMPLE: Specialized model-sample matching for domain adaptation," in *Eleventh International Conference on Learning Representations (ICLR 2023)*, 2023.

[4] Z. Li, K. Ren, X. Jiang, B. Li, H. Zhang, and D. Li, "Domain generalization using pretrained models without fine-tuning," 2022. [Online]. Available: https://arxiv.org/abs/2203.04600

[5] J. R. Hobbs, M. E. Stickel, D. E. Appelt, and P. Martin, "Interpretation As Abduction," *Artificial Intelligence*, vol. 63, no. 1-2, pp. 69–142, Oct. 1993.

[6] A. S. Gordon, "Commonsense Interpretation of Triangle Behavior," in *Thirtieth AAAI Conference on Artificial Intelligence*. Palo Alto, CA: AAAI Press, 2016, pp. 3719–3725.
[7] ——, "Interpretation of the Heider-Simmel Film using Incremental Etcetera Abduction," *Advances in Cognitive Systems*, vol. 6, pp. 1–16, 2018.
[8] A. Gordon and A. Feng, "Searching for the most probable combination of class labels using etcetera abduction," in *Proceedings of the 57th Annual Conference on Information Sciences and Systems (CISS-2023), March 22-24, 2023, Baltimore, MD*, 2023.
[9] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.
[10] M. S. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat, and G. Mori, "A hierarchical deep temporal model for group activity recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1971–1980.
[11] A. Krizhevsky, "Learning multiple layers of features from tiny images," 2009. [Online]. Available: https://www.cs.toronto.edu/ kriz/learning-features-2009-TR.pdf
[12] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *2008 Sixth Indian Conference on Computer Vision, Graphics Image Processing*, 2008, pp. 722–729.
[13] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101 – mining discriminative components with random forests," in *European Conference on Computer Vision*, 2014.
[14] H. Xiao, K. Rasul, and R. Vollgraf. (2017) Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. [Online]. Available: https://arxiv.org/abs/1708.07747
[15] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 8748–8763.
[16] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On Calibration of Modern Neural Networks," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, 06–11 Aug 2017, pp. 1321–1330.