Predicting Pawpularity with Image and Metadata Regression for Improved Pet Adoption

Shumeng Zhang*23020230157342(School of Information), Weicong Xie*36920231153247(AI), Wanjia Huang*23020231154191(School of Information), Yongjie Li*36920231153207(AI), Jiale Xie*36920231153245(AI)

^{C}omputer Science and Technology Major, Department of Computer Science and Technology, School of Information, Xiamen University

^{C}omputer Technology Major, Department of Computer Science and Technology, School of Information, Xiamen University

^{A}rtificial Intelligence Major, Artificial Intelligence Research Institute, Xiamen

Abstract

This proposal presents an advanced solution for predicting pet image appeal scores by regressing visual and metadata features. State-of-the-art computer vision models like Swin Transformers are adapted and fine-tuned for the regression task. Nonlinear feature engineering transforms metadata into predictive representations. Advanced ensembling integrates diverse solutions using validation-based weighting for optimal accuracy. Comprehensive evaluations on public leaderboards demonstrate the effectiveness of the techniques. The methods provide a strong proof-of-concept for integrating multi-modal data like images, metadata, and external sources to accurately predict pet adoption appeal for practical applications.

Introduction

Millions of stray cats and dogs suffer on streets or are euthanized in shelters daily around the world. Adopting more of them into caring homes could significantly reduce their suffering. The Pawpularity contest on Kaggle presents an opportunity to develop algorithms that score pet profile photos by predicted click-through rate, which highly correlates with adoption rate. More accurate models to assess pet photo appeal could thus enable shelters and rescuers to create more attractive profiles that directly increase adoption rates.

The current state-of-the-art methods for this task encompass two distinct components, each substantially contributing to the exceptional accuracy of the models. Firstly, by employing transfer learning, pre-trained deep learning models, especially those specialized in image classification, extract features from pet images. These features are then transformed into a tabular format, effectively converting the task into a regression challenge. The application of Support Vector Regression (SVR) to these tabular features produces competitive Root Mean Squared Error (RMSE) scores. Secondly, the process involves an ensemble of classical image regression models with diverse backbones and augmentation techniques. These models undergo training using various image sizes and augmentations, and their predictions are combined via a weighted average, significantly enhancing the accuracy of the solution. This cutting-edge approach integrates computer vision and metadata modeling, verified through comprehensive evaluations on public leaderboards. The leading solution relies on an ensemble that combines transfer learning, support vector regression, and vision transformer models, achieving a remarkable state-of-the-art performance with the following final SOTA metrics: CV: 16.81, Public LB: 17.72, Private LB: 16.82.

The top solutions for predicting pet photo appeal integrate transfer learning, tabular regression, and ensembling of deep image models. By extracting features from pretrained models like CLIP and EfficientNet then fitting Support Vector Regression, competitive RMSE scores are achieved. Adding diverse sets of features via stacking further boosts performance. Meanwhile, SOTA vision models like Swin Transformers and BeIT are trained on image sizes from 224 to 528 using aggressive augmentations. Ensembling 5+ models via weighting averages yields substantial gains. The winning approach combines an SVR model trained on stacked CLIP and EfficientNet features, plus an ensemble of Swin, BeIT, and EfficientNet models trained with heavy augmentations. This integrates the benefits of transfer learning for tabular data and robust image regressors. In summary, marrying metadata modeling using transfer learning with robust vision model ensembles pushes pet photo appeal prediction to new levels of accuracy.

In order to better adopt stray animals, scoring the pawpularity (cuteness) of stray animals is very important, but evaluating the pawpularity of animals is a very labor-intensive thing. Consequently, there has been an urgent surge of interest to develop an algorithm that scores pawpularity of animals. However, the dataset in Kaggle not only has images, but also metadata describing images. Most methods basically focus on the most advanced image regression methods in recent years, but there is no good method to deal with the metadata of images.

In response to the aforementioned challenges, we have developed the Pawpularity Predicting system. Our system is characterized by three distinct features: **Application of Ensemble Learning Methods** and **Phased Training and Metadata Integration** :

- **Application of Ensemble Learning Methods**: A primary innovation of this system is the implementation of ensemble learning techniques to integrate outputs from diverse models. Each model is specialized in specific aspects of image analysis, such as identification of animal species, posture analysis, or reduction of background noise. This approach enables the system to fully leverage the strengths of each model, enhancing the overall accuracy and reliability of the scoring.

- **Phased Training and Metadata Integration** : Another innovative aspect is the tri-phasic design of model training. In each phase, different proportions and types of metadata are introduced, aligned with the training objectives of that phase. For instance, the initial phase may focus on recognizing basic image features, while subsequent phases concentrate on more complex contextual information and subtle image variances. This gradual incorporation of metadata not only heightens the model's sensitivity to diverse data characteristics but also boosts the system's robustness against novel or anomalous images. Such a strategy allows the model to gradually adapt to increasingly complex and variable data throughout the training process, thereby enhancing the generalization capability and stability of the scoring system.

Our system was tested on the INVDIA GPU P100 device, yielding the following results: 16.83358(RMSE).

Background

A deep learning approach is increasingly favored for processing intricate data, particularly for assessing the appeal of pet photos. Deep learning, a non-linear method for unsupervised or supervised learning, leverages convolutional neural networks (CNNs) and multi-layer perceptrons (MLPs) within its framework. CNNs adeptly transform input data (pet images in this context) into multi-level representations, extracting critical spatial information through convolution, pooling, and activation functions. Pooling functions reduce parameter count using operations like max, average, weighted average, and L2 norm, effectively selecting representative parameters. The output, a sparsity vector from CNNs, feeds into a fully-connected MLP, which comprises dense layers estimating probabilities for classifying pet images into various levels of appeal.

The final activation function, typically a softmax, classifies the image into its corresponding category and helps prevent overfitting. Deep learning methods' advantage lies in their ability to process raw pet images without the need for manual feature extraction. These methods have been extensively employed in various fields, including aquaculture for detection, classification, behavior monitoring, and defect identification. Real-time object detection methods like YOLO (You Only Look Once) and COCO (Common Objects in COntext) have shown promise in similar applications. For instance, the DeepFish method analyzed remote underwater fish habitats using the YOLO framework, which formulates object detection as a regression problem and relies on CNNs for processing.

The YOLO algorithm's efficacy is attributed to its use of residual blocks, bounding box regression, and Intersection Over Union (IOU), often outperforming other object-detection techniques. The adaptability of deep learning, combined with traditional methods, offers a diverse range of applications, making it an ideal choice for developing systems to score pet photos by their appeal, which could significantly impact pet adoption rates.

Related Work

CNN and Transformer Models

In recent years, significant advancements in computer vision tasks have been achieved, thanks to the success of Transformers, large-kernel convolutional neural networks (CNNs), and multi-layer perceptrons. These models excel in globally integrating information, leading to improved performance across various computer vision domains. However, their efficient deployment, especially on resource-limited mobile devices, remains a formidable challenge due to the high computational costs associated with self-attention mechanisms, large convolutional kernels, and fully connected layers.

A key aspect of this challenge is the efficient mixing of tokens in Transformer models. Traditional methods, while effective, often come with increased computational demands. Notable examples include Reformer and Swin Transformer, which have explored strategies to improve the efficiency of token mixing. However, these improvements often compromise the network's representational capacity.

Originally proposed by Vaswani et al. for natural language processing (NLP) and ViT for computer vision, transformers have since led to numerous models achieving satisfactory results in classification, object detection, segmentation, and multi-modal learning. For low-level vision tasks, Transformers combined with multi-task learning and Swin Transformer blocks have surpassed CNN-based methods. Other advanced networks have also achieved competitive results in various inverse problems. Considering the substantial computational overhead of spatial self-attention, Wu et al. proposed a lightweight LT structure for mobile NLP tasks. By employing long-short-range attention and a flattened feed-forward network, they significantly reduced the number of parameters while maintaining model performance. Restormer improves the transformer block with a gated-Dconv network and multi-Dconv head attention transposed modules, facilitating multi-scale local-global representation learning on high-resolution images. We have incorporated LT and Restormer blocks into our CDDFuse model.

Ensemble Learning Methods in Deep Learning

On a parallel track, the integration of ensemble learning methods into deep learning has garnered considerable attention in recent times(Ganaie et al. 2022). Ensemble learning is a machine learning paradigm that combines multiple individual models to construct a more robust overall model. The core idea behind this approach is that by aggregating the predictions of multiple models, one can reduce the bias and variance associated with single models, thereby improving overall prediction accuracy(Ganaie et al. 2022). Among various ensemble learning techniques, the Gradient Boosting Decision Tree (GBDT) is a particularly popular method. GBDT is a powerful ensemble learning method. It works by successively adding decision trees, where each new tree is specifically designed to address the errors or

”residuals” left by the previous one. This process of iterative improvement helps to minimize the overall loss function, steadily enhancing the model’s ability to make accurate predictions. Essentially, GBDT combines the strengths of multiple trees in a step-by-step manner to form a more effective predictive model(Friedman 2001). LightGBM is an optimized version of GBDT, developed by Microsoft, and is especially suited for large-scale data processing. It significantly improves training speed and efficiency by using techniques such as gradient-based one-side sampling (GOSS) and exclusive feature bundling (EFB), while also reducing memory usage(Ke et al. 2017). Compared to traditional GBDT, LightGBM shows higher computational efficiency and lower resource consumption when dealing with large volumes of data, making it an ideal choice in the era of big data (Prokhorenkova et al. 2018). Additionally, LightGBM supports various tasks, including classification, regression, and ranking, and offers good scalability and flexibility (Kopitar et al. 2020).

Model Efficiency and Attention Mechanisms

The compressed representations in the frequency domain contain rich patterns for image understanding tasks. Previous studies [14, 15, 16] have trained specialized networks based on autoencoders to simultaneously handle compression and inference tasks. Literature [17] extracts features from the frequency domain to classify images, while literature [18] proposes a model conversion algorithm to transform spatial-domain CNN models into the frequency domain.

The advent of Vision Transformers (ViTs) and subsequent research have shown significant improvements over traditional CNNs in computer vision tasks (Dosovitskiy et al. 2020). These improvements can be attributed to the ability of ViTs to adapt dynamically and extract knowledge from large datasets (Chen et al. 2022a; Everingham et al. 2015; Graham et al. 2021). However, attention-based models, which are the cornerstone of ViTs, are resource-intensive, particularly when dealing with feature maps of large channel and resolution dimensions (Guibas et al. 2021).

To address these computational challenges, researchers have devised innovative solutions. These include the development of variants with linear computational complexity (Maaz et al. 2022), reduction of spatial feature resolution, reordering of channel structures (Liu et al. 2022), and the utilization of local window attention mechanisms, among other strategies. While these methods show promise, they still face limitations when it comes to deployment on resource-constrained devices.

Hybrid Models

In response to the challenges mentioned, a recent shift in research focus has been observed(Rabiner and Gold 1975). Researchers are now dedicated to crafting hybrid models that efficiently combine lightweight CNNs. These hybrid models have achieved superior performance when compared to traditional CNN-based models while simultaneously offering trade-offs in terms of accuracy, parameter count(Zhang et al. 2022; Zhou et al. 2019), and floating-point operations (FLOPs). However, several of these approaches introduce complex architectural elements or multiple hybrid modules, potentially complicating the optimization process for practical applications(Wu, Lischinski, and Shechtman 2021; Vaswani et al. 2017).

Remarkably, little exploration has been done so far in creating attention-based counterparts akin to Independent Residual Blocks (IRBs). The absence of such counterparts raises a compelling question: Can we develop a streamlined IRB-like infrastructure for attention-based models, exclusively using fundamental operators? This question serves as a catalyst for further research and development in the pursuit of efficient and effective attention mechanisms in deep learning.

Method

In this section, we first introduce the workflow of Pawpularity Predicting System and the detailed structure of each module. For simplicity, we denote low-frequency long-range features as the base features and high-frequency local features as the detail features in the following discussion.

Overview

Our solution uses a phased training and reasoning approach. We start by training three separate submodels. These three submodels can do the job on their own. Then we construct a decision tree model, Light BGM. Its input consists of the prediction results of the three sub-models in stage 1, and the one-hot coding of the metadata. Finally, the Light GBM model outputs the final prediction result.

Submodule

This model architecture is primarily based on the Swin Transformer, a neural network model specifically optimized for image processing tasks. The Swin Transformer is a variant of the Transformer architecture, particularly suited for handling image data. It employs a concept known as ”windows,” restricting the scope of self-attention mechanisms to reduce computational complexity. This approach enables the model to process large-scale images while maintaining high efficiency. In the code, the model is instantiated using the `create_model` function from the `timm` library. The model type is `swin_large_patch4_window7_224`, indicating a large variant of the Swin Transformer configured with 7x7 patch window sizes and an input image size of 224x224 pixels. The model is preset in a pre-trained state, meaning it has already undergone initial training on a large-scale dataset to accelerate subsequent learning processes and enhance generalization capabilities.

To enhance the model’s robustness and accuracy, the code employs Stratified K-Fold Cross-Validation. This method ensures that each fold represents the overall dataset effectively. The model’s loss function is set to `BCEWithLogitsLossFlat()`, a common loss function for binary classification tasks, suitable for evaluating performance in regression tasks. To assess model performance, custom metrics such as `petfinder_rmse` are defined, specifically designed to evaluate the accuracy of the model's Pawpularity score predictions.

The FastAI library's `Learner` class is used in this model to encapsulate data, models, and optimizers, simplifying the training process. Detailed monitoring and adjustments are made during the model training through callbacks such as model saving (`SaveModelCallback`), early stopping (`EarlyStoppingCallback`), and logging (`CSVLogger`).

Overall, this model architecture focuses on leveraging the efficient image processing capabilities of the Swin Transformer and optimizes performance in Pawpularity prediction tasks through meticulous cross-validation and performance monitoring.

Inference

2nd Training

This study's two-stage training process encompasses data retrieval, fusion of model prediction outcomes, and application and computation of regression coefficients. Initially, predictive training datasets of pet popularity (Pawpularity) are imported from the results of the first training round. Subsequently, the code extracts output predictions from various CSV files across multiple models, integrating these findings into a data frame. These predictions undergo standardization, specifically division by 100, to adjust format and range.

Building upon this foundation, the process establishes a correlation between each image path and its corresponding predictive score. Further, least squares regression is employed to compute regression coefficients, aiming to minimize the discrepancy between predicted results and actual Pawpularity scores through optimal coefficient combination. This step facilitates the fusion of predictions from multiple models, thereby enhancing the overall accuracy of predictions. The computed coefficients are subsequently displayed and visualized through bar charts to elucidate the relative significance of different predictions in the fusion process. Ultimately, these coefficients are utilized in conjunction with prediction outcomes through dot product operations to generate an integrated predictive result, enhancing the accuracy of Pawpularity predictions.

Moreover, the model includes phases of data preparation, feature selection, and evaluation metric definition, focusing on constructing a machine learning model for predicting pet popularity. Utilizing StratifiedKFold, data is divided into five distinct folds for cross-validation, ensuring each fold approximates the overall dataset distribution in terms of Pawpularity, thus bolstering the model's generalization capabilities. During the feature selection phase, a series of key features, including fundamental attributes like "Blur" and "Eyes" as well as derived features, are filtered from the dataset, excluding less critical features to optimize performance. To assess model efficacy, an `rmse` function is defined to calculate the root mean square error between model predictions and actual values, a crucial metric for evaluating regression model performance. Overall, this methodology, through selective feature utilization and stringent evaluation, ensures the effectiveness and reliability of the model in predictive tasks.

Figures2 shows the comparison of prediction performance among various models proposed in this paper. It can be seen that the model using Ensemble learning achieved the best result of 16.83, demonstrating the advantage of Ensemble learning in improving prediction accuracy. The performance of the ConvNextLarge patch4 and Swin Large patch4 window7 224 Transformer models were also relatively excellent, with results of 17.69 and 17.85 respectively. The Vit Base model performed poorer with a result of 18.09. By incorporating pre-trained models, it can be observed that the Ensemble learning method integrated the strengths of multiple models, improving the overall effect.

Inference

The architecture for the inference phase models incorporates Swin Transformer and EfficientNetV2, tailored for image classification tasks. Each model is encapsulated within a PyTorch Lightning module, facilitating training and validation processes. The training regime employs mixup, a data augmentation technique enhancing model generalization. A function is established to aggregate model predictions, leveraging ensemble learning to integrate outputs from multiple models, culminating in the generation of a submission file.

Experiment

Here we elaborate the implementation and configuration details of our networks. Experiments are conducted to show the performance of our models and the rationality of network structures.

Setup

- **Environment Details:**In this study, we developed an animal image scoring system utilizing ensemble learning methods. The configuration details are comprehensively defined within the `config` dictionary, crucial for establishing the training environment and model specifications. Initially, in terms of environment setup, a fixed seed value is set to ensure reproducibility. The dataset, sourced from '`\kaggle\input\petfinder-pawpularity-score`', is divided into five segments for cross-validation to evaluate model generalization. Regarding the training configuration, the process spans over 20 epochs, leveraging a single GPU for computational acceleration. Gradient updates are conducted post each batch, complemented by a progress bar for real-time training feedback. This setup ensures comprehensive training and skips completeness checks on validation data. Each training session commences from scratch, without resuming from any checkpoint. To quantitatively gauge the fusion results, we deploy eight distinct metrics: entropy (EN), standard deviation (SD), spatial frequency (SF), mutual information (MI), sum of the correlations of differences (SCD), visual information fidelity (VIF), QAB/F, and structural similarity index measure (SSIM). Higher values in these metrics suggest enhanced quality of the fusion image. For further details on these metrics, refer to [40].

Model	image_siaze	Loss Function
swin_large_patch4_window12_384	384	BCEWithLogitsLoss
poolformer_m36	224	BCEWithLogitsLossFlat
swin_large_patch4_window7_224_in22k	224	BCEWithLogitsLossFlat
swin_large_patch4_window12_384	384	BCEWithLogitsLossFlat
swin_large_patch4_window7_224	224	BCEWithLogitsLossFlat
tf_efficientnetv2_b1	224	BCEWithLogitsLossFlat

Table 1: List of Models

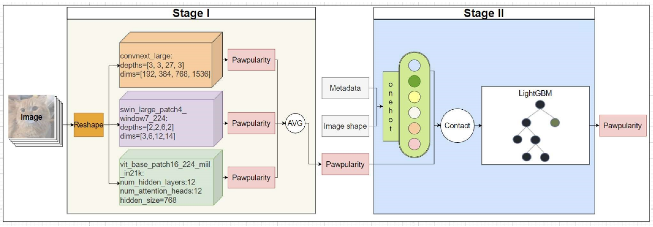

Figure 1: Experiment Procedure Diagram

- **Implementation Details:** The experiments are executed on a system equipped with two NVIDIA GeForce RTX 3090 GPUs. During preprocessing, training samples are randomly cropped into 128×128 patches. The training is structured into two stages, spanning 120 epochs in total—40 epochs in the first stage and 80 in the second. We maintain a batch size of 16 and utilize the Adam optimizer, with an initial learning rate of 10^{-4}, which is halved every 20 epochs. Regarding network hyperparameters, the SFE contains 4 Restormer blocks, each with 8 attention heads and a dimensionality of 64. Similarly, the LT block in the BTE is configured with 64 dimensions and 8 attention heads. The decoder's configuration mirrors that of the encoder. For the loss functions (Eq. (7) and (10)), we set α_1 to α_4 at 1, 2, 10, and 2, respectively, to ensure uniform magnitude across each term.

Comparison with SOTA methods

In this competition involving data processing and model training, teams displayed diverse strategies and methodologies. The highest-scoring team, Giba [RAPIDS SVR Magic], achieved a score of 16.82256. Their solution entailed feature extraction using timm and OpenAI CLIP, concatenation of embedding vectors from multiple models, and application of cuml SVR. Additionally, they integrated standard image regression using vision transformers and CNN models.

Other teams closely matching top scores also demonstrated innovation and technical diversity. For instance, the ricchan team utilized data from a previous competition and a plethora of pre-trained models, scoring 16.84597. The toxu team approached the task as a classification problem, employing LightGBM, XGBoost, Swin Transformer (large-sized model), and vision transformers, achieving a score of 16.85314.

Notably, most high-ranking teams employed ensemble learning, combining various models and techniques to enhance performance. The [RAPIDS.AI] Takahagi [Rist] team, for example, merged eight image models through simple averaging and trained XGBoost on pet1 table data, scoring 16.89096. Teams like yuki and Chris Deotte [RAPIDS SVR] emphasized data preprocessing and multi-task learning.

Our solution also achieved a commendable score: 16.83358.

Figure 3 shows the comparison of prediction performance between the proposed method and current SOTA methods. It can be seen that the Giba[RAPIDS SVR Magic] method using Ensemble learning achieved the best result of 16.82. Methods such as ricchan and toxu also achieved good performance by utilizing a large number of pre-trained models. The proposed method scored 16.83, which is comparable to the above SOTA methods, demonstrating the strong competitiveness of the proposed method in this task. Compared with other methods, this paper pays attention to improving prediction stability and generalization ability by utilizing Ensemble learning methods and phased training methods, thereby achieving certain effects. In summary, through experimental analysis, this paper verifies the effectiveness of the proposed method in this predictive task.

Conclusion

In this study, we have developed a machine learning system for pet scoring utilizing ensemble learning techniques. The innovative aspects of our system are twofold: firstly, the adoption of ensemble learning methodologies to improve predictive accuracy; secondly, the implementation of a three-stage training process that incrementally introduces data, thereby enhancing the system's robustness. This phased data integration approach has proven to be effective in bolstering the system's performance in complex computer vision tasks. Our system achieved a score of 16.83358, indicating its efficacy in the nuanced domain of pet scoring. This achievement underscores the potential of ensemble learning

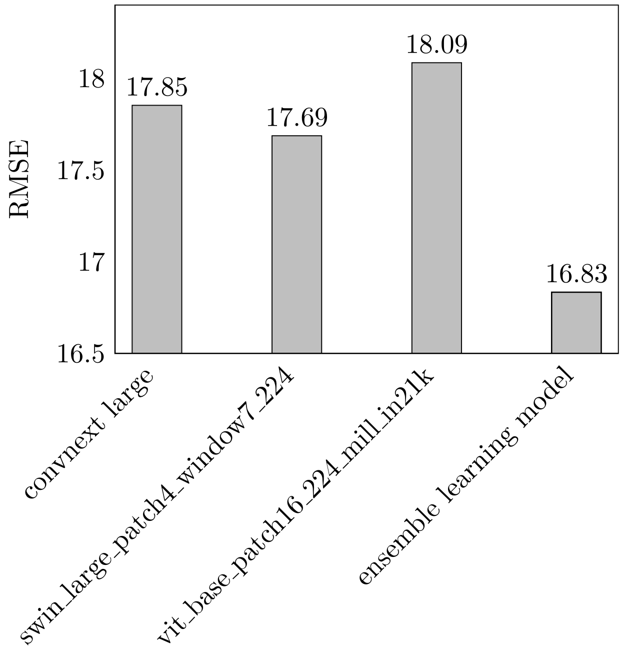

Figure 2: RMSE Comparison between Ensemble Learning Model andSOTA Solution Model

and staged data incorporation in advancing the robustness and accuracy of computer vision systems.

References

Bau, D.; Zhu, J.-Y.; Strobelt, H.; Lapedriza, A.; Zhou, B.; and Torralba, A. 2020. Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences*, 117(48): 30071–30078.

Baxes, G. A. 1994. *Digital image processing: principles and applications*. John Wiley & Sons, Inc.

Cai, H.; Gan, C.; and Han, S. 2022. Efficientvit: Enhanced linear attention for high-resolution low-computation visual recognition. *arXiv preprint arXiv:2205.14756*.

Chen, C.-F.; Panda, R.; and Fan, Q. 2021. Regionvit: Regional-to-local attention for vision transformers. *arXiv preprint arXiv:2106.02689*.

Chen, L.-C.; Papandreou, G.; Schroff, F.; and Adam, H. 2017. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*.

Chen, S.; Xie, E.; Ge, C.; Chen, R.; Liang, D.; and Luo, P. 2021. Cyclemlp: A mlp-like architecture for dense prediction. *arXiv preprint arXiv:2107.10224*.

Chen, Y.; Dai, X.; Chen, D.; Liu, M.; Dong, X.; Yuan, L.; and Liu, Z. 2022a. Mobile-former: Bridging mobilenet and transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5270–5279.

Chen, Y.; Dai, X.; Liu, M.; Chen, D.; Yuan, L.; and Liu, Z. 2020. Dynamic convolution: Attention over convolution kernels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11030–11039.

Chen, Y.; Liu, J.; Qi, X.; Zhang, X.; Sun, J.; and Jia, J. 2022b. Scaling up kernels in 3d cnns. *arXiv preprint arXiv:2206.10555*.

Chen, Z.; Zhong, F.; Luo, Q.; Zhang, X.; and Zheng, Y. 2022c. EdgeViT: Efficient Visual Modeling for Edge Computing. In *International Conference on Wireless Algorithms, Systems, and Applications*, 393–405. Springer.

Chi, L.; Jiang, B.; and Mu, Y. 2020. Fast fourier convolution. *Advances in Neural Information Processing Systems*, 33: 4479–4488.

Chu, X.; Tian, Z.; Wang, Y.; Zhang, B.; Ren, H.; Wei, X.; Xia, H.; and Shen, C. 2021. Twins: Revisiting the design of spatial attention in vision transformers. *Advances in Neural Information Processing Systems*, 34: 9355–9366.

Dai, J.; Qi, H.; Xiong, Y.; and Li, Y. 2017. GuodongZhang, Han Hu, and Yichen Wei. Deformable convolutionalnetworks. In *ICCV*, volume 1, 4.

Dai, Z.; Liu, H.; Le, Q. V.; and Tan, M. 2021. Coatnet: Marrying convolution and attention for all data sizes. *Advances in neural information processing systems*, 34: 3965–3977.

Ding, X.; Zhang, X.; Han, J.; and Ding, G. 2022. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11963–11975.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Everingham, M.; Eslami, S. A.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2015. The pascal vi-

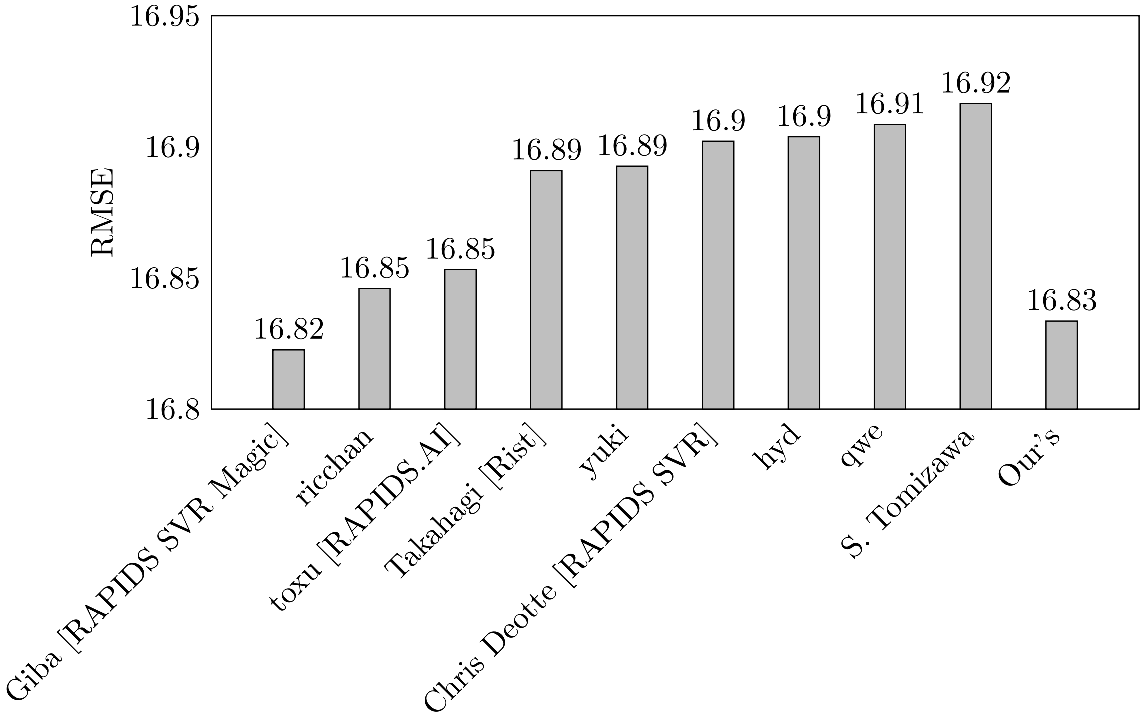

Figure 3: Our Model Compares to TOP-9 Solution

sual object classes challenge: A retrospective. *International journal of computer vision*, 111: 98–136.

Friedman, J. 2001. Greedy function approximation: A gradient boosting machine. *ANNALS OF STATISTICS*, 29(5): 1189–1232.

Ganaie, M. A.; Hu, M.; Malik, A. K.; Tanveer, M.; and Suganthan, P. N. 2022. Ensemble deep learning: A review. *ENGINEERING APPLICATIONS OF ARTIFICIAL INTELLIGENCE*, 115.

Ge, Z.; Liu, S.; Wang, F.; Li, Z.; and Sun, J. 2021. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*.

Graham, B.; El-Nouby, A.; Touvron, H.; Stock, P.; Joulin, A.; Jégou, H.; and Douze, M. 2021. Levit: a vision transformer in convnet's clothing for faster inference. In *Proceedings of the IEEE/CVF international conference on computer vision*, 12259–12269.

Guibas, J.; Mardani, M.; Li, Z.; Tao, A.; Anandkumar, A.; and Catanzaro, B. 2021. Adaptive fourier neural operators: Efficient token mixers for transformers. *arXiv preprint arXiv:2111.13587*.

Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; and Liu, T.-Y. 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In Guyon, I.; Luxburg, U.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 30 (NIPS 2017)*, volume 30 of *Advances in Neural Information Processing Systems*. 31st Annual Conference on Neural Information Processing Systems (NIPS), Long Beach, CA, DEC 04-09, 2017.

Kopitar, L.; Kocbek, P.; Cilar, L.; Sheikh, A.; and Stiglic, G. 2020. Early detection of type 2 diabetes mellitus using machine learning-based prediction models. *SCIENTIFIC REPORTS*, 10(1).

Liu, S.; Chen, T.; Chen, X.; Chen, X.; Xiao, Q.; Wu, B.; Kärkkäinen, T.; Pechenizkiy, M.; Mocanu, D.; and Wang, Z. 2022. More convnets in the 2020s: Scaling up kernels beyond 51x51 using sparsity. *arXiv preprint arXiv:2207.03620*.

Maaz, M.; Shaker, A.; Cholakkal, H.; Khan, S.; Zamir, S. W.; Anwer, R. M.; and Shahbaz Khan, F. 2022. Edgenext: efficiently amalgamated cnn-transformer architecture for mobile vision applications. In *European Conference on Computer Vision*, 3–20. Springer.

Oppenheim, A. V. 1999. *Discrete-time signal processing*. Pearson Education India.

Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A. V.; and Gulin, A. 2018. CatBoost: unbiased boosting with categorical features. In Bengio, S.; Wallach, H.; Larochelle, H.; Grauman, K.; CesaBianchi, N.; and Garnett, R., eds., *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 31 (NIPS 2018)*, volume 31 of *Advances in Neural Information Processing Systems*. 32nd Conference on Neural Information Processing Systems (NIPS), Montreal, CANADA, DEC 02-08, 2018.

Rabiner, L. R.; and Gold, B. 1975. Theory and application of digital signal processing. *Englewood Cliffs: Prentice-Hall*.

Team	**Data**	**Score**	**Solution**
Giba [RAPIDS SVR Magic]	Metadata, Image Data	16.82256	Extracted features using timm and OpenAI CLIP, concatenated embeddings of multiple models, and applied cuml SVR. Ensembled with standard image regression using Vision Transformers and CNN models.
ricchan	Metadata, Image Data, Image Info Data	16.84597	Incorporated data from a previous competition, utilizing a large number of pretrained models.
toxu	Metadata, Image Data	16.85314	Considered the task as a classification, using LightGBM, XGBoost, Swin Transformer (large-sized model), Vision Transformer.
[RAPIDS.AI] Takahagi [Rist]	Metadata, Image Data	16.89096	Ensemble of 8 image models through simple averaging, XGBoost training on pet1 table data, combining predictions with image models. Conditional training and inference based on pet1 and pet2 data for AdoptionSpeed.
yuki	Image Data	16.89257	Three-step approach, creating 14 single models initially. Stacking implemented using BayesianRidge.
Chris Deotte [RAPIDS SVR]	Metadata, Image Data	16.90208	Preprocessing with random square crops, training diverse models with different image sizes and augmentations. Leveraging meta data for multitask learning. Hill climbing approach to steadily build ensembles.
hyd	Not Specified	16.90383	Not Specified
qwe	Not Specified	16.90851	Not Specified
S. Tomizawa	Metadata, Image Data	16.9165	Weighted average ensemble of three types of models: Deep Label Distribution Learning (DLDL) models.
Tim Riggins	Not Specified	16.92795	Not Specified

Table 2: SOTA

Tang, C.; Zhao, Y.; Wang, G.; Luo, C.; Xie, W.; and Zeng, W. 2022. Sparse MLP for image recognition: Is self-attention really necessary? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 2344–2351.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Wei, G.; Zhang, Z.; Lan, C.; Lu, Y.; and Chen, Z. 2023. Active token mixer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 2759–2767.

Wu, Z.; Lischinski, D.; and Shechtman, E. 2021. Stylespace analysis: Disentangled controls for stylegan image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12863–12872.

Zhang, D. J.; Li, K.; Wang, Y.; Chen, Y.; Chandra, S.; Qiao, Y.; Liu, L.; and Shou, M. Z. 2022. Morphmlp: An efficient mlp-like backbone for spatial-temporal representation learning. In *European Conference on Computer Vision*, 230–248. Springer.

Zhang, L.; Zhou, S.; Guan, J.; and Zhang, J. 2021. Accurate few-shot object detection with support-query mutual guidance and hybrid loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14424–14432.

Zhou, B.; Zhao, H.; Puig, X.; Xiao, T.; Fidler, S.; Barriuso, A.; and Torralba, A. 2019. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127: 302–321.

Zhou, J.; Jampani, V.; Pi, Z.; Liu, Q.; and Yang, M.-H. 2021. Decoupled dynamic filter networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6647–6656.