Enhancing Dataset Sufficiency for Attributes Through Text-Driven Generative Data Augmentation

Masatoshi Sekine[1], **Daisuke Shimbara**[1], **Tomoyuki Myojin**[1], **Eri Imatani**[2]

[1]Research & Development Group, Hitachi, Ltd., Japan[1]
[2]Digital Systems & Services Division, Hitachi, Ltd, Japan.[2]
{masatoshi.sekine.ck, daisuke.shimbara.gk, tomoyuki.myojin.fs, eri.imatani.no}@hitachi.com

Abstract

Training datasets for deep learning models, including foundation models, must be both diverse and comprehensive. Therefore, the dataset should be improved to ensure overall mean accuracy and mean accuracy per classification class as well as fine-tuning attributes in the dataset. The accuracy of conventional evaluation methods based on class-wise classification can degrade for attributes other than the class, even if each class achieves high classification accuracy. Therefore, in this study, a novel evaluation metric called attribute-wise classification accuracy was proposed for classification tasks. In this model, an automated data augmentation method that constructs subsets based on individual words in image captions was used to compute and maximize classification accuracy. The proposed method generates captions corresponding to the original image dataset and expresses attribute values in the text format. Furthermore, we introduced generative automated image data augmentation based on text-driven attribute manipulation (GANDAM), an automated data augmentation method that generates interpretable new data by manipulating text. GANDAM manipulates attribute values to ensure sufficient and complete data coverage for attribute values. By learning an optimal policy to manipulate text in a manner that maximizes classification accuracy for each attribute value and maintains the naturalness of the generated text data, GANDAM optimizes data augmentation. Performance evaluation confirmed that the proposed method improved the attribute-wise classification accuracy and its mean.

Introduction

The importance of data quality in AI has been discussed in numerous studies. Data in AI is highlighted as critical, especially in high-risk applications in which predictions can have downstream effects (Sambasivan and Aroyo 2021). Furthermore, rigorous evaluation of data quality is necessary for improving the quality of machine learning (Chen and Ding 2021). Problems associated with data collection and quality in deep learning, where feature engineering is minimal, as well as fairness metrics and bias mitigation techniques have been discussed (Whang and Lee 2023). Furthermore, in natural language processing, the construction of robust and linguistically competent models requires both data curation and algorithmic solutions (Rogers 2021). Machine learning guidelines, such as (of Advanced Industrial Science and AIST) and (of Quality Assurance for Artificial-Intelligence-based products and services 2022) emphasize the importance of confirming dataset coverage for each subdivided domain as well as dataset uniformity to ensure that data are included evenly and without bias across the entire dataset.

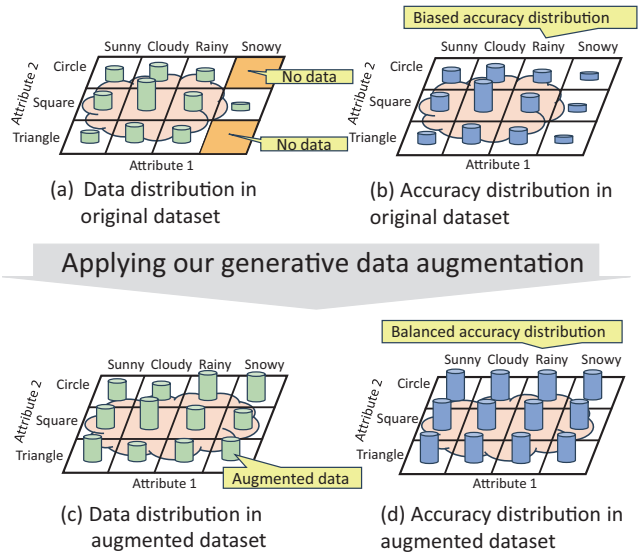

Figure 1: Data augmentation considering precision for multiple attribute values

Generally, the mean accuracy and mean average precision per class (mAP) typically do not accurately reflect the classification accuracy of data with low sample counts or data with attribute values other than class names. Figure 1 illustrates the necessity for data set sufficiency for each attribute value and its combination by using road sign image data used in autonomous driving as an example. The road sign image data includes attributes such as “round,” “square,” and “triangle” for shape, and “sunny,” “cloudy,” “rainy,” and “snowy” for weather conditions. Figure 1(a) depicts the number of samples per attribute combination for shape and weather conditions. These data counts are vari-

Presented at the Workshop on Preparing Good Data for Generative AI: Challenges and Approaches (Good-Data) in conjunction with AAAI 2025.

able, and some combinations, such as "round sign in snowy weather" or "triangular sign in snowy weather" may not exist. This data imbalance could be attributed to the difficulty of obtaining comprehensive data for all attribute combinations. Thus, as depicted in Figure 1(b), classification accuracy could decrease for certain attributes or attribute combinations with low sample counts. This phenomenon highlights the importance of ensuring sufficient data coverage for attributes beyond just the "classification class" attribute.

Generally, insufficient data leads to increased detection uncertainty and unstable model behavior (Gawlikowski 2023). Reference (Bertossi 2020) discusses the relationship between data quality and explainability in AI systems, asserting that improving data quality improves model explainability. By contrast, data typically contains complex interdependencies that can be difficult to explain and are not always suitable for assessing overall data quality. Data enhancement for images includes common transformations such as rotation and cropping as well as techniques such as random erasing, which randomly masks rectangular areas with noise (Zhong and Yang 2020). However, these transformations are performed on a trial-and-error basis without optimization. Automated data augmentation methods proposed include AutoAugment (Cubuk, Mane, and Le 2019) and faster methods (Lim and Kim 2019)(Hataya and Nakayama 2020), as well as GA3N (Chinbat and Bae 2022), in which GAN is used for data augmentation, and GALIP (Tao and Xu 2023) in which GAN with CLIP are combined (Radford and et. al. 2021) for fast processing. Text AutoAugment (Ren and Zhou 2022) was proposed as an automatic data augmentation method for text data by using sequential model-based global optimization (SMBO) for optimal policy search. However, they do not consider attribute-wise accuracy improvement for text or nontext data, similar to GA3N.

In data augmentation research focused on improving dataset quality, methods have been devised to improve dataset quality by generating images from text. For example, ALIA (Dunlap. and Darrell 2023) incorporates diffusion models to increase dataset diversity by automating image processing with natural language guidance, especially to address data scarcity in fine-grained classification tasks (e.g., animal classification). Prompt augmentation (PA) (Bodur and Kim 2024) introduces a technique called "prompt augmentation" for text-guided image processing, generating multiple target prompts from a single-input prompt to perform precise image processing. Furthermore, contrast loss is applied to distinguish between edited areas and preserved areas in the original image. Although PA is focused on high-precision editing of specific areas of an image, it does not optimize for dataset uniformity and coverage through attribute-based image generation.

In addition to discrete attributes such as words, continuous attributes such as thickness or shape in handwriting should be considered. Multi-facet clustering variational autoencoders (MFCVAE) (Falck 2021) allows the extraction of latent variables from multiple perspectives. A dataset quality evaluation method using MFCVAE, and a visualization tool for obtaining various attribute information in datasets, such as thickness and shape in handwritten characters (Sekine and Imatani 2023) have been proposed for dataset quality evaluation.

In this study, we proposed a novel method to improve dataset quality in classification tasks by extracting and manipulating attribute information from the original dataset for automatic data augmentation, while considering the improvement of classification accuracy for each attribute. By applying GANDAM, the imbalance in the number of samples per attribute is mitigated, as depicted in Figure 1(c), and the coverage of representative samples is improved, resulting in improved attribute-wise accuracy, as depicted in Figure 1(d). Attributes are extracted using methods that generate text data describing the data, and handling them in text format allows flexible manipulation and generating data faithful to specified attribute values. By evaluating the naturalness of the text data, GANDAM ensures the validity of the data generated by the policy. Furthermore, we introduce word-wise classification accuracy as a novel evaluation metric and demonstrated the improvement of attribute-wise classification accuracy.

Section 2 discusses related research on automated data augmentation, Section 3 details the proposed text-driven generative automated data augmentation based on attribute manipulation, and Section 4 presents the conclusion and future directions.

GANDAM

Overview

In GANDAM, the original image data is converted into text (caption) describing the image. To generate image data with maximized classification accuracy for each combination of attribute values, the policy optimization model learns to manipulate attribute values in the text domain and generates corresponding image datasets from text generated according to the policy. Furthermore, the naturalness of the text data generated based on the policy is evaluated, and the policy optimization model is trained to achieve high evaluation scores.

GANDAM consists of the following five features.

Function 1: Generation and Part-of-Speech Tagging of Text Data Corresponding to Original Image Data

In Function 1 of Figure 2, attribute information is first extracted from images by generating text data describing the images by using image-to-text conversion.

Function 2: Generating Image Data from Text Data for Augmented Data and Reassigning Label

In Function 2 of Figure 2, image data corresponding to the text data for augmented image data generation created based on the policy output by the policy optimization model (explained in function 5) is generated. In the generated image, the class name can change if the original class name word is replaced with another word. Therefore, a text classification model is created to estimate the class name from the caption

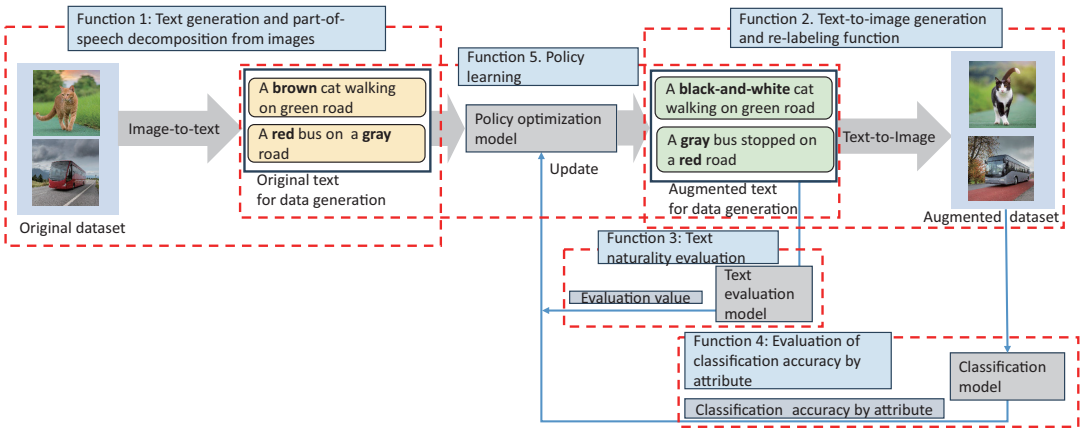

Figure 2: Functions and configuration of GANDAM

corresponding to the new image data and to assign labels to the generated data.

Function 3: Naturalness Evaluation of Text Data for Augmented Data Generation

In function 3 of figure 2, during the training process of the policy optimization model, the naturalness evaluation of the generated text data for augmented data generation based on the policy is performed and evaluation scores are obtained. Here, naturalness refers to the contextual appropriateness, grammatical correctness, and realism of the generated text data. For example, the text data "a person walking" would receive a high evaluation score, while "a car walking" would receive a low evaluation score. The naturalness of text is evaluated using the similarity method BERTScore (Zhang and Artzi 2020).

Function 4: Attribute-Wise Accuracy Evaluation of the Classification Model Using Generated Image Data

In Function 4 of Figure 2, the classification model trained on the generated image dataset is used to infer the class on a validation image dataset. The attribute value combinations, as presented in the example in Figure 2, include the accuracy per attribute for character shapes, as well as per weather condition, and the attribute-wise accuracy per combination of shape and weather condition. We computed the minimum attribute-wise accuracy and adjust the model to mitigate bias. In this study, the word-wise classification accuracy for words in captions corresponding to the inference data images is calculated, with the mean serving as the evaluation score. This evaluation metric was defined as mean accuracy by attribute value (MAAV), where AAV denotes accuracy by attribute value. Let X be the set of word types (excluding articles) appearing in the captions of all image datasets. For a given word x (where x is an element of X, e.g. "bird," "zoo," "sky," etc.), let $N_{x,all}$ be the number of all datasets that include the word x in their captions, and let $N_{x,correct}$ be the number of correct data sets among them. Further, let $|X|$ be the number of elements in the set X , that is, the total number of word types in all the captions. Then for any given x, AAV and MAAV are expressed as follows.

$$AAV(x) = \frac{N_{x,correct}}{N_{x,all}} \quad (1)$$

$$MAAV = \frac{1}{|X|}\sum_{x\in X} AAV(x) \quad (2)$$

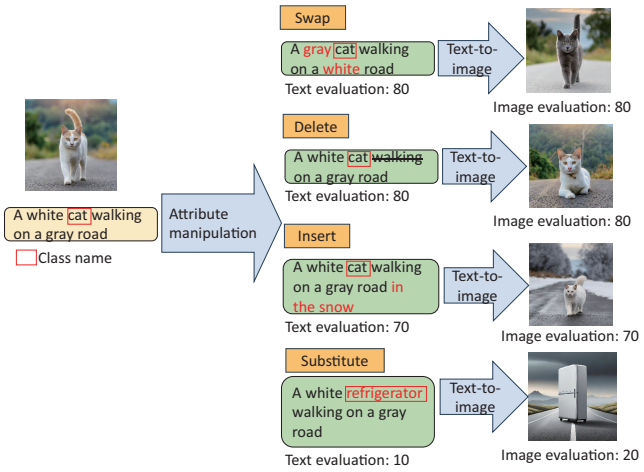

Figure 3: Manipulation of attribute information in text domain

Function 5: Policy Optimization Based on Text-Driven Attribute Manipulation

In Function 5 of Figure 2, the policy is optimized to generate text data for the augmented data by manipulating the original text data set. Sequential model-based global optimization (SMBO), a type of Bayesian optimization, is used to optimize the policy. First, the text data corresponding to the original image data extracted in function 1 is input to the policy optimization model. The policy optimization model outputs a policy consisting of probabilities for performing each operation such as "swap,""delete," "insert" or "replace" on the original text data. These operations are performed based on the probabilities in the policy.

Figure 3 illustrates the manipulation of attribute information in the text domain. For example, assume the original

text data is “A white cat walking on a gray road” with “cat” as the classification class. The words can undergo operations such as “swap,” “delete,” “insert,” or “replace.”

Based on the policy, text data are created for data generation, and extended data are generated according to Function 3. Naturalness scores for text data and attribute-wise classification accuracy of enhanced data are obtained in Functions 3 and 4, respectively. If the generated text data is unnatural or results in low classification accuracy, then the policy is less likely to be issued.

The objective function of the policy optimization model is set to the weighted sum of the naturalness score of the generated text data and the classification accuracy of the data. By training the policy optimization model to maximize the objective function, the naturalness of the generated text and the attribute-wise classification accuracy are maximized. The objective function is represented by the MAAV and naturalness score TN, with weights k_1 and k_2.

$$J = k_1 \cdot MAAV + k_2 \cdot TN \tag{3}$$

Experiments

To evaluate the performance, the following experiments were performed. First, we investigated how the number of training data per attribute affects accuracy. Next, we compared the proposed method with Faster AutoAugment. We used the COCO dataset (Lin, Dollár, and Zitnick 2014) as the original training and inference datasets, with 100 images in ten classes for training data and 1000 images in ten classes for inference data. All image files were set to 160 pixels $times$ 160 pixels.

Naturalness Evaluation of Text

The naturalness of generated text for image data was evaluated based on the average similarity to captions in the COCO dataset. For example, the natural sentence “a man is walking’ has a score of 0.870, whereas “a car is walking” has a score of 0.847, indicating that the latter could be less natural. Table 1 shows the naturalness scores and examples of generated text. The top 5 scores indicate plausible, natural sentences with corresponding scores.

Item	Score	Generated Text for Data Augmentation
Best 5	0.944	a close up of an elephant walking in the grass
	0.924	a young boy looking at a cow in a pen
	0.920	a herd of animals grazing on a dry grass field
	0.919	two giraffes in a cage
	0.915	a clock tower in the middle of a busy street
Worst 5	0.827	two zebras grazing on grass in an enclosure
	0.836	a silver metal toilet seat on the floor
	0.851	a black yak standing in a grassy field
	0.852	a white toilet in a small bathroom stall
	0.852	a train sitting on the tracks next to a water tower

Table 1: Naturalness evaluation scores and examples of generated text for data augmentation

Attribute-Wise Accuracy Evaluation

For the attribute-wise classification accuracy, we calculated the average accuracy for each word contained in the caption corresponding to the validation data image. We excluded words with fewer occurrences than the threshold to ensure reliability and set the threshold to 10. The weights for the text naturalness factor k_1 and the attribute-specific accuracy factor k_2 in the objective function were each set to 0.5.

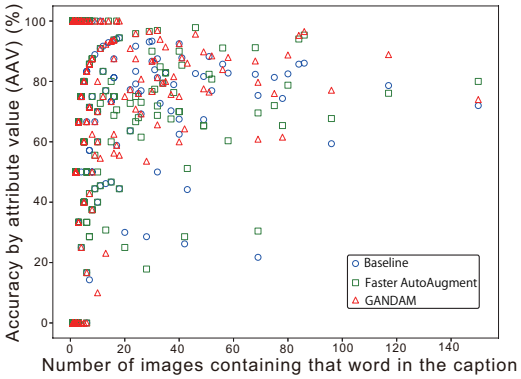

Figure 4: Relation between the number of images containing that word in the captionword and Accuracy by Attribute Value (AAV)

Baseline	Faster AutoAugment	GANDAM
74.2	74.9	78.6

Table 2: Mean Accuracy by Attribute Value (MAAV)(%)

Table 2 details the MAAV. GANDAM outperformed both the original dataset (baseline) and Faster AutoAugment by 4.5% and 3.7%, respectively. This improvement could be attributed to GANDAM’s inclusion of attribute-wise classification accuracy in the objective function, which improves data augmentation to maximize this metric.

Figure 4 depicts the relationship between the number of validation data per word and the classification accuracy for the baseline without data augmentation, Faster AutoAugment, and the proposed method GANDAM. The proposed method exhibited improved accuracy.

Conclusion and Future Work

In this study, we proposed a novel method to improve the quality of datasets by manipulating various attributes in the text domain, transforming them back into dataset format, and training a policy optimization model to maximize attribute-wise classification accuracy. GANDAM can address cases in which preparing sufficient data is difficult for each combination of attribute values, such as when data acquisition costs are high, by correcting imbalances in the number of samples per attribute combination.

In future, we plan to extend the capabilities of the model from single-class classification to multi-class and multimodal classifications. In addition, to reduce the computational cost of learning, the amount of data must be reduced while simultaneously maintaining the accuracy of classfication. Furthermore, a key challenge in developing explainable AI models is the automatic extraction of attributes and attribute values that influence accuracy.

References

Bertossi, F., L.;and Geerts. 2020. Data Quality and Explainable AI Journal of Data and Information. *Journal of Data and Information Quality (JDIQ)*, 12.

Bodur, B., R.; Bhattarai; and Kim, T.-K. 2024. Prompt Augmentation for Self-supervised Text-guided Image Manipulation. In *CVPR*.

Chen, J., H.; Chen; and Ding, J. 2021. Data Evaluation and Enhancement for Quality Improvement of machine learning.

Chinbat, V.; and Bae, S.-H. 2022. GA3N: Generative adversarial AutoAugment network. *Pattern Recognition*, 127.

Cubuk, B., E.D.; Zoph; Mane, V., D.;Vasudevan; and Le, Q. V. 2019. AutoAugment: Learning Augmentation Policies from Data. In *CVPR*.

Dunlap., A. Z. H. Y. J. G. J., L.; Umino; and Darrell, T. 2023. Diversify Your Vision Datasets with Automatic Diffusion-Based Augmentation. In *NeurIPS*.

Falck, H. W. M. N. G. Y. C. H. C., F.; Zhang. 2021. Multi-Facet Clustering Variational Autoencoders. In *NeurIPS*.

Gawlikowski, J. e. a. 2023. A Survey of Uncertainty in Deep Neural Networks.

Hataya, Z. K., R.; Jan; and Nakayama, H. 2020. Faster AutoAugment: Learning Augmentation Strategies Using Backpropagation. In *ECCV*.

Lim, I. K. T. C., S.; Kim; and Kim, S. 2019. Fast AutoAugment. In *NeurIPS*.

Lin, M. B. S. H. J. P. P. R. D., T-Y.; Maire; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

of Advanced Industrial Science, A. I. R. C. N. I.; and (AIST), T. 2021. Machine Learning Quality Management Guideline.

of Quality Assurance for Artificial-Intelligence-based products, C.; and services. 2022. Guidelines for Quality Assurance of AI-based Products and Services.

Radford, A.; and et. al. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*.

Ren, J. L. S.-X., S.; Zhang; and Zhou, J. 2022. Text AutoAugment: Learning Composition Augmentation Policy for Text Classification. In *EMNLP*.

Rogers, R. 2021. Changing the World by Changing the Cata. In *Annual Meeting of the Association for Computational Linguistics*.

Sambasivan, S. H. H. A.-D. P. P., N.; Kapania; and Aroyo, L. M. 2021. Everyone wants to do the model work, not the data work. In *CHI Conference on Human Factors in Computing Systems*.

Sekine, D. M. T., M.; Shimbara; and Imatani, E. 2023. Visualization Tool for Extraction of Various Attributes and Corresponding Data for Dataset Quality Assessment. In *IEEE International Conference On Artificial Intelligence Testing (AITest)*.

Tao, B.-K. T. H., M.; Bao; and Xu, C. 2023. GALIP: Generative Adversarial CLIPs for Text-to-Image Synthesis. In *CVPR*.

Whang, Y. S.-H., Y. S. E.; Roh; and Lee, J.-G. 2023. Data Collection and Quality Challenges in Deep Learning: a Data-Centric AI Perspective.

Zhang, V. W. F. W.-K., T.; Kishore; and Artzi, Y. 2020. BERTScore: Evaluating Text Generation with BERT. In *ICLR*.

Zhong, L. K. G. L. S., Z.; Zheng; and Yang, Y. 2020. Random Erasing Data Augmentation. In *AAAI*.

Appendix

Pseudocode of our proposed data augmentation (GANDOM)

The pseudocode for the operation is shown in Algorithm 1. The relation to functions 1 to 5 is indicated by ”func 1”, ”func 2” and so on.

Algorithm 1: GANDAM: Generative Automated Data Augmentation using Text-Driven Attribute Manipulation

Input: Original image dataset X_{img}
Output: Augmented image dataset X_{gen}, Trained classification model M_{trained}
Step 1: (func 1) Convert images to text (captions)
$T_{\text{caption}} \leftarrow f_{\text{caption}}(X_{\text{img}})$
Step 2: (func 2) Initialize policy model P
$P_{\text{init}} \leftarrow$ Initialize policy model
Step 3: (func 3, func 4, func 5) Optimize policy model based on naturalness and attribute accuracy using SMBO
for each training iteration **do**
    Generate text using current policy P
    $T_{\text{gen}} \leftarrow$ Generate text based on P_{init}
    Evaluate naturalness TN based on reference texts ▷ **func 3**
    $TN \leftarrow \frac{1}{N}\sum_{i=1}^{N} \text{Sim}(T_{\text{gen}}, T_{\text{ref}_i})$
    Evaluate attribute accuracy $MAAV$ ▷ **func 4**
    $MAAV \leftarrow$ Calculate mean accuracy by attribute value
    Update policy P to maximize $J = k_1 \cdot MAAV + k_2 \cdot TN$ ▷ **func 5**
end for
Step 4: (func 2) Generate augmented image dataset based on optimized policy P_{opt}
$X_{\text{gen}} \leftarrow g_{\text{policy}}(T_{\text{caption}}, P_{\text{opt}})$
Step 5: (func 4) Train classification model using augmented dataset
$M_{\text{trained}} \leftarrow$ Train model using X_{gen}
Return: Augmented image dataset X_{gen}, Trained classification model M_{trained}

Figure 5 shows a bar graph between the accuracy for each word and attribute value shown in Figure **??**.

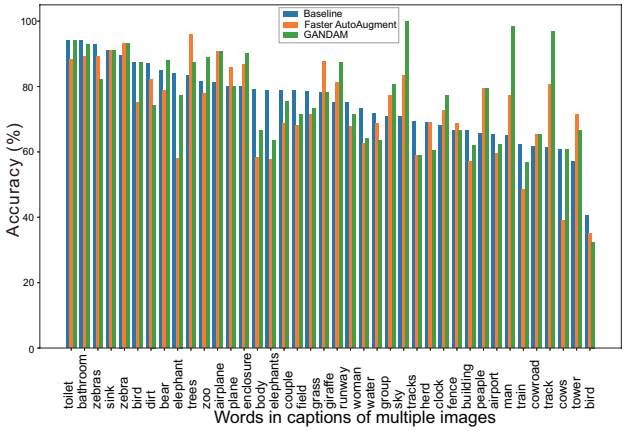

Figure 5: Accuracy by attribute aalue (AAV)