

Real or really fake? Synthetic vs. genuine image classification

Jacob Suchardt

LT2318 HT24 Artificial Intelligence: Cognitive Systems
MLT, FLov, University of Gothenburg, Sweden
gussucju@student.gu.se

Abstract

This project aims to explore the task of detecting AI-generate images using CNNs and an SRM channel attention layer trained on the CIFAKE dataset. This dataset is explored and reviewed in terms of the presented task's nature, difficulty, and in particular the suitability of its data and the features inherent therein as a training set for models intended to be deployed in downstream tasks on varied data. This transfer suitability and general model performance is additionally examined extrinsically by application to the test sets of CIFAR-100 and AGIRI.¹

1. Introduction

With the rise of artificial intelligence (AI), so has its presence in spheres of cultural and social interaction. Especially on social media, we see users being outnumbered by 'bots', AI-based accounts, sharing AI-generated text, image and video contents. Discerning synthetic from genuine data is becoming a key skill in navigating online spaces. Its importance is exacerbated by recent decisions of Meta to discontinue third-party fact checking of contents on their online platforms (Facebook, Instagram), instead leaving this role up to users themselves – similar to the current guidelines of X (form. Twitter). Yet, the struggle of many users to acquire this skill against growing technical advances has grave implications, the risk of misinformation and deception increasing.

The coherence and quality of synthetic image data in particular has improve greatly since the first freely available services were released. They now achieve, to the human eye, near photographic realism. It is this category of hyper-realistic generated images that this project focuses on. When speaking about recipients' skills and associated risks, we must consider not only the ability to weed out fake images consciously, i.e. when explicitly trying to ascertain the nature of an image, but also their instinctive ability: Whether users, especially those with lesser familiarity with synthetic images, are more likely to accept lower quality generated images when not on the



Figure 1. Samples from Bird & Lotfi's (2024) demonstration of AI-generated images in CIFAKE.

active look-out for them, e.g. during daily activities, scrolling by.

For instance, recipients for the most part passively accepted the images in figure 1, but when prompted to view the image for longer, noted mangled writing on the number plate and strange textures by the car roof (left), as well as 'nonsensical branches' and discrepancies in object styles ('cartoonish' bird vs. 'realistic foliage').

Our aim is to explore the usability of machine learning methods in this area, by training a model to assess the nature of an image as being AI-generated/synthetic ('fake') or genuine ('real'), and label it as such. Further, the generalisational ability and thus applicability of a model trained on one specific dataset ought to be examined to gain insights into its usability in a real-world context and the difficulty of the task at hand.

The selected datasets for real and fake images are described in section 2. The section thereafter details the model architecture and modifications thereof, training parameters, and the experiment setup used to test model performance (section 3). In Section 4, the results of these are reported, analysed, and subsequently interpreted in section 5. Finally, section 6 closes with a conclusion and notes regarding future work.

2. Data

The main dataset underlying this project is CIFAKE (Bird, 2023), described in Bird & Lotfi (2024). CIFAKE comprises of a total of 120k

¹ <https://github.com/jshrdt/aics-project>

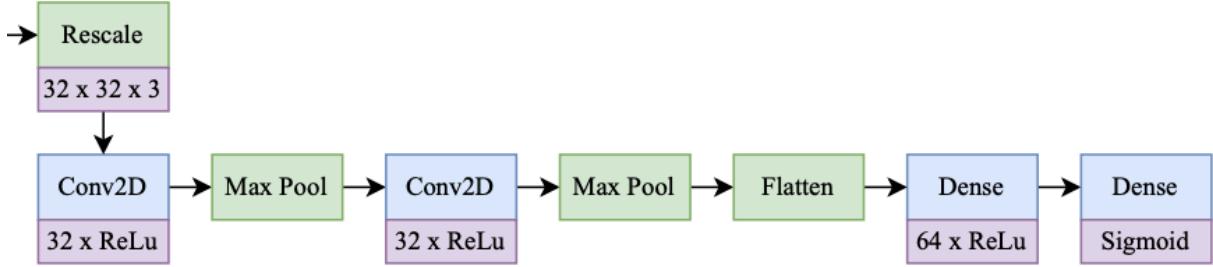


Figure 2. For reference: Architecture and parameters with 'best performance' in Bird & Lotfi (2024, p.9 fig.5).

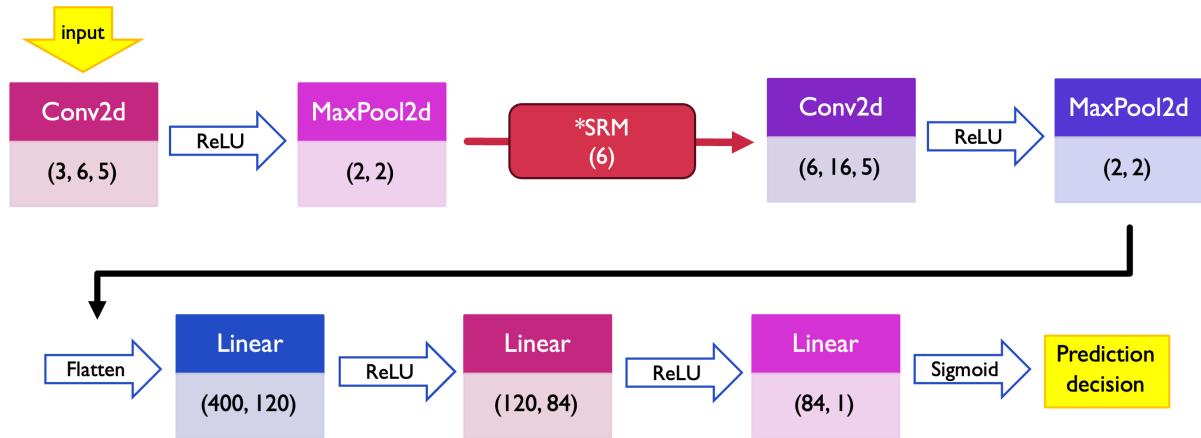


Figure 3. Base model architecture implemented in this project, modified from cifar10_tutorial.ipynb. SRM layer only included in the attention model.

images at a resolution of 32x32, half (train: 50k, test: 10k) of which are genuine, and the other half having been generated by Lotfi & Bird (2024) for use in the dataset. The 'real' images are the entirety of the CIFAR-10 dataset (Krizhevsky, 2009; available at <https://www.cs.toronto.edu/~kriz/cifar.html>) which can be separated into ten mutually exclusive content categories (airplane, automobile, bird, cat, deer, dog, frog, horse, ship and truck), each being represented by 6k sample images. The 'fake' images were generated using CompVis Stable Diffusion v1.4 (<https://huggingface.co/CompVis/stable-diffusion-v1-4>). Bird & Lotfi (2024) express the intent to create a dataset similar to but more diverse than CIFAR-10. This was done by designing the prompts for image generation to be prefaced by "a photograph of a(n) {content label}" in which the content label referred to CIFAR-10's class labels or synonyms thereof. In addition, a pre- or post-nominal modifier could be included to vary the generated outputs' focal object or surrounding (e.g. *fire truck*, *bird flying*). Images were generated with a resolution of 512px and then downsampled to 32px using bilinear interpolation in order to match the resolution of the CI-

FAR-10 images. Moreover, heuristics were employed to denoise the synthetic images.

To test how suitable models trained on CIFAKE are for application we additionally include the test datasets of CIFAR-100 (<https://www.cs.toronto.edu/~kriz/cifar.html>) and 'AGIRI' (*ai-generated-images-vs-real-images*; Zhang, 2024). CIFAR-100 closely resembles CIFAR10, but consists of entirely new content categories, differentiated into 20 superclasses and 100 fine-grained classes. The test set contains 10k real photographs (scaled to 32px), 100 per fine-grained class, spread equally across superclasses. There is no overlap of categories or images between CIFAR-100 and CIFAR10, and thus our CIFAKE training data.

Lastly, we include evaluation on AGIRI which includes 12k images in the test set, one half being genuine, the other being synthetic. AGIRI presents the most challenging dataset for CIFAKE models, as images diverge in content as well as style: They are neither restricted to the CIFAR10 content classes, nor required to feature one focal object/animal at all and are often non-photographic in style. Images were also generated using various resources, i.e. of the total 60k images in AGIRI, 10k were generated by Stable

Diffusion, Midjourney, and DALL-E each, while the real images stem from Pexels, Unsplash (22.5k) and WikiArt (7.5k). The datacard makes no mention of the split depending on the image origin, a balanced split is assumed.

3. Methods

This section presents the implemented model architecture, code resources and training process (3.1), as well as the intended research areas and experiments (3.2).

3.1. Models

The task to be learned on CIFAKE is that of a binary image classification with the classes 'real' (1) and 'fake' (0). In accordance with the original paper (fig. 2), our architecture builds on a CNN with MaxPooling and fully connected layers (fig. 3) using PyTorch (<https://pytorch.org/>).² We employ ReLU as the primary activation function and the Sigmoid function to scale the final output to make a binary prediction decision.³

This project also constitutes an extension of the original Bird & Lotfi (2024) paper, as the base network structure is later expanded upon with Style-based Recalibration Module (SRM) channel attention (Lee et al., 2019), according to the implementation by Misra (2021). The SRM layer is added after the first pooling layer, just before the second convolutional layer.⁴ We presuppose that a key feature in distinguishing genuine from synthetic photographic images is their texture, i.e. a content-independent low-level style feature. Thus, we aim to test how SRM, which according to Lee et al. (2019) enhances a CNN's ability to capture style by recalibrating input feature maps, affects our architecture. If denoising of the images generated for CIFAKE was successful, SRM should have little to no effect on performance. If SRM leads to an improvement, however, low-level style features may still be present across CIFAKE samples.

Both models are trained on the CIFAKE training set (50k real, 50k fake) for ten epochs with a batch size of 32. The order of the images is randomised once at the beginning of the batching process (seeded). During training, the sequence of batches as well as of the images within each batch is shuffled once on each epoch. Image pix-

Gold/ prediction	Real image	Synthetic image	
Real	True positive (TP)	False positive (FP)	Precision $\frac{TP}{TP + FP}$
Fake	False negative (FN)	True negative (TN)	
	Recall $\frac{TP}{TP + FN}$		Accuracy $\frac{TP + TN}{TP + FP + TN + FN}$

Table 1. Combination of gold and predicted labels into categories of true/false positives/negatives along with common evaluation metrics.

el values are initially scaled to a range of (0, 1) and then normalised with a mean of 0.5 across all dimensions. The models are optimised using Binary Cross-Entropy Loss (BCE) and Stochastic Gradient Descent (SGD) with a momentum of 0.9 and a learn rate of 0.001. As an indicator of the learning progress, besides the total loss per epoch, the average loss per batch was also logged at equidistant time steps four times per epoch.

3.2. Experiments

The models are evaluated on various datasets in order to gain insights into their performance and ability to generalise to unseen and dissimilar data. First, we evaluate the models **intrinsically** on the CIFAKE test set (10k real, 10k fake) to ensure that models have learnt the intended task of this dataset. Using the gold and predicted labels (fig. 1) of the test set, we assess their overall performance via their accuracy and F1-Scores ($2 \times \frac{precision \times recall}{precision + recall}$). Furthermore, we quantify the reliability of 'real' and 'fake' classifications using precision and recall, respectively. Evaluation measures are generally created with the default decision threshold for classification as 'real' being $p \geq 0.5$. Visualisations of the effect of this threshold size on the evaluation metrics are also created, to find performance peaks and to assess the diversity of the dataset in terms of classification difficulty. High scores in balanced metrics (e.g. F1-Score) with extreme threshold values would suggest a high confidence of the model in its prediction, thus a possibly low difficulty. Due to the similarities between Bird & Lotfi's (2024)

² Due to the scope of this project, and the lack of reference regarding the number of training epochs or batch sizes in Bird & Lotfi (2024), we attempt neither an exact replication of the model created in figure 2, nor a parameter search for our model.

³ The code base is taken from the course's 'cifar10_tutorial.ipynb' and modified for a binary classification task by altering the loss functions, final linear layer's output dimension, and lastly applying the Sigmoid function.

⁴ No explicit information on where to place the SRM layer in a CNN were given. Comparing the performance of a CIFAKE model with SRM placed just before the second convolutional layer against one with SRM implemented just before flattening revealed slightly better performance for the former (F1: 92,07% vs. 91,81%).

architecture and that of our base model (fig. 3), we take similar performance on CIFAKE as an indication that a similar task and features have been learned by the models.

In addition, we also trained a small version of the base model for only one single epoch on the CIFAKE training data. To monitor the training progress, we section off 10% of the test data as a validation set. During training, every four batches the averaged loss per batch on the training data, validation data, as well as accuracy on the validation data is logged. These values and the mini model's performance on CIFAKE's test set are expected to shed light on the inherent difficulty of the task presented within the dataset: If learning and performance progress extremely quickly, it may be that a majority of the samples in CIFAKE are trivial in such a way that we must assume their classification does not present a learnable challenge. In this case, we must lower our expectations for the transfer task accordingly. For example, it might be possible that the specific image generation process devised for CIFAKE marked many images with a specific fingerprint that is easily detectable for the model but renders them non-representative instances of synthetic images in general.

Second, to examine whether, if present, the task of real vs. fake image detection was learned on CIFAKE, the models are tested on the CIFAR-100 test set (10k real images). If the system predicts an image to be real, we treat it as though it had correctly predicted the content labels. We insert a dummy content label of -1 to assign in case of a 'fake' prediction. By calculating recall for each of the coarse and fine content class labels in CIFAR-100, we may gain insights into whether CIFAKE implicitly encoded content bias (e.g. say the colour purple or clouds were only present in the AI-generated images they would be salient features in the dataset, but detrimental for its ability to generalise as intended). This also allows for a closer look at how often the models produce false negatives, i.e. incorrectly assert an image to be synthetic.

Lastly, the models are also evaluated on AGIRI's (Zhang, 2024) test set (6k real, 6k fake from mixed sources), down-scaled to match CIFAKE's 32px resolution. This constitutes in no way a 'fair' evaluation for the models trained on CIFAKE which not only consists of images at a very low resolution, but also only includes photographs and imitations thereof with very specific contents (e.g. lack of humans, landscape) and styles (e.g. no illustrations, digital 'real' images).

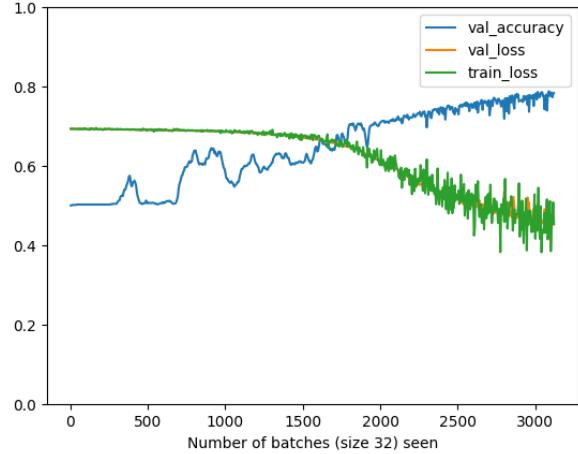


Figure 4. Learning progress of base architecture within the first epoch.

Nevertheless, Bird & Lotfi (2024, p. 8) postulate that:

“The CIFAKE dataset provides the research community with a valuable resource for future work on the social problems faced by AI-generated imagery. The dataset provides a significant expansion of the resource availability for the development and testing of applied computer vision approaches to this problem.”

It is therefore the secondary research goal of this project to test whether the task learnt in CIFAKE is transferable to the wide variety of both 'real' and 'fake' image data that recipients may face in the wild.⁵ If performance on CIFAKE does not correlate with performance on more general data, the suitability of the dataset as training data for downstream tasks may be highly limited.

The mini model is applied to these two datasets as well, in order to establish a baseline which would be a point of reference for the base and attention models that can be used to eliminate doubts about overfitting. Ultimately, should CIFAKE allow for the fully trained models to learn generalisable features of real and synthetic images, they must necessarily perform better than the mini model on the external datasets. Worse performance on the other hand, could also be a sign of overfitting.

4. Results

This section reports on the results of the described experiments on CIFAKE (4.1), CIFAR-100 (4.2) and AGIRI (4.2), as well as their implications for the research areas.

⁵ We leave the philosophical discussion of what constitutes a 'real' image as opposed to a 'fake' up to the reader's discretion and merely distinguish between whether an image is indicated as having been 'generated entirely by AI', or not.

	Acc.	F1	Prec.	Rec.	Loss	
					first	last
Base	91,70	Real	94,02	89,05	2016,3	661,0
		Fake	89,60	94,34		
		ø	91,81	91,7	0,645	0,212
SRM	92,06	Real	92,15	91,96	2048,8	667,7
		Fake	91,97	92,17		
		ø	92,07	92,07	0,656	0,214
Mini	76,57	Real	70,93	90,02	-	-
		Fake	86,35	63,11		
		ø	78,64	76,57	-	0,455

Table 2. Performance of models on CIFAKE (n=20k).

4.1. CIFAKE

The learning and training progress within the base architecture's first epoch is visualised in figure 4. Average batch loss and accuracy conveniently fall within a similar range for plotting but are not directly comparable. The average batch losses, initially around 0,65, decline only marginally until about 1750 batches (56k images) have been seen, at which point the slope intersects with the validation accuracy and begins declining more steeply. Similarly, validation accuracy does not reliably rise above the 50% baseline until c.a. 700 batches (22.4k images). From that point onwards until the intersection with the average batch loss, the learning progress is inconsistent but overall fast, rising by nearly 20%. After the intersection point, validation accuracy rises more slowly, yet also more steadily. We take this as an indication that while a number of salient features for this dataset appear to be easily extractable, the task presented within the dataset is overall not trivial. The perceived difficulty of distinguishing real from fake images appears to vary within this dataset, as it seems to be the case from the few sample images presented in the resource paper, but this is not necessarily a point of critique, since this characteristic theoretically applies to real world data as well.

Once its training is complete, the mini model performs far above chance (accuracy 76,57%) and quite below the fully trained models still (table 2). Most notable are the low fake-recall (63,11%) and comparatively high real-recall (90,02%). This combination is to be expected, due to the fact that the task is still being learnt and that the model's default prediction at the beginning of training is to classify any sample as being a real image.

The fully trained base model on the other hand, even without any tuning of (hyper-) parameters or variation of the architecture, reaches an accuracy score of 91,7% and an F1-score of 91,47%. The SRM-attention model slightly improves on these by +0,36% (92,06%) and +0,6% (92,07%), respectively. The accuracies of these naively designed and trained models do not fall far behind the maximum accuracy of 92,98% score reported by Bird & Lotfi (2024, p.8) who employed a parameter search across 36 different model configurations (p. 5). We note again the lack of information regarding the number of epochs for which training was run and report that training the base model for 50 epochs lead to an accuracy of 91,98% (+0,28%). It is thus not possible to determine whether their scores are due to a more suited model architecture and parameters, or whether the model was trained to overfit on the particularities of the CIFAKE dataset – a hypothesis to which we return in 4.3.

The per-class measures reveal that the attention model is even more balanced between the classes than the base model. Whereas the latter has a slight tendency to over-predict fake images (precision on fake: 89,60%, recall on real: 89,05%), the attention model sees both these measures rise above 90% as well (91,97%, 91,96%), albeit for slightly lower scores on real-precision and fake-recall. From this we may gather that the addition of style-based channel attention strengthens the reliability of a 'fake'-prediction slightly. However, this is accompanied by some synthetic images, which presumably lack this noise element, being mislabelled as 'real', thus decreasing precision on in favour of recall in the 'real' class. It is therefore probable, that the denoising attempt made in the generation of the synthetic images could be improved upon, or, as the gradient class activation maps (Grad-CAM) in Bird & Lotfi (2024) show, that background of synthetic images tend to be of poorer quality in terms of emulating a photographic style.

Finally, the variation of the decision threshold (default $p \geq 0,5$) used to transform the models' predictions, saw the patterns in figure 5 emerge. Intersection points of evaluation metrics lie around 0,5 ($\pm 0,1$) for the fully trained models.⁶ For these, performance is still fairly high (>75%) even at extreme threshold values. In accordance with the possible conclusions regarding the diversity of sample difficulty from figure 4, this too appears to indicate that for a sizeable amount of the data, classification is made with a high

⁶ Performance peaks of the base model are observed for $p \geq 0,4$ with a F1-Score of 91,90%. Most commonly across multiple training runs, however, peaks aligned with the default threshold.

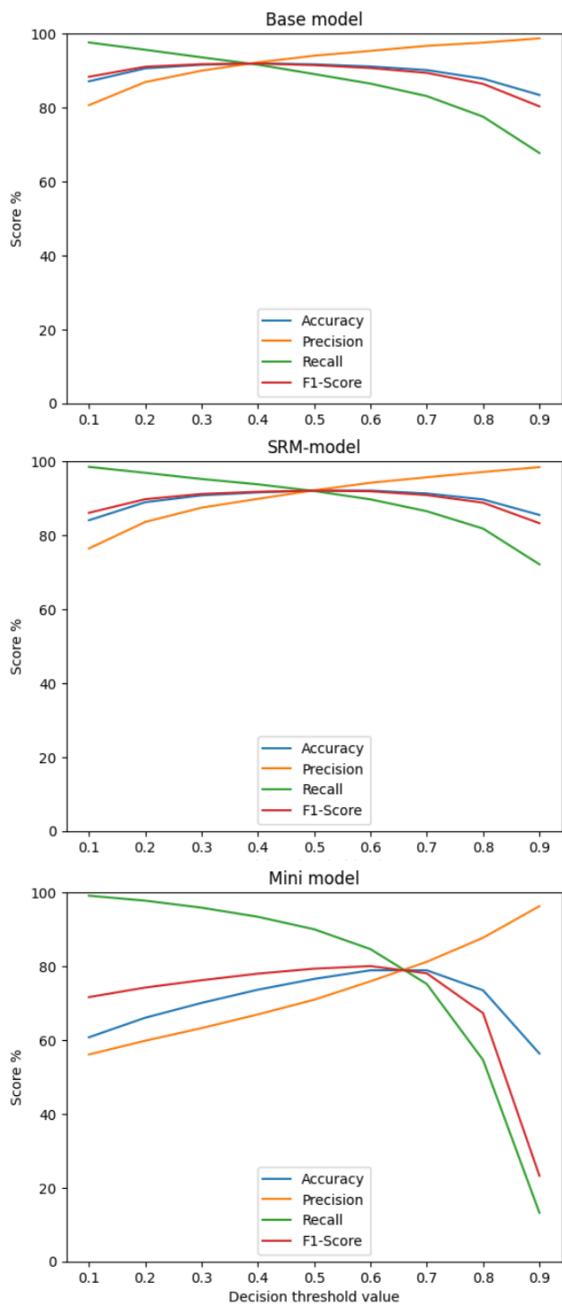


Figure 5. Effect of decision threshold size on evaluation metrics computed for base, attention, and mini model on CIFAKE.

confidence ($p \leq 0.1 | p \geq 0.8$). This in turn leads to suspect that only a small fraction of CIFAKE consists of challenging images simulating CIFAR10 very closely; or conversely falls very short of this goal. This hypothesis is underlined by the results of a base model which was trained for 100 epochs: Though overall accuracy does not improve greatly (91,72%), loss is still decreasing (total loss of epoch 100: 1,203, batch average: 0,0004) and accuracy scores calculated with extreme threshold values remain very high

	Acc.	Rec.	
		coarse	fine
Base	92,91	max 96,8 trees min 88,2 flowers	100 ray 76,0 butterfly
		max 97,2 small mammals min 83,8 large natural outdoor scenes	99,0 castle
SRM	91,31		
		max 98,4 small mammals min 65,4 flowers	100 chimpanzee 35,0 sunflower
Mini	90,37		
			62,0 sea

Table 3. Performance of models on CIFAR-100 ($n=10k$).

($p \geq 0.01$: 90,61%; $p \geq 0.99$: 91,14%). Nevertheless, in comparison with the mini model it is visible that the task and high confidence are only learnt over time, indicating the presence of the expected classification task in CIFAKE, though with an asymmetrical distribution of difficulty across its samples.

4.2. CIFAR-100

When applying the three models on the CIFAR-100 test set (table 3), accuracy scores are near equally high (90,3%–92,9%) with the base model performing best, followed by the attention model and finally the mini version. Overall, this is an indicator that the models did not learn another arbitrary hidden task from the CIFAKE dataset. Models generally do not appear to be influenced by the new content categories as accuracy of the base and attention models are quite similar on both datasets, with the exception of the base model scoring slightly higher on CIFAR-100.

Upon closer inspection of the recall values however, the presence of some interaction between content classes and predictions becomes visible. The ranges of these values in the coarse content labels conditions are smaller for the base and attention models (base: 88,2%–96,8%, SRM: 83,8%–97,2%) than for the mini model (65,4%–98,4%). The latter also displaying both the highest and lowest scores of any model across both content label categories. But the fine label condition reveals that models flag no false negatives in some categories (maximum of one category with a perfect score shown, multiple such categories possible), but visibly struggle in others, e.g. the attention model only being slightly more accurate than chance (62,0%) for *sea*. For this category in particular, it is likely that texture had a sizeable effect on the predictions. Especially given that samples in this category does not seem to include a focal object, unlike CIFAR10 and

	Acc.	F1		Prec.	Rec.
Base	58,16	68,09	Real	65,30	71,14
			Fake	42,74	36,30
			\emptyset	54,02	53,72
SRM	57,14	67,94	Real	64,03	72,36
			Fake	40,35	31,51
			\emptyset	52,19	51,93
Mini	67,09	76,92	Real	68,69	87,4
			Fake	60,76	32,87
			\emptyset	59,79	60,14

Table 4. Performance of models on AGIRI (n=12k).

CIFAKE. This calls to question how representative CIFAKE and thus how suitable its training data is for transfer to more varied data. Inanimate content classes most often present challenges for our models, but this also applies to insects and people. Categories that occur at in the minimum/maximum score categories multiple times are *small mammals* (max. scores for SRM and mini) and *flowers* (min. scores for base and mini). In particular the former of these two and other vertebrates tended to be labelled as 'real' more consistently, possibly indicating that content influences cannot be ruled out.

4.3. AGIRI

Lastly, on transfer to the AGIRI dataset, the observable drop in accuracy scores is most striking (table 4). Both the base and attention models' predictions are only marginally above chance (base: 58,16%, SRM: 57,14%). This pattern repeats across the F1-scores which, across all models, are generally higher than their accuracy scores by about 10,19% on average (SD $\pm 0,377\%$). This is due to the models oftentimes falsely flagging images as fake, but over-classifying images as real at the same time.

It is now the mini model which reaches the best accuracy (67,09%), F1 (76,92%), and per-class measures for precision and recall of real images. Additionally, the mini model outperforms both other models significantly on the precision of fake images (60,76%) by about 20%. Its recall on fake images is rather poor (32,87%), however, and even the best performing model in this measure, the base model, only achieves 36,3% in this metric. Thus, high scores on CIFAKE do not correlate with high scores on AGIRI. Moreover, the performance of the mini model rather indicates that complex features learnt from CIFAKE may be misleading on more diverse datasets.

Though differences between the base and attention model are minimal on AGIRI, the latter ranks as the lowest of the three models in each

metric, except for the recall of real images (72,36%) where it performs marginally better than the base model by 1,22%. Therefore, low-level features based on style as have likely been extracted from CIFAKE, are not reliable indicators for image authenticity overall.

5. Discussion

From the CIFAKE dataset we have seen that the task of synthetic image detection is contained within the dataset, learnable by our CNN architecture with initially promising results ($>90\%$), and non-trivial, as attested by the mini model's training progress. Yet, combining insight from the mini model, variations of the prediction decision threshold, and experiments with a much larger number of training epochs have shown that the learning curve and thus the difficulty of the task is not distributed equally amongst the sample images. While one portion of the dataset can be classified correctly quite quickly, after only one epoch, our architecture eventually hits a ceiling at around 92% accuracy.

Expansion of this architecture with SRM channel attention hit the same ceiling, yet lead to slight but notable improvements, indicating that some remainder of noise due to the image generation process is likely still contained within the synthetic images in CIFAKE. Although it is possible that the models are also sensitive to the content class of the focal entity, influences from low level features are likely. This hypothesis is strengthened on the one hand by the performances across content categories in CIFAR-100, where categories similar to CIFAR10 tend to receive higher scores, while novel or dissimilar categories are more often falsely labelled as fake. On the other hand, Bird & Lotfi (2024) themselves demonstrate using Grad-CAMs that the focal entity oftentimes does not contribute to the prediction.

Ultimately, the suitability of CIFAKE as a training set for the applied task of synthetic image detection is tested by the transfer of the models onto the AGIRI test set. This transfer reveals that high evaluation metrics on CIFAKE do not appear to correlate positively with performance on AGIRI, i.e. learning the task presented within CIFAKE does not enable our systems to generalise features of AI-generated images to a degree that would allow the, to reliably perform this distinction in the wild. This result strongly emphasises the main shortcoming of CIFAKE: its limited highly limited variety in image content, textures, artistic style, and composition. On AGIRI, it is the mini model that outperforms the fully trained models. Consequentially, this creates the impression that any

features suitable for generalisation are learnt early on, while further training on CIFAKE yields a model that is either a) highly specific for this dataset, or b) highly specific for the general style of synthetic images in CIFAKE that is rare at least AGIRI and perhaps overall.

In considering the stylistic diversity that the overarching category of synthetic images exhibits, it stands to reason that design choices regarding model architecture, features, and dataset composition must be considered more closely to allow for 1) the creation of a dataset that is representative of this variety, and 2) the subsequent training of a model capable to handle a task this difficult and diverse in downstream applications.

6. Conclusion and future work

In this project, we replicated the implementation of a CNN model for the task of synthetic image detection on the CIFAKE dataset. We further extended this approach by expanding the architecture with the light-weight channel attention structure of a Style-Recalibration module. Both these models, along with a baseline version of the former, were initially tested CIFAKE in order to examine the nature and difficulty of the dataset's task, revealing a non-linear learning curve. In applying these models to CIFAR and AGIRI and comparing their performances, suspicions concerning the task learnt in CIFAKE were raised, as well as such concerning the proposed suitability of CIFAKE as a training set for the synthetic image detection task. Models performing well on CIFAKE did not transfer similarly to AGIRI, the most varied and thus realistic dates for downstream applications, either due to failure to learn universal features of synthetic images, or due to none of this kind being existent in CIFAKE.

To pinpoint underlying causes for these findings, future work may create, analyse and compare heat maps of the SRM layer to the Grad-CAMs in Bird & Lotfi (2024) extrapolate whether content- or texture-based features were more salient to this model. From the mini model's performance on AGIRI, it appears likely that some manner of generalisable features that characterise synthetic images are present within CIFAKE. Thus, exploring the ideal cutoff point for training on CIFAKE within the goal of application transfer may be desirable. Another research angle in light of this goal is to test the ability of the presented architecture for the downstream task by training directly on the AGIRI training set, though more costly. Alternatively, it is plausible that the remaining 10% of samples beyond the established performance ceiling contain the more salient and universal

features. In this case, improving the model architecture, e.g. by employing different attention mechanisms, ought to be focused on.

7. References

- Bird, J. J. (2023, March 28). *CIFAKE: Real and AI-generated synthetic images*. Kaggle. <https://www.kaggle.com/datasets/birdy654/cifake-real-and-ai-generated-synthetic-images/data>
- Bird, J. J., & Lotfi, A. (2024). CIFAKE: Image classification and explainable identification of AI-generated synthetic images. *IEEE Access*, 12, 15642–15650. <https://doi.org/10.1109/access.2024.3356122>
- Krizhevsky, A., & Hinton, G. (2009). Learning multiple layers of features from tiny images.
- Lee, H., Kim, H.-E., & Nam, H. (2019). SRM: A style-based recalibration module for Convolutional Neural Networks. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. <https://doi.org/10.1109/iccv.2019.00194>
- Misra, D. (2021, April 22). *SRM Channel Attention explained: Paperspace Blog*. Paperspace by DigitalOcean Blog. <https://blog.paperspace.com/srm-channel-attention/>
- Zhang, T. (2024, May 23). *ai-generated-images-vs-real-images*. Kaggle. <https://www.kaggle.com/datasets/tristanzhang32/ai-generated-images-vs-real-images/data>