Investigating the role of context in classification tasks with CNN

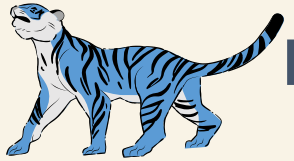
Fabio Ceruti, Fabio Martino, Alex Lucchini, Jacopo Biggiogera

Research Questions

RQ1: Does contextual information in pictures help a CNN classifier perform better?

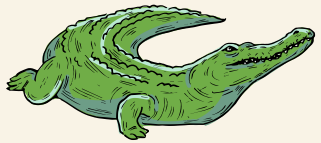
RQ2: If so can we redirect the CNN to better pay attention also to this information?

AGENDA



Part 1: How did we construct the dataset

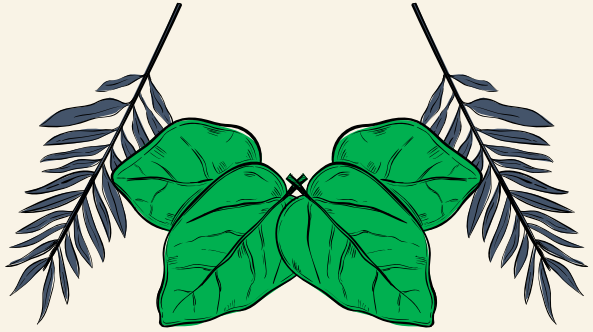
Part 2: Model performance analysis



Part 3: Did we achieve targeted attention redirection

Part 4: Conclusions and applications

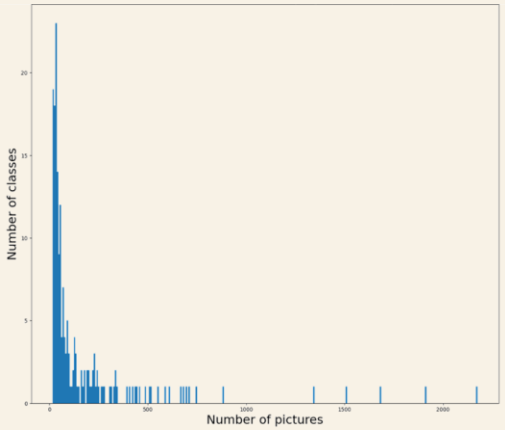
Part 1: How did we construct the dataset



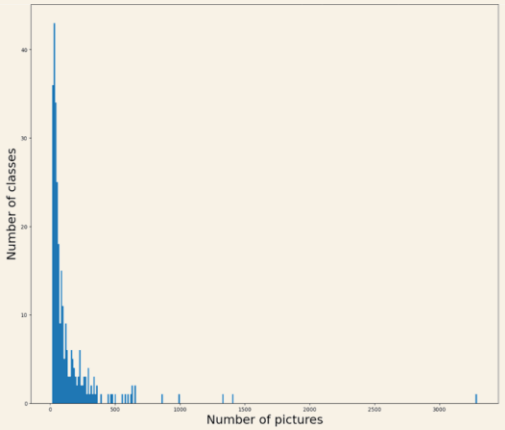
Starting distribution of the Image Classes

- iNaturalist dataset
- Constraints:
  - At least 10k pictures to be able to robustly train a model
  - Avoid an excessive amount of classes
  - Have a diverse range of contexts that would allow us to address our research questions

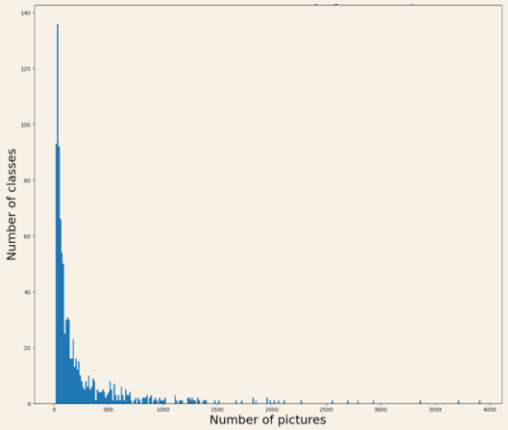
Mammalia



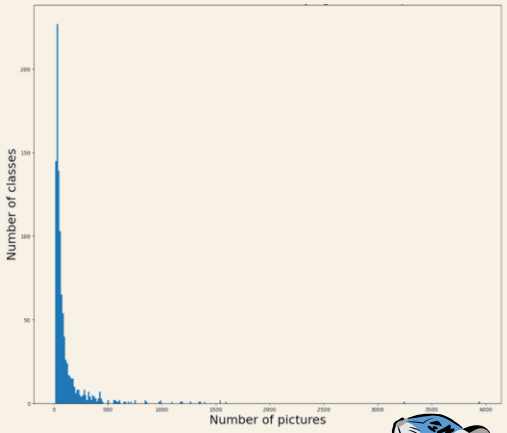
Aves



Insecta



Reptilia



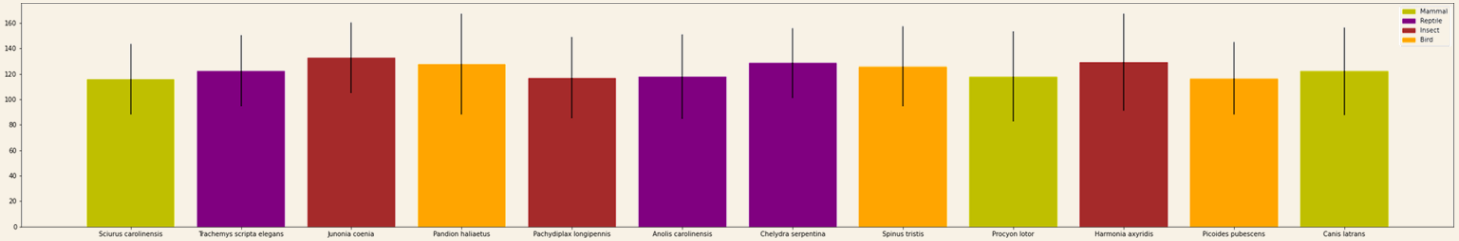
Our final classes distribtution

Final Dataset Dimension:
17439 images

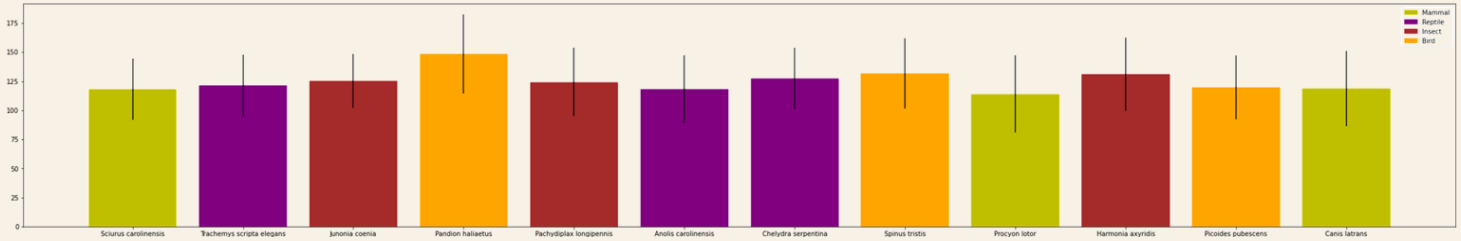
Colour Distribution

Mean pixel color distribution

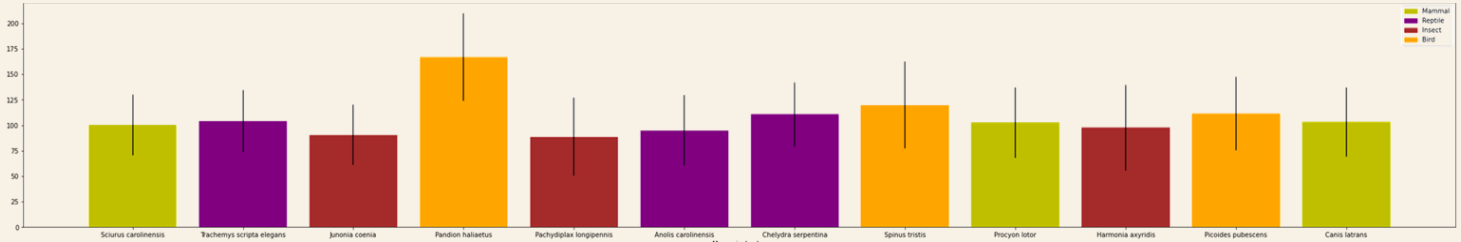
RED



GREEN



BLUE



AVARAGE

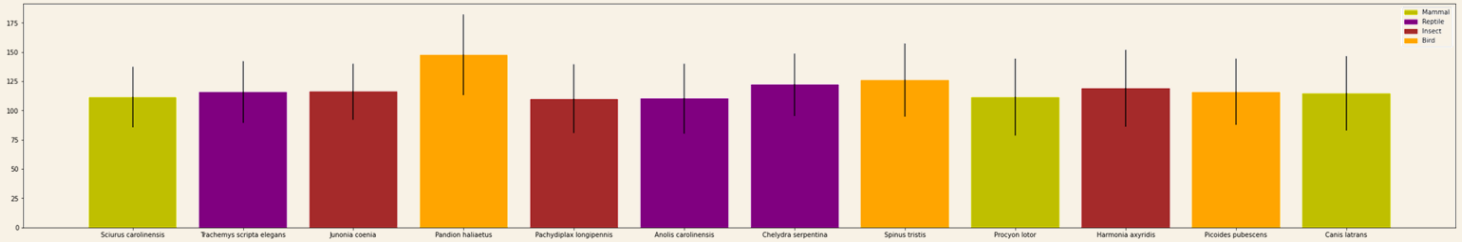
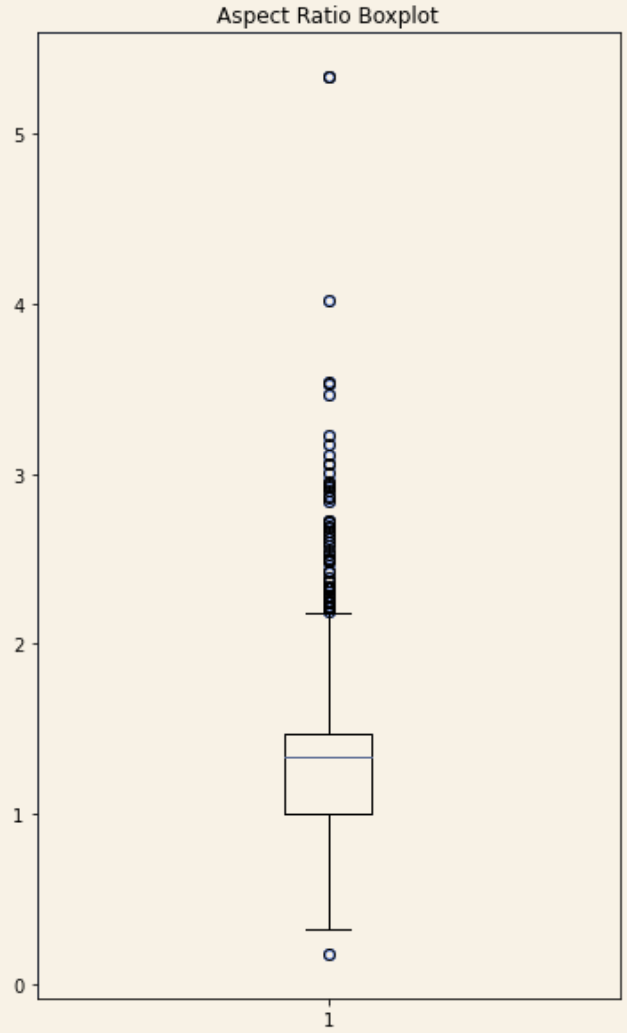
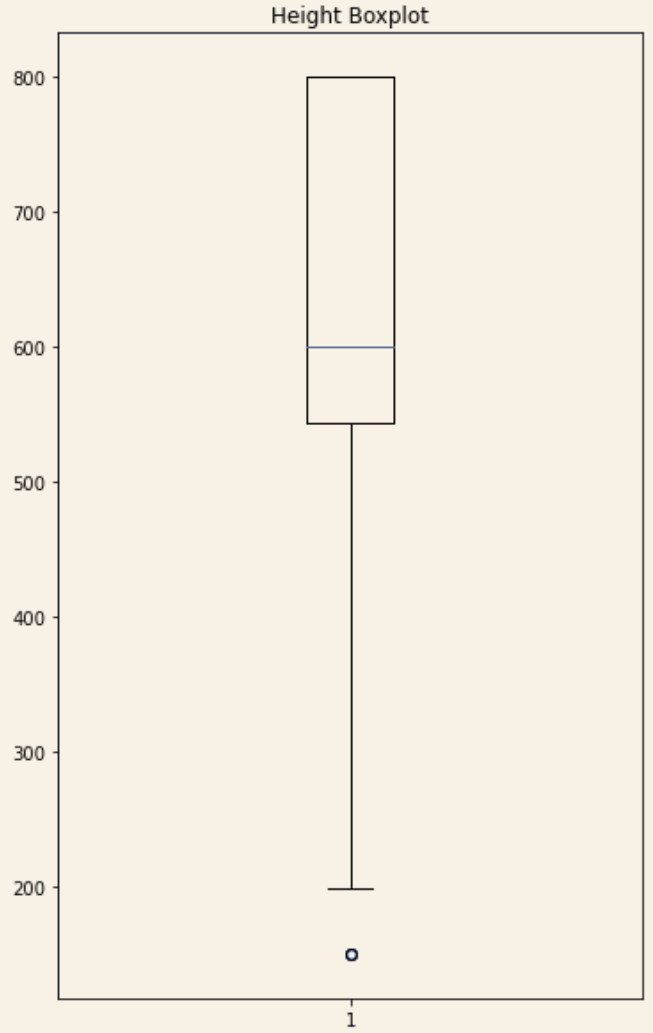
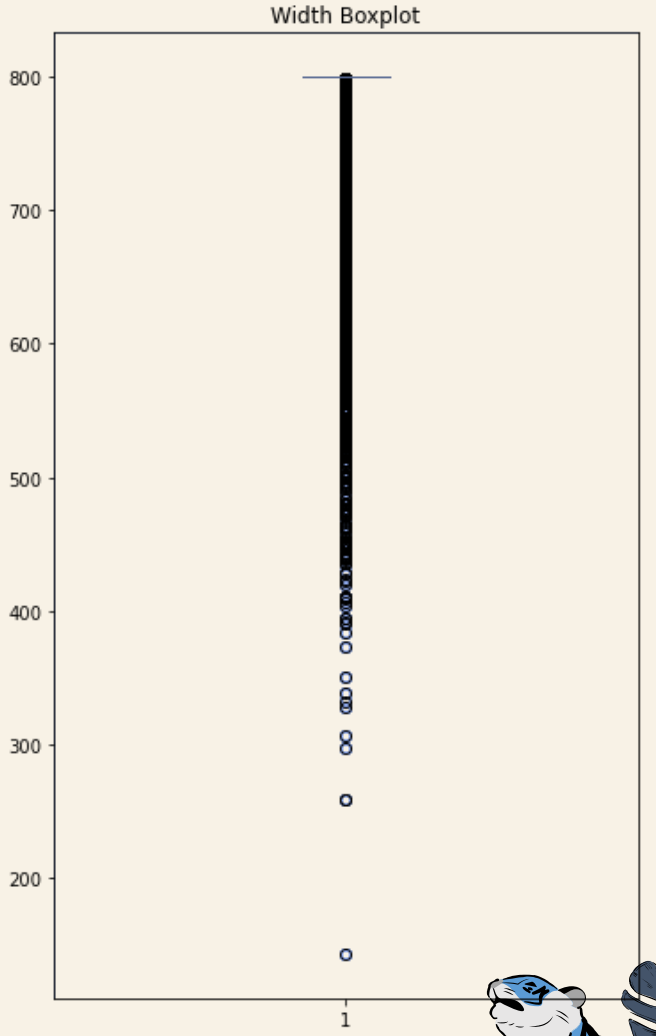


Image Dimensions

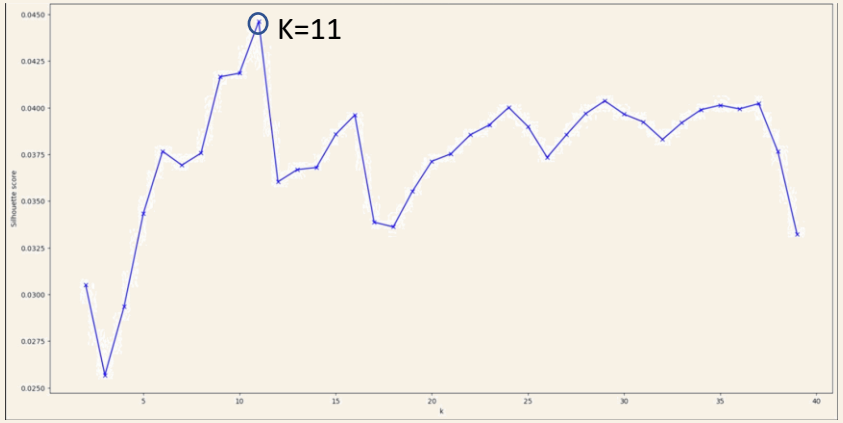




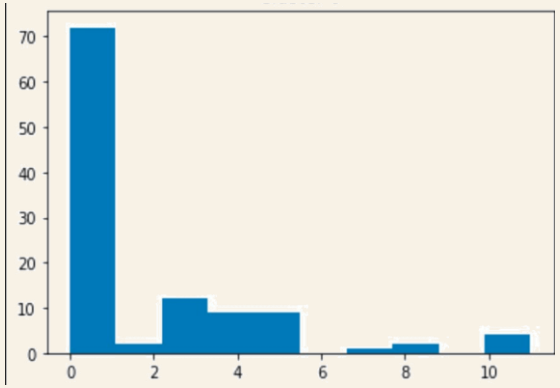


Descriptive Analysis Clustering Analysis

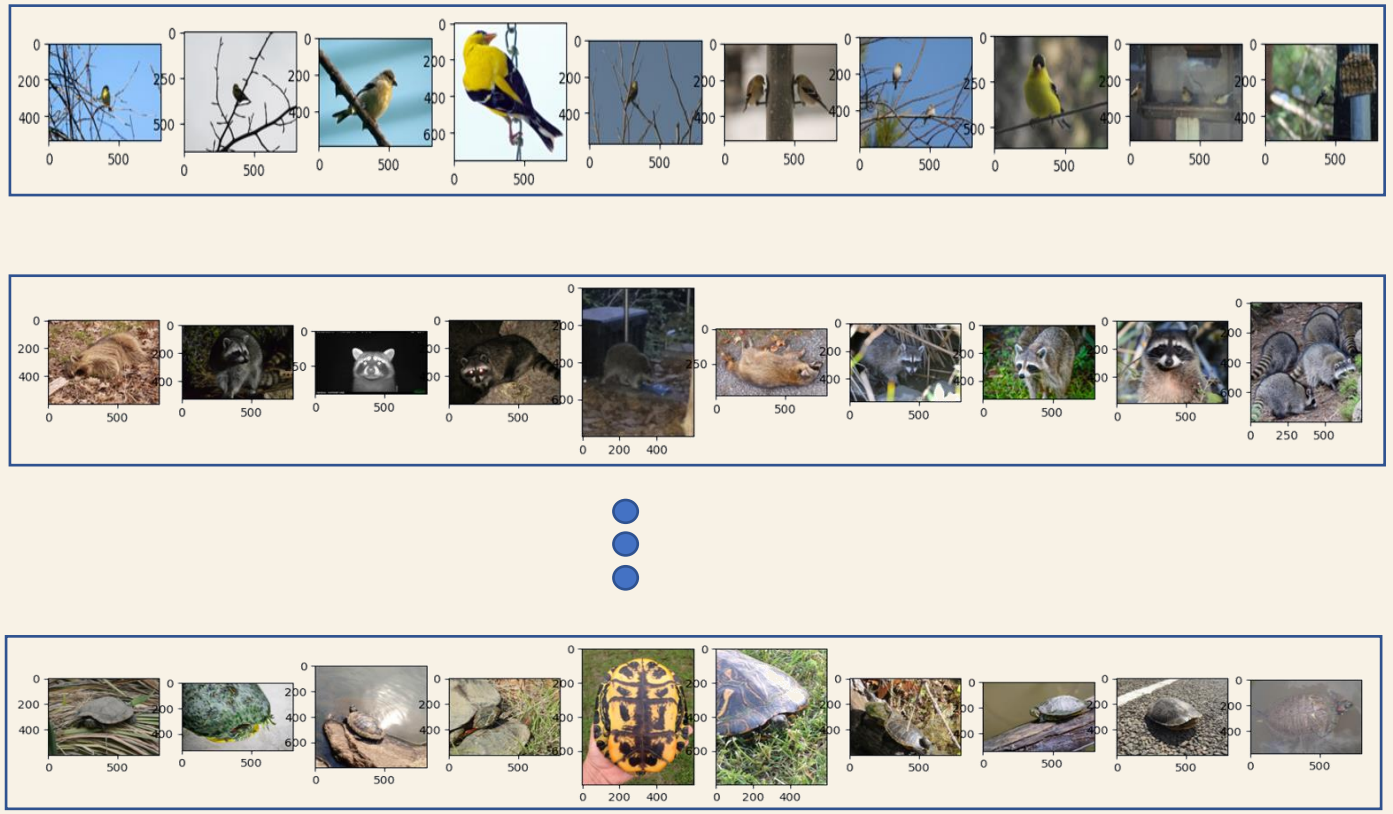
Silhouette Score



K: Number of clusters



Most clusters contain only one class with a high probability



Addressing the Research Questions

First of all, why did we use iNaturalist?

...

Because it has the boxes coordinates !!!

What is Context?

The circumstances that form the setting for an event, statement, or idea, and in terms of which it can be fully understood

The portion of the image exceeding the bounding box

Vanilla

First, we train the model on the original images.

Context

We filled the pixels that were contained in the bounding box with black ones

Bounding Box Content

Used only the portion of the image that was inside the bounding box

Mixed Model

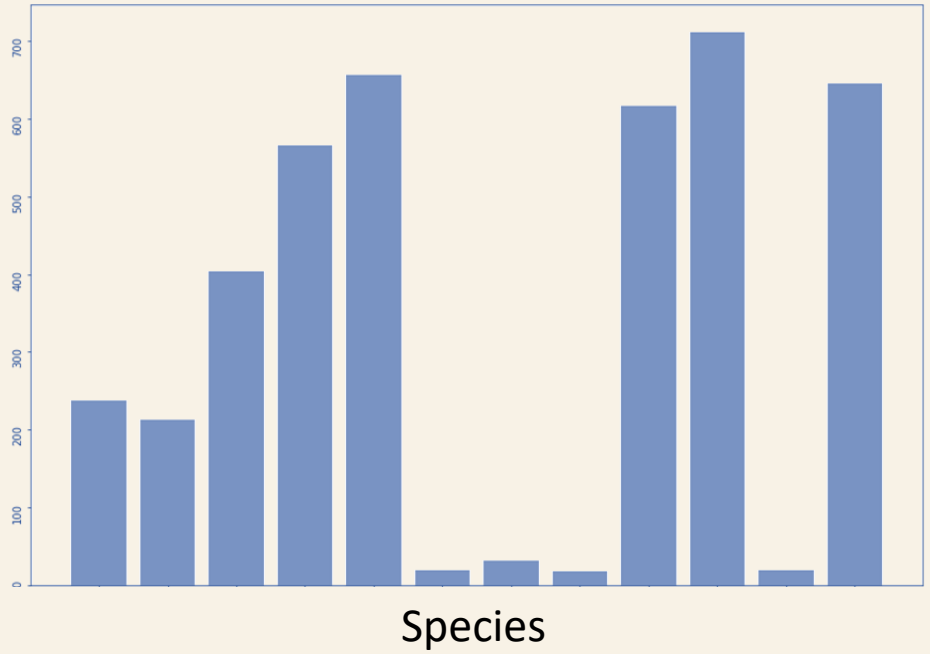
Fed the model with both the context image and the context of the box image

+

Test Split

When we went on splitting the dataset into train and test we notice the presence of several images not having the bounding boxes:

- First thought of using such images as our Test Dataset



So, we decided to generate the new boxes for such images: How ?

CONTRASTIVE LEARNING!

CLIP MODEL

Label: Canis latrans

"A picture of a coyote"

Visual Encoder

Text Encoder

Similarity score

Reference: Contrastive Language-Image Pre-training for the Italian Language Bianchi et al. (2021)

CLIP MODEL

Label: Canis latrans

“A picture of a coyote”

Visual Encoder

Text Encoder

***Sim(pic,text)* = 0,1**

CLIP MODEL

Add a little description of the model

Label: Canis latrans

"A picture of a coyote"

Visual Encoder

Text Encoder

Sim(pic,text) = 0,5

CLIP MODEL

Add a little description of the model

Label: Canis latrans

"A picture of a coyote"

Visual Encoder

Text Encoder

Sim(pic,text) = 0,9

CONTRASTIVE LEARNING

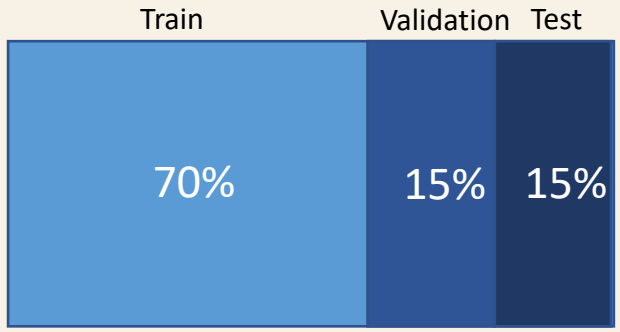
- The Model was able to match the text description to the pixels most similar to the text encoding
  -> **localisation map**

- We built an algorithm to invert the Clip output and resemble a bounding box

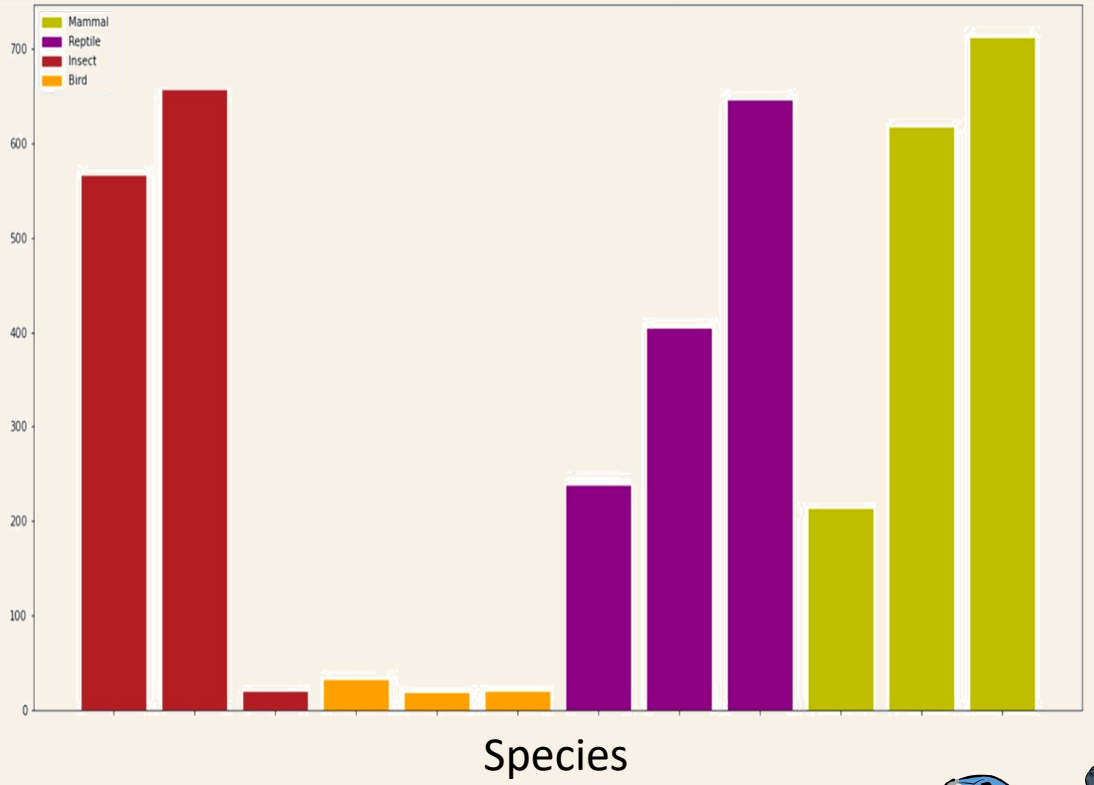
- Finally, we smooth the result into a quadratic shape to reproduce the original bounding box format

Train-Validation-Test

- After having used the clip model we subsequently splitted our dataset:

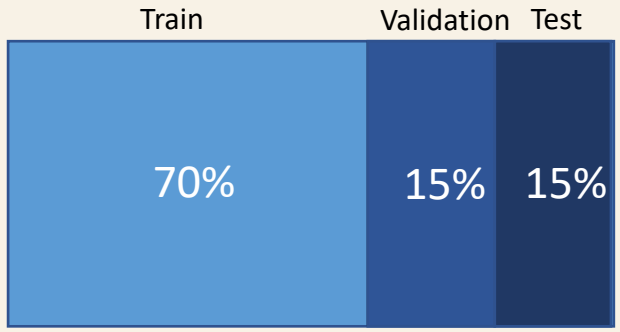


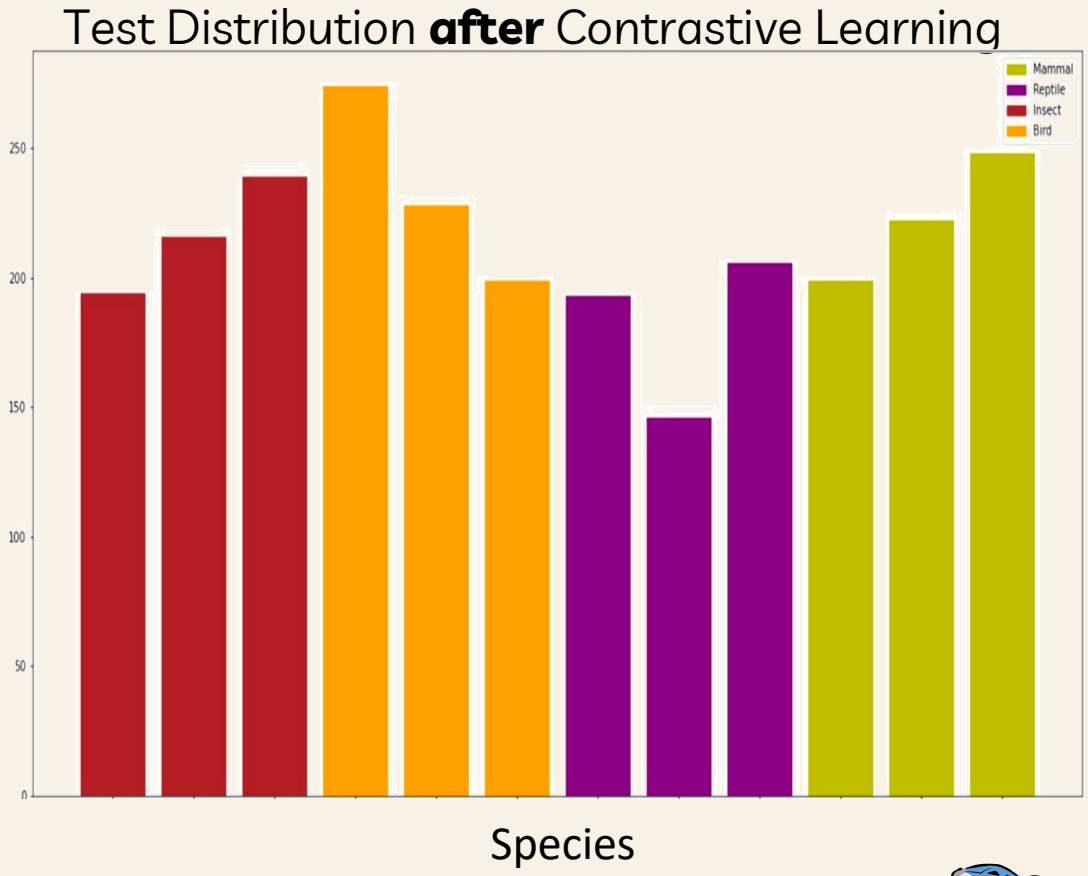
Test Distribution **before** Contrastive Learning

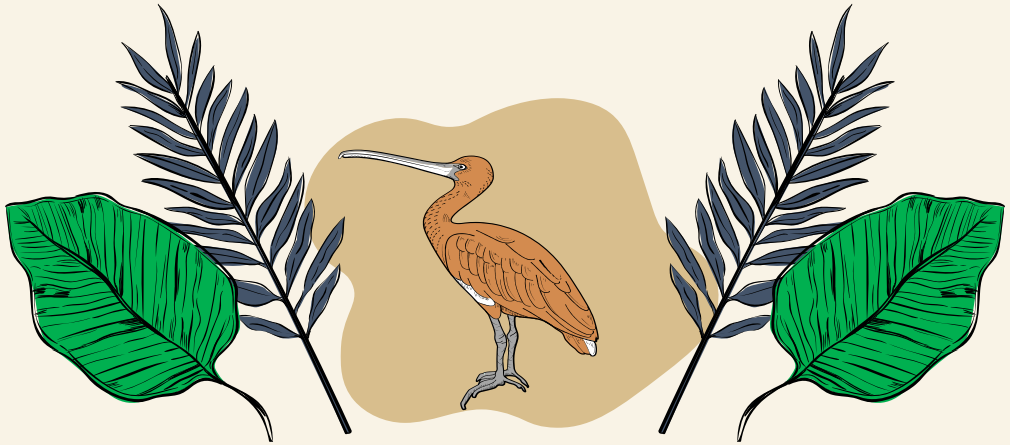


Train-Validation-Test

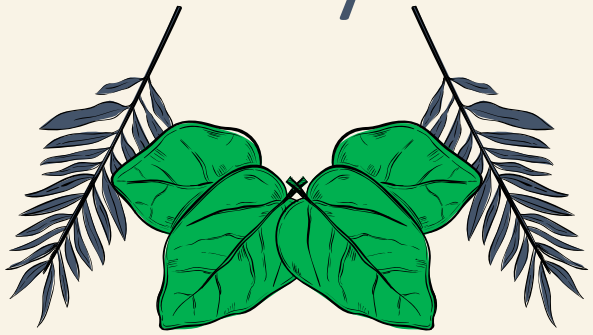
- After having used the clip model we subsequently splitted our dataset:







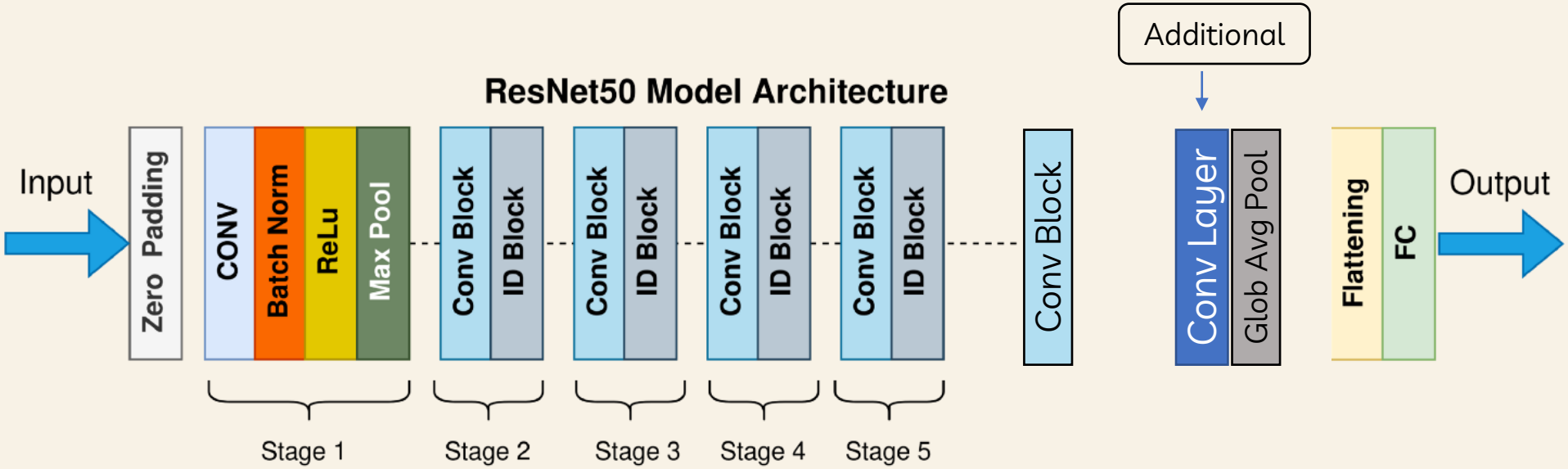
Part 2: Model Performance analysis



Model Architecture

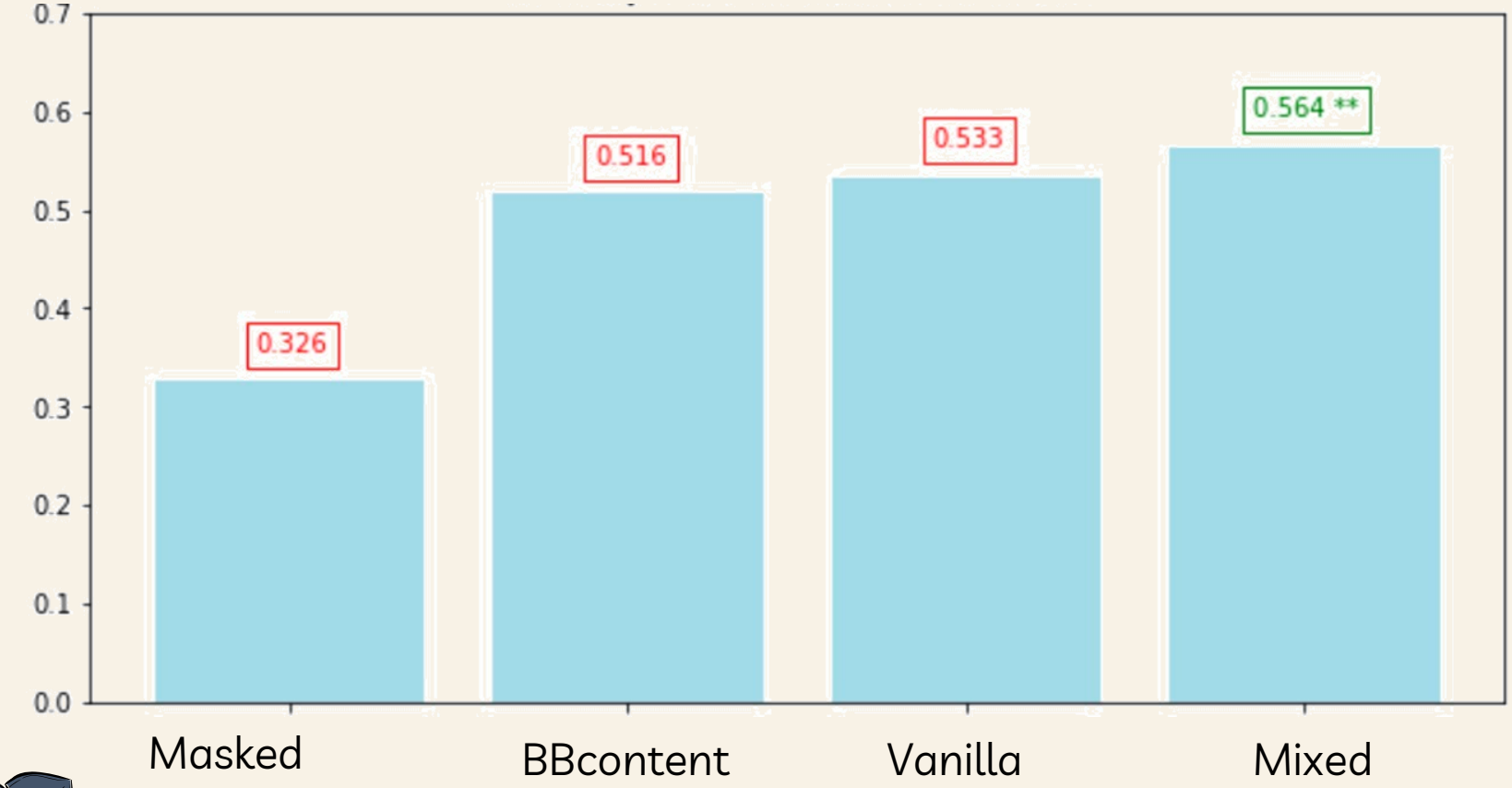
We first tested some custom made CNN configurations we came up with

However, ResNet50 with an additional final convolution layer was the best



We kept this as our final architecture, the one we used to train all the aforementioned combinations

Model Accuracies on test



- Mixed model is significant better
- No significant differences between vanilla and bbcontent

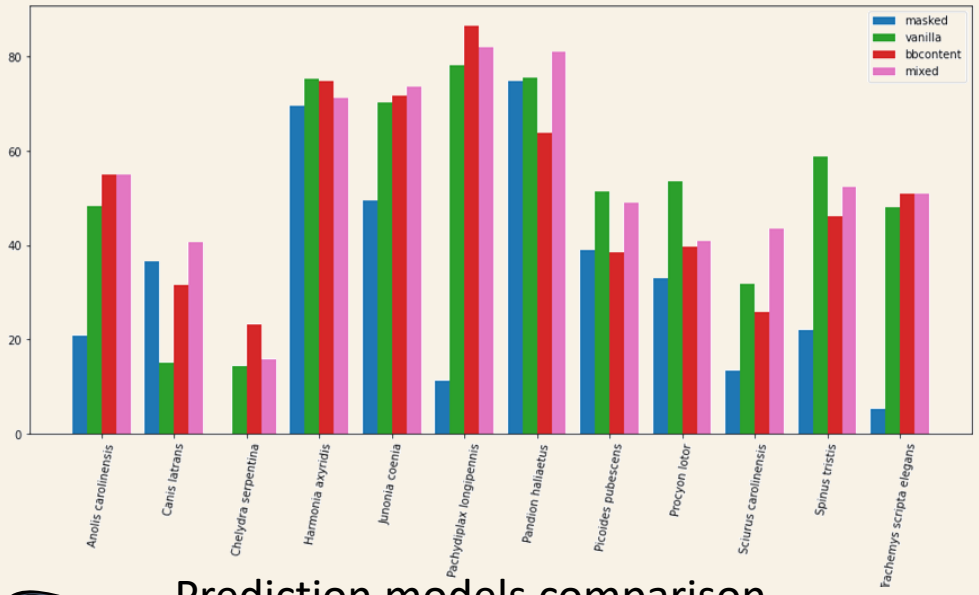
Additionally we used Bootstrapped samples of the test set to ensure robustness of our results.

ERROR ANALYSIS

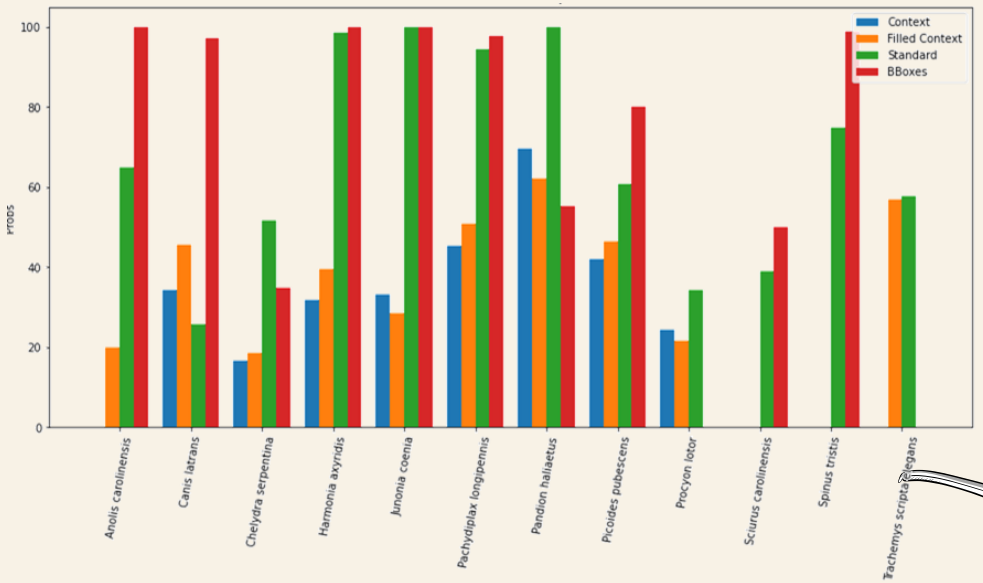
Models performances on Species

To have a deeper understanding of **context** model performance we furtherly check each model accuracy on the single specie

Species Accuracies across models



Prediction models comparison

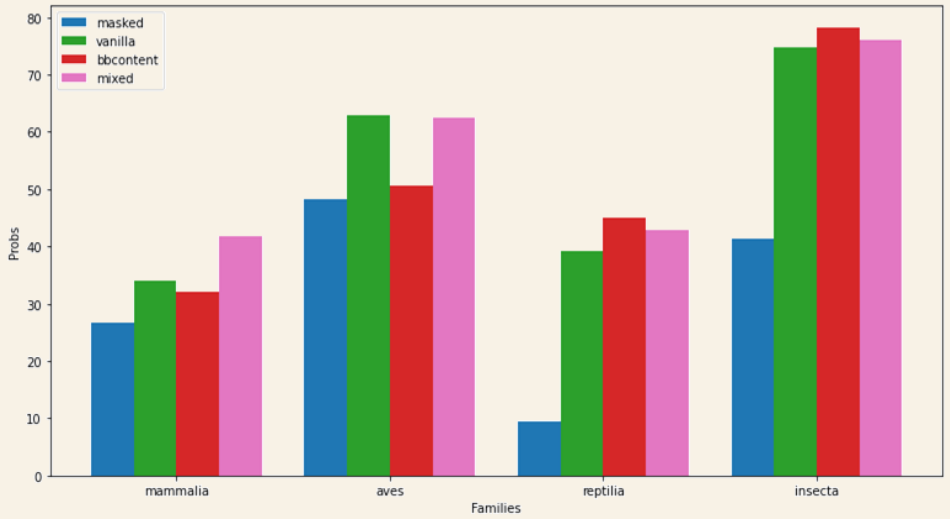


Clustering Analysis

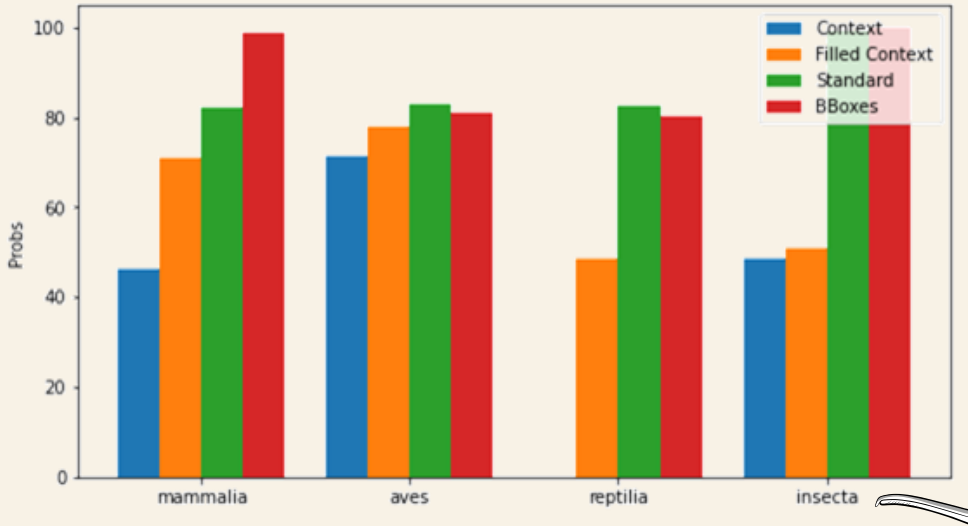
Models performances on Families

To have a deeper understanding of the models' performance we furtherly check for each model its performance on the single families

Families Accuracies across models



Simple model comparison



Clustering Analysis

Why are the accuracies changing so much across species/families ?
Let's evaluate models at the picture feature level !

Feature Engineering

To have deeper understanding of how the models were performing we decided to explore some of the image features.

- *bbox_ratio:* Fraction of the image covered by a bounding box
- *Contrast:* Contrast level of the
- *Bright:* Brightness of the image
- *Log_sift:* Quantity of SIFT features that are present in the picture
- *diff_cont_red* Difference between the average red channel in the context and the average red channel of all test data photos contexes
- *diff_cont_green* Difference between the average green channel in the context and the average red channel of all test data photos contexes
- *diff_cont_blue* Difference between the average blue channel in the context and the average red channel of all test data photos contexes

Logit Regression Results

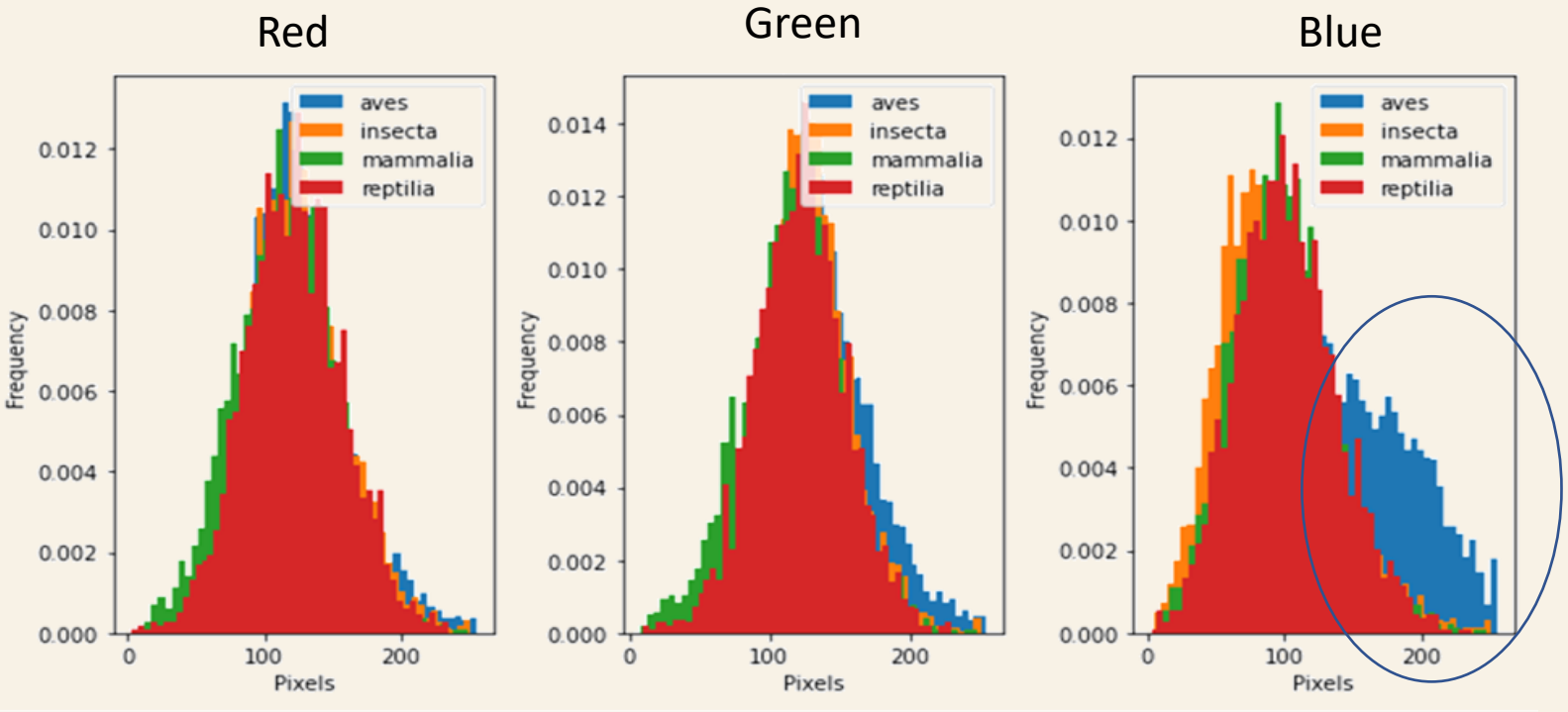
At this point we ran a **logit regression**, where the dependent variable was a binary variable taking value 1 If the image was correctly predicted and 0 otherwise. In this way we aimed at finding potential differences amongst the rightly and wrongly predicted images.

	Vanilla	Masked	Bbcontent	Mixed
bbox_ratio	.0013***	-.0025***	.0052***	.003***
log_sift	-.0726***	-.0277***	-.0869***	-.0651***
contrast	.0007	-.0001	-.0011	-.0007
bright	.0023***	.0025***	.0015***	.0021***
diff_cont_red	-.0005	.0002	.0002	-.0012**
diff_cont_green	.0015	-.0021***	.0011	.0003
diff_cont_blue	**.0016*** **	**.0035*** **	.0002	**.002*** **

Note: Values in the table are Marginal Effects

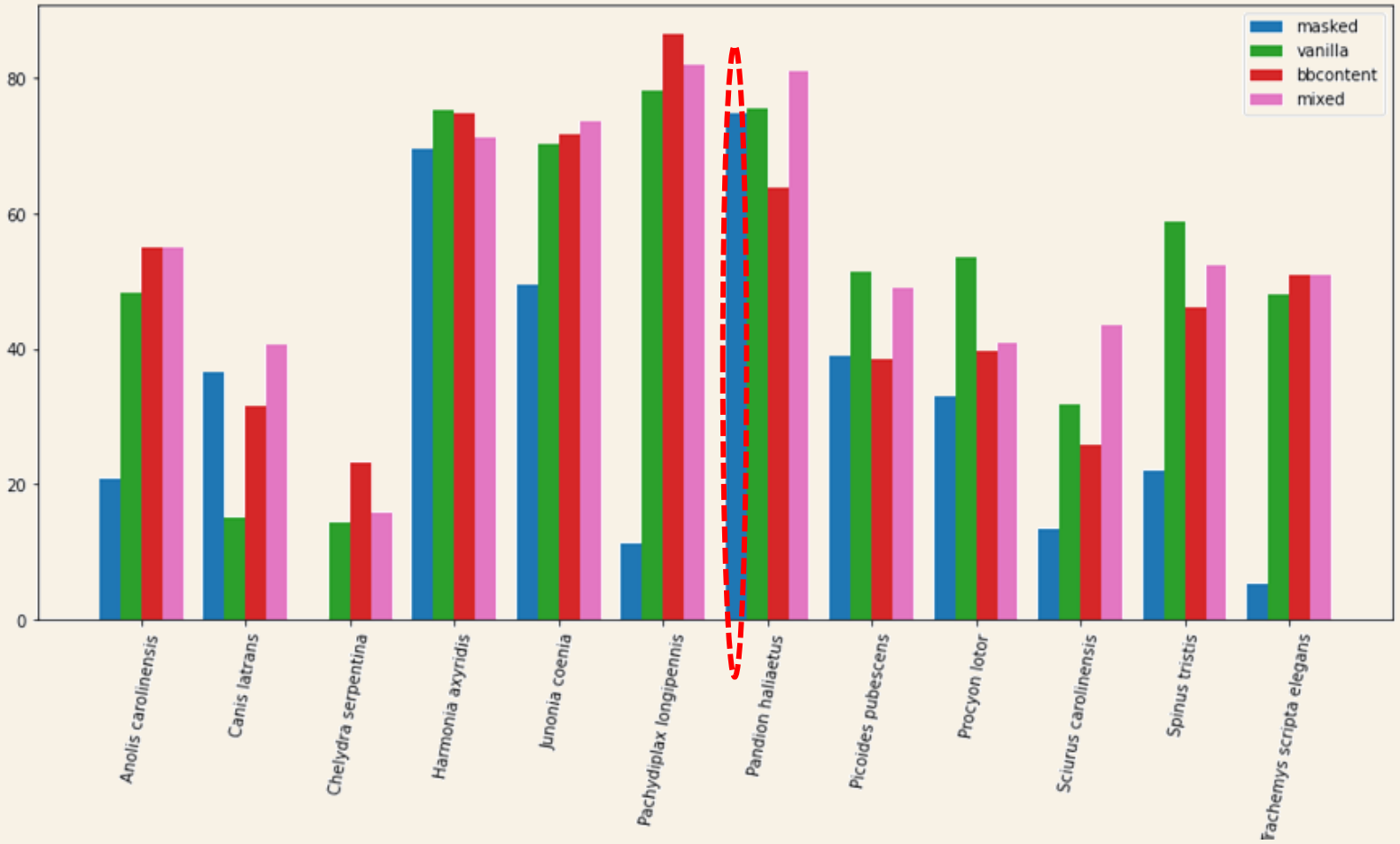
Getting deeper into our Test Images

To furtherly explore the context model performance we specifically analysed the families colour analysis



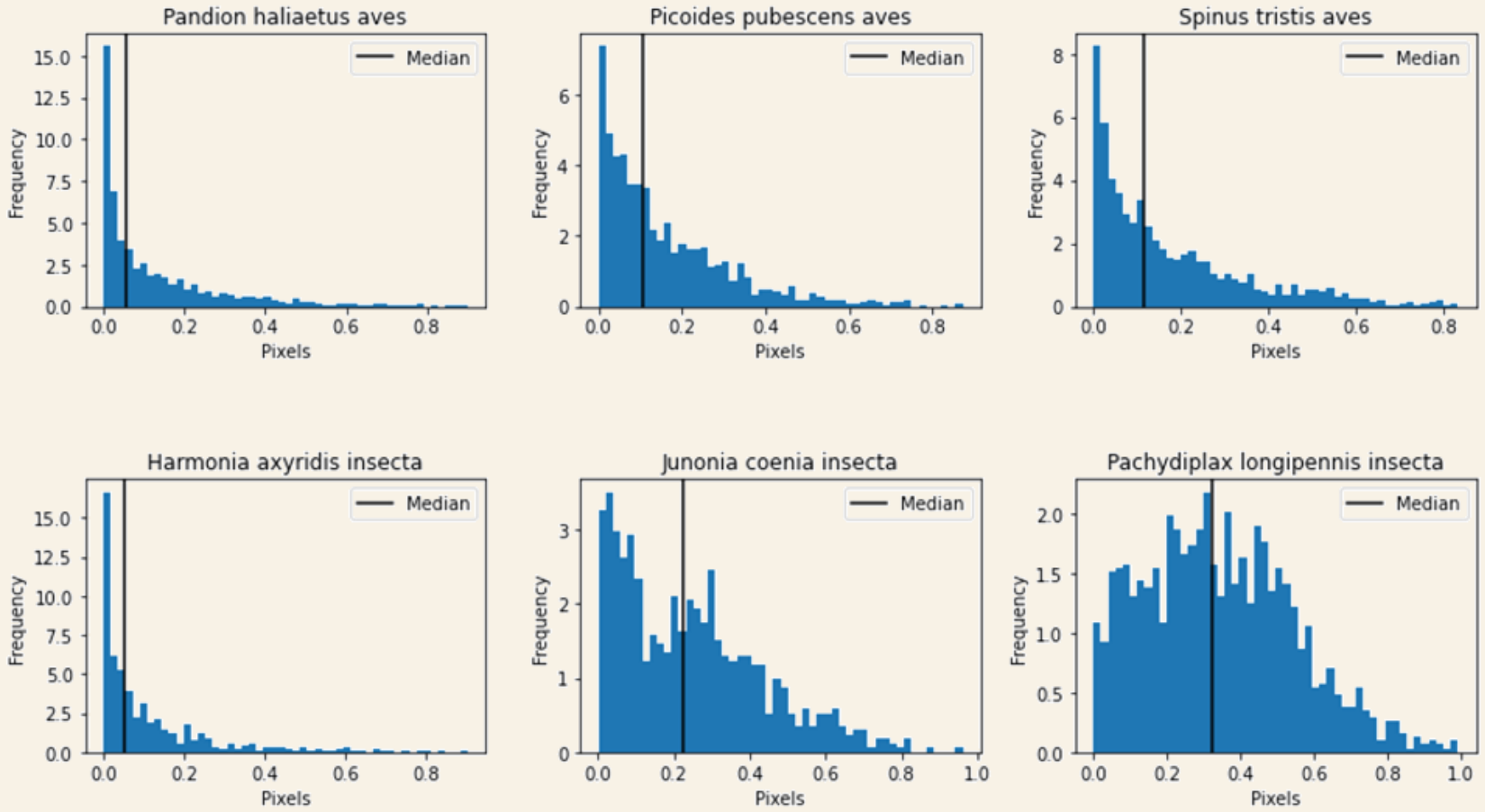
Aves blue channel distribution is strongly skewed to the left: Aves is indeed well predicted by context
This was expected because of how the dataset was built, indeed an higher presence of blue will more easily guide the models towards the aerial or aquatic species.

Species Accuracies across models



Box Content Model Results

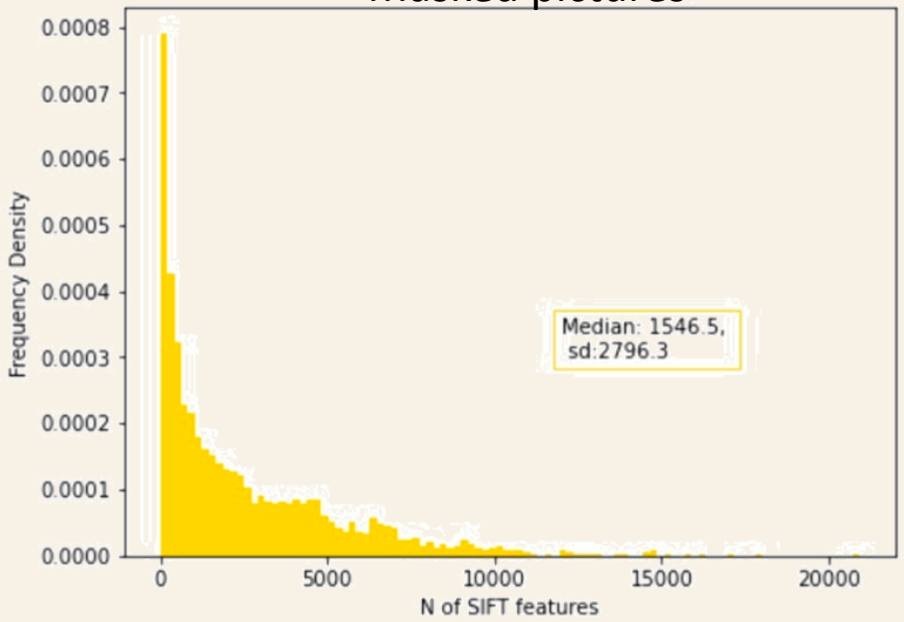
BBox Ratio Analysis



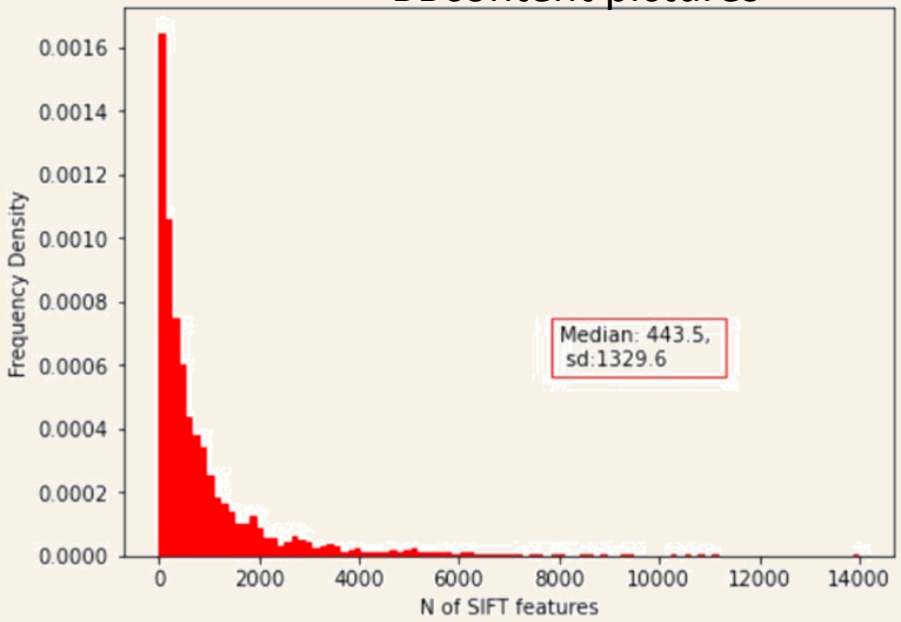
The previous result was furtherly confirmed by the analysing of the bbox ratio: the Aves have indeed the lower bbox ratio (a small portion of the image contained in the bounding box), hence as expected the bbox model performs better

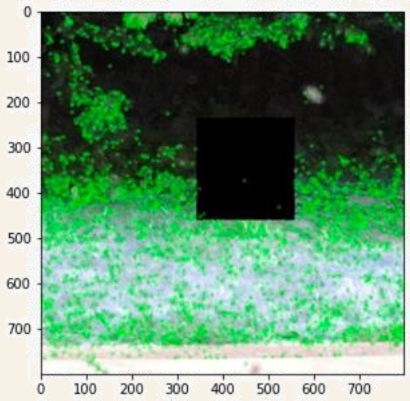
Sift Features Distribution

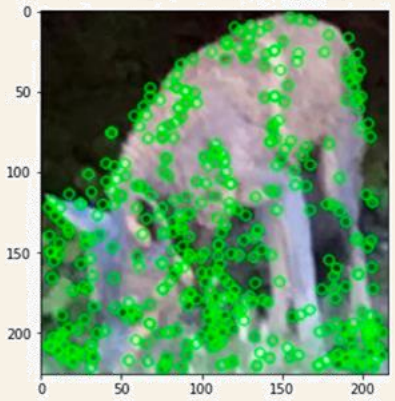
Masked pictures



BBcontent pictures

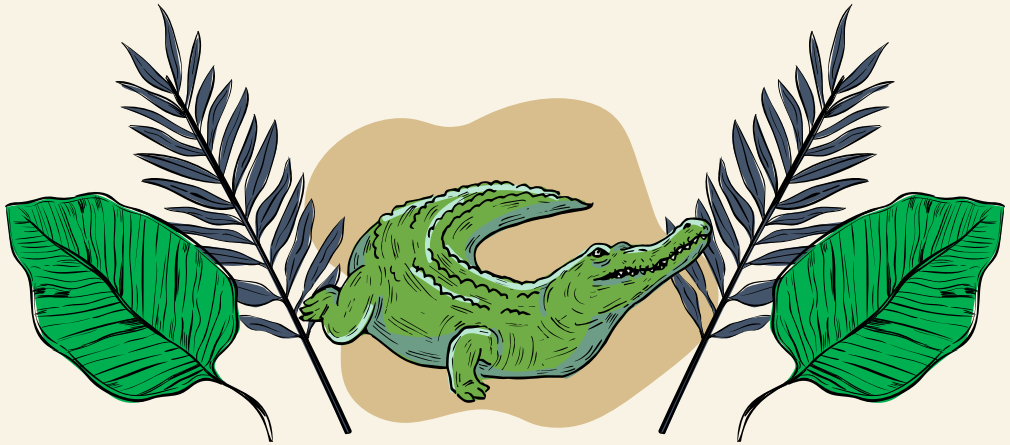




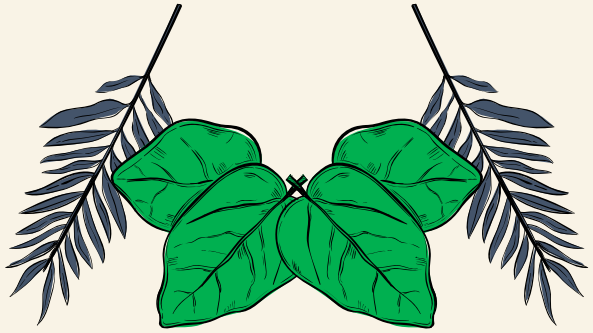


Brightness

As expected the brighter the image the better the prediction!

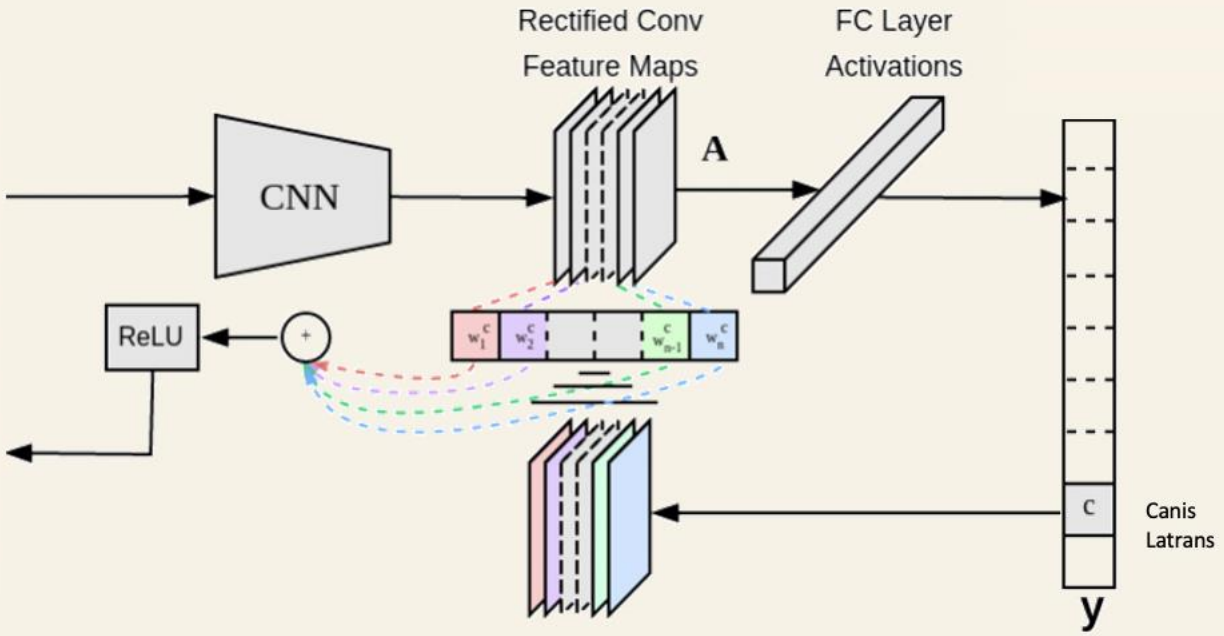


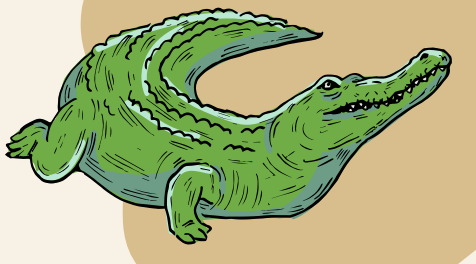
Part 3: Did we achieve targeted attention redirection



Heatmaps

To understand what the model focuses on, we used the Grad-Cam algorithm which captures the gradient of the top predicted class for our input image with respect to the activations of the last convolution layer





Reference: Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization Ramprasaath R. Selvaraju et al. (2019)

Qualitative Heatmaps Analysis

Vanilla

Captures lots of noise coming from the entire image

Context

Focuses mostly on the image borders

Box content

Captures mostly the animal traits

Mixed

Reduce the noise of the all image, capturing the key elements of the image

Focus redirection

To assess whether models were actually focusing on different portions of the image we built an error measure:

- Built a proxy measure: a Matrix of ones having same dimension as the input imahe
- Subtracted the colour matrix resulted from the Grad-Cam algorithm
- Finally took the sum of all the values of the resulting matrix to obtain an observation for the reference model

Numpy.sum

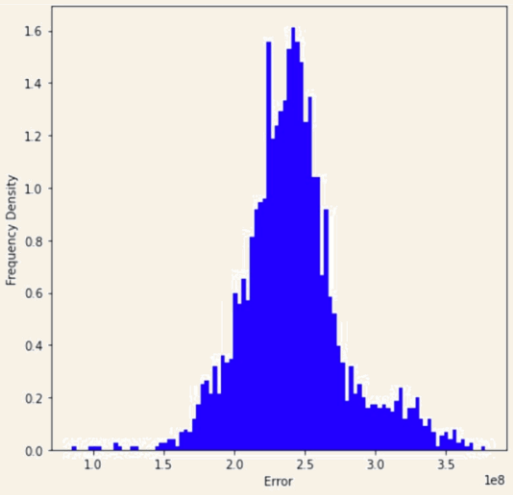
1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1

-

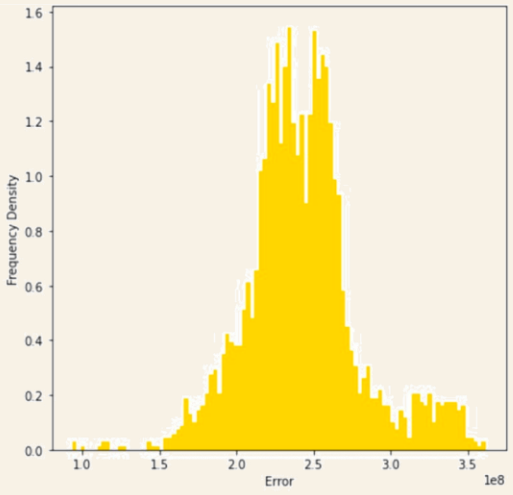
Error distribution

- Iterated across all the test images, obtaining four distributions for all the different models

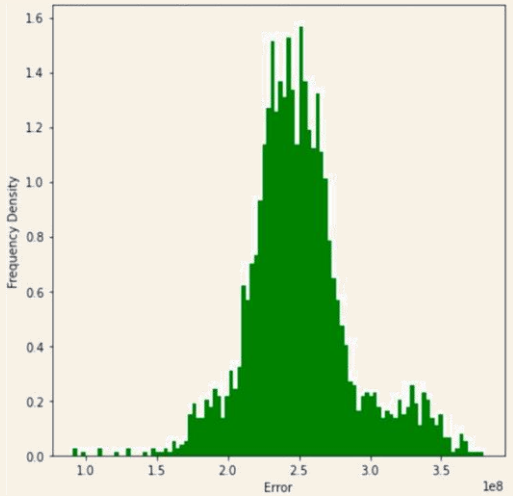
Vanilla



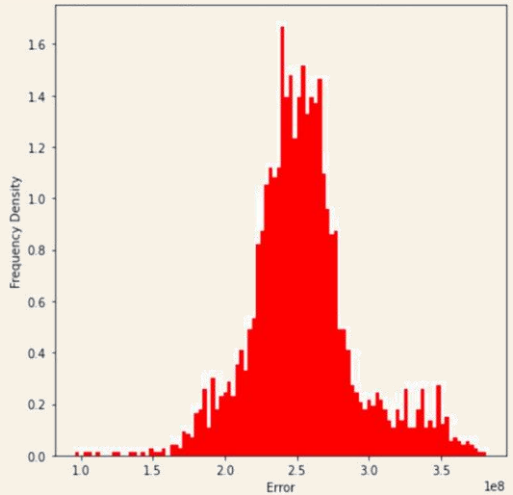
Context



Box content

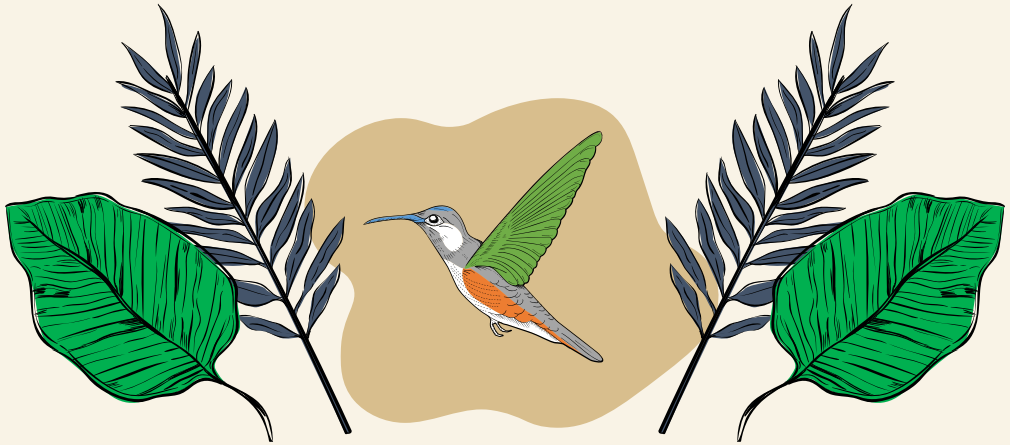


Mixed

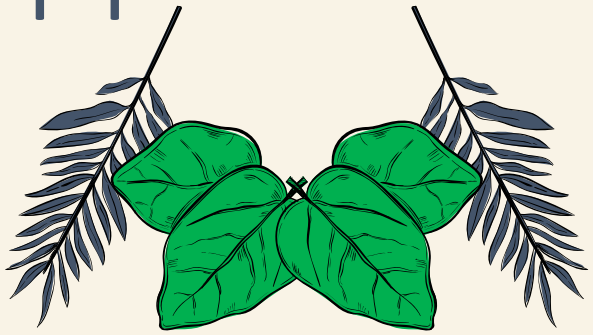


- We finally performed Kolmogorov Smirnov Tests

K-test results	P-Values	Significance
Vanilla-Masked	0.57e-16	***
BBcontent-Vanilla	2.37e-13	***
BBcontent-Masked	0.0115	**
Vanilla-Mixed	3.65e-06	***



Part 4: Conclusions and applications



Answering the research questions

1RQ: Overall contextual information does help a CNN classifier to better predict, however its predictive power mostly resides in its prevalence in the image, and its distinguishing colours.

2RQ: The models' way of learning is obviously strongly linked to the images it is trained on, hence by decomposing such images into the main structures we want the model to focus on we are indeed able to redirect the model attention

Potential Applications

1. Advices for photographers

Low bbox ratio

High bbox ratio

Reducing the distance with the target improves significantly the probability to correctly identify the desired animal

Potential Applications

1. Advices for photographers

Low brightness

High Brightness

Increasing the Brightness of the image improves significantly the probability to correctly identify the desired animal

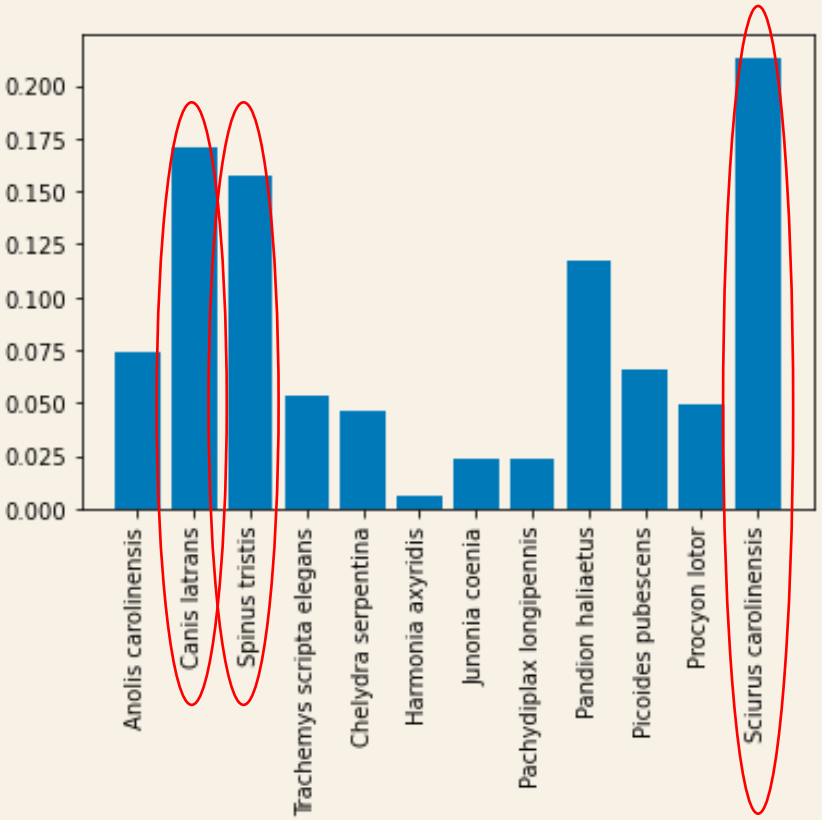
Potential Applications

2.Predicting Animals in Context

- Final Goal: Retrieving the probability distribution of animals potentially living in certain areas just by a picture of the context

Desert of Sonora (Mexico)

Habitat of: Squirrels, Hawks and Coyotes



THANK YOU FOR LISTENING !!!

Anolis

Tristis

Pandion

Picoides