Bird Identification using Deep Learning Techniques

Presentation by Elias Sprengel
University: ETH Zürich
Group: Data Analytics Lab
http://da.inf.ethz.ch
Outline

1 Quick overview of our approach

2 BirdCLEF competition results

3 Dealing with the dataset
   3.1 Pre-processing
   3.2 Data Augmentation

4 Conclusion and Outlook
Overview

- Convolutional neural network (CNN)
  - Five convolutional / max-pooling layers, one dense layer.
  - Employing centering, batch normalization and drop-out.
- Trained on a big dataset (24'607 audio recordings, 999 bird species).
  - Pre-processed data to make it more consistent.
  - Augmented data to avoid over-fitting.
  - Roughly 35 millions weights, trained for a week (GPU).
- Fine-tuning of super parameters paid off.
  - First place in the 2016 BirdCLEF challenge.
Contest Results

- Official score: Mean Average Precision (with background species)
- Mean Average Precision (only main species, same queries BirdCLEF2015)
- MAP with background species (only queries 2016 "Soundscape")
Contest Results [Submissions]

- “Run 1” was an early submission (no fine tuning of parameters).
  - Shows how important it is, to get all the parameters right.

- “Run 2” and “Run 3” were the same architecture but “Run 2” was trained on resized spectrograms.
  - Results are very close (0.536 and 0.522 official MAP scores) but not resizing seems a bit better.

- “Run 4” was just the average of Run 2 and 3 (Ensemble).
  - Suggests that boosting/bagging of CNNs could improve the performance of the system even further.

- Overall, very high scores when targeting foreground species, but slightly lower scores when considering background species as well.
Pre-Processing [Overview]

● To understand contest results, we need to understand the system.

● Pre-Processing in short: We compute the spectrogram (short-time Fourier transform) of the sound file - use image to train CNN.

● Two main obstacles:

  – The quality of the recordings varies drastically:
    ○ Some files contain no audible bird, other contain multiple birds singing at the same time.
    ○ A lot of background noise.

  – Different sound file lengths:
    ○ 30 files in the dataset are shorter than 0.5 seconds, others are as long as 45 minutes.
Pre-Processing [Noise/Signal Separation]

- To remove unnecessary information, split sound file into a signal and noise part.
  - Heuristic, inspired by Lasseck (2013), that extracts segments where at least one bird is audible.
Pre-Processing [Noise/Signal Separation]

• Benefits:
  – Helps the CNN focus on the important parts.
  – Noise part can be used later as a background-noise augmentation method.

• Possible Drawbacks:
  – Can create artefacts in the spectrogram.
    ○ The CNN seems to handle these very well (we create even more in the data augmentation phase without problems).
  – Can miss less audible birds.
    ○ Might be one reason why our scores drop when also considering, less audible, background species.
Pre-Processing [Chunks]

- Second issue was the varying length of the sound files (different widths of the spectrograms).

- Solved by splitting each spectrogram into chunks (fixed-length) and padding the last chunk with zeros.
  
  - We removed the noise part → no “empty” chunks.
  
  - While testing: Multiple predictions from the CNN (for each chunk) → average them to create a more robust prediction.
    
    - Tried other techniques to combine predictions, none of them worked better.

  - Chunk length of 3 seconds was optimal.
Data Augmentation

- Not a lot of samples (average 25 samples per class) → Data Augmentation is super important.
- Time invariant: shift in time!
- Add noise part from other sound files.
  - Great because, eventually, the networks gets to see every bird sound combined with every possible background variation.
- Mix files that have the same class assigned (Takahashi et al. 2016).
  - Class label should stay the same, adding files is equivalent to having multiple birds sing/call at the same time.
  - Helps the CNN to see more relevant patterns at once → faster convergence.
Augmentation

- Augmentation and Drop-Out are the key ingredients to train on a small dataset.
- Apply the augmentation every time → never show the same example twice.
  - Exception: Show the true value (without augmentation) every so often (here, $1/3$ of the cases).
- Combine multiple background-noises (we add three background-noise samples on top of the signal sample) to increase diversity even further.
Conclusion

• We are able to train a CNN (35 million weights) without over-fitting.
  – Works well, even though we have only 25 samples per class.
  – When trained/tested with only 50 random species (1’250 sound files), the network reached a validation accuracy over 90%.
  – Without the use of any external dataset.
  – Without using any meta data values.

• Shows the power of CNNs, even for small datasets (not only bird identification).
  – Requires a lot of care when fine-tuning super parameters as well as good pre-processing and data augmentation methods.
Outlook

- Lots of meta data (Season, Time, Location).
  - Build a model for each region, time, ...
  - CNN reaches higher scores when the number of bird species is low (see tests on 50 bird species).

- Use ensembles (bagging/boosting).
  - Contest results showed potential (simple average of two predictions performed better).
Outlook

- Need to incorporate background species (multi-label).
  - Problem: Pre-processing can remove background species, augmentation methods train the network to ignore everything in the background.
  - One solution: Incorporating background species in training (loss) function (not done for contest submissions).
  - Alternatively, train two CNNs, one for foreground- the other for background-species.
    - Would also help dealing with sound-scape recordings.
Final words

- Some of the ideas might help advance other fields.
  - Example: Acoustic event recognition.

- Showed the power of pre-processing and data augmentation methods.
  - Especially when the number of samples is low and the number of bird species is high (Amazonas acts as the worst case scenario).

- Scores on sound-scape recordings should improve with updated loss function and separate networks, targeting only background species.
  - Even easier if training set would include any examples.
Thank you

• That’s all for now. Thank you for your attention.

• Feel free to ask questions, not about birds though. I can not recognize a single species myself.

• Come to my poster and challenge my results. E.g. How do you compare the performance of two networks?

---


3Image from: http://www.acuteaday.com/blog/tag/fuzzy-bird/
References