CHI – Chair of Health Informatics
TUM School of Medicine and Health
Technical University of Munich

EXPLORING META INFORMATION FOR AUDIO-BASED ZERO-SHOT BIRD CLASSIFICATION

Alexander Gebhard[1,2], Andreas Triantafyllopoulos[1,2], Teresa Bez[2], Lukas Christ[2], Alexander Kathan[2], Björn W. Schuller[1,2,3]

[1]CHI – Chair of Health Informatics, MRI, Technical University of Munich, Germany
[2]Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
[3]GLAM – Group on Language, Audio, & Music, Imperial College London, UK

Motivation

- Plethora of audio recordings for bioacoustics
- Experts have **limited time and resources**
- Utilisation of only auxiliary information could help
- **Annotation** of recorded data **without previous labelling effort**
- Beneficial for **data scarcity** and **underrepresented species**
- Good availability of rich and diverse meta information of birds

Dataset

- **95 European bird species** based on Jung et al. [1]
- **Audio** data gathered from Xeno-Canto in MP3 format
  - ~725 hours
- **Textual descriptions** of bird sounds [2]
  - Example for phoenicurus ochruros (black redstart):

    "Call a straight, slightly sharp whistle, 'vist', often repeated impatiently. When highly agitated, a discreet clicking is added, 'vist, tk-tk-tk'. Song loud, frequently given at first light from high perch, usually consists of four parts: starts with a few whistles and a rattling repetition of same note, followed by a pause c. 2 sec. long, then a peculiar crackling sound (not very far-carrying), after which the verse terminates with some brief whistled notes, e. g. 'si-srü TILL-ILL-ILL-ILL-ILL....... (krschkrschkrsch) SRÜsvisvi'; the sequence of the four components may sometimes be switched around."
- **Functional** traits
  - **AVONET** [3]: ecological parameters, continuous morphological traits, and information on range and location, etc.
  - **Bird life-history** (BLH) [4]: morphological, reproductive, behavioural, dietary, and habitat preference characteristics, etc.

Experimental Setup

- **Non-exhaustive five fold cross-validation** with a training (80%), development (10%), and test (10%) set
- The **dev** and **test** sets among the splits are **disjoint**
- **Training**
  - 30 epochs
  - Stochastic gradient descent (SGD) optimiser
  - Learning rate of .0001
  - Batch size of 16
- Evaluation metric is **unweighted F1-score**

Results

- **Mean results** over the five development (Dev) and test (Test) splits
- **Best** performance is marked **bold**, the *second best* is marked *italic*
- **Displayed metrics**
  - Accuracy (ACC)
  - Unweighted average recall (UAR)
  - Unweighted F1-score (F1)
- Main evaluation metric is the **F1-score**

	Dev			Test		
Embeddings	**ACC**	**UAR**	**F1**	**ACC**	**UAR**	**F1**
BERT	.220	.195	.169	.188	.208	.167
AVONET	.372	**.298**	*.262*	.267	.215	.191
BLH	**.384**	*.288*	**.265**	**.289**	*.286*	*.221*
SBERT	.306	.238	.219	.197	.185	.163
BERT+AVONET+BLH	.181	.175	.154	.175	.168	.151
BERT+AVONET	.254	.193	.178	.169	.158	.141
BERT+BLH	.198	.183	.164	.164	.178	.141
AVONET+BLH	*.335*	.281	.244	*.287*	**.295**	**.233**

Features

Audio

- Audio spectrogram transformer (AST) embeddings
- Resample audio to 16kHz
- Average 2D features over time ⇒ vector with size 768

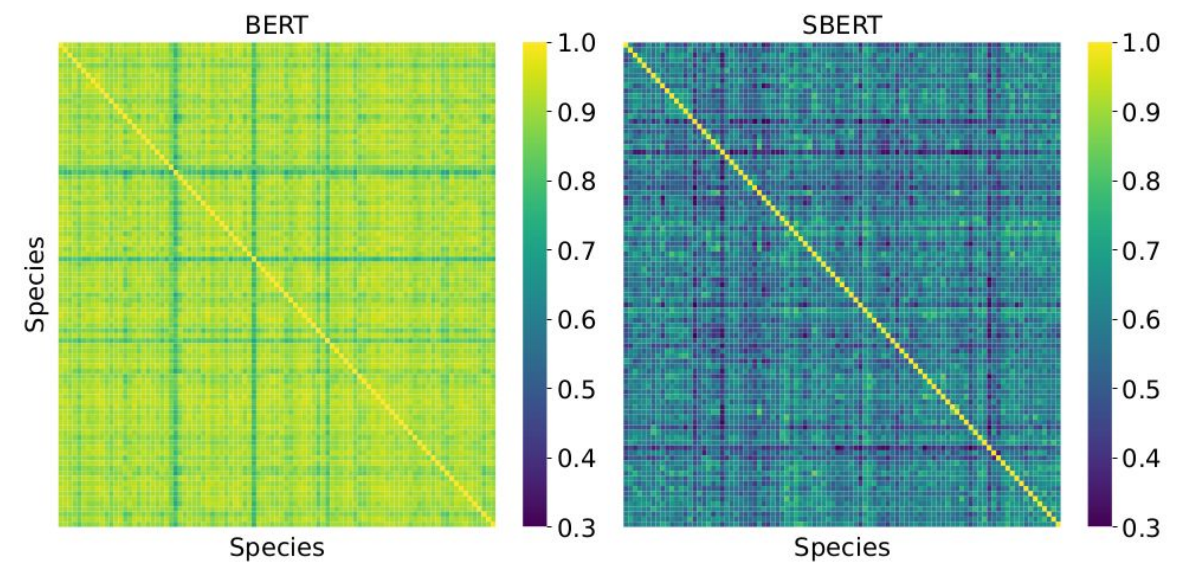
Textual

- BERT [5] and Semantic BERT (SBERT) [6]
- Both with an embedding vector size of 768

Functional

- String labels are encoded to numerical values
- Scaling values to the range [0, 1] via min/max normalisation

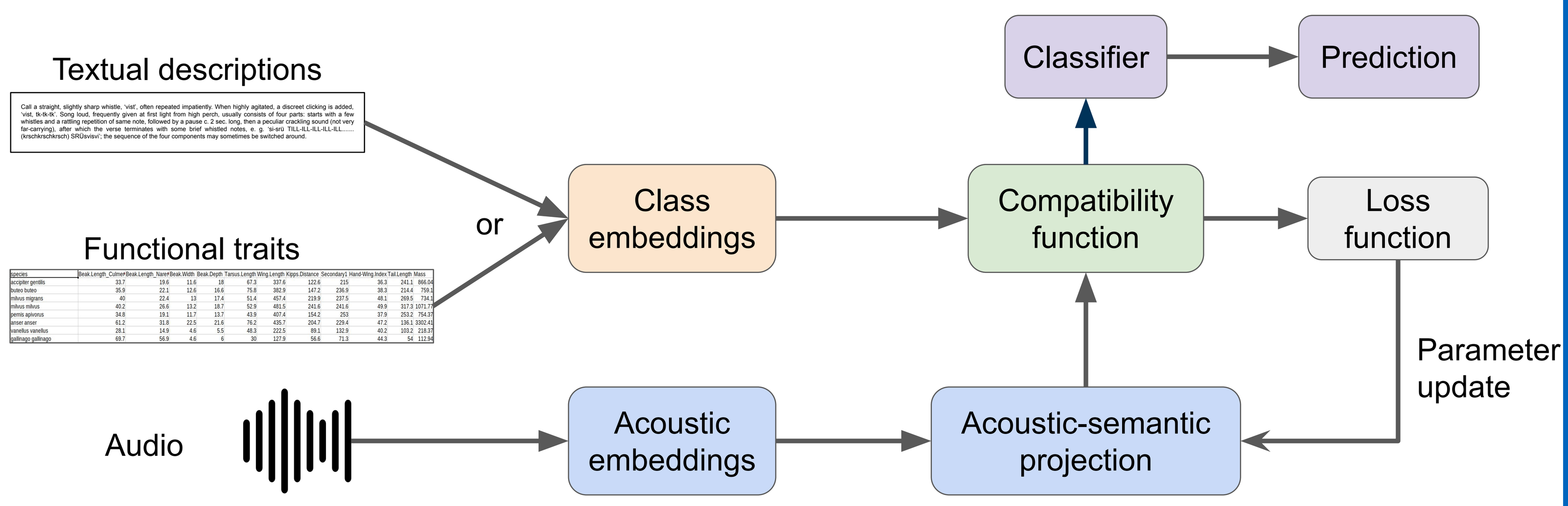
Textual features - cosine similarity



SBERT creates stronger distinctions ⇒ we expect a better performance

Zero-Shot Bird Classification

- Applying a **compatibility function** to an **acoustic-semantic projection**
  - **Project** the acoustic embeddings to the class embeddings with a single **linear layer**
  - **Dot product** as compatibility function
- Standard zero-shot learning **ranking hinge loss** based on [7]
- **Goal:** The **highest ranked** class embeddings **best describe** the audio sample, so that the **class with the highest compatibility** is considered as the **correct prediction**



Conclusion

- The **functional traits outperformed** the encoded bird **sound descriptions**
- **Concatenation of AVONET and BLH** achieve the **best performance**
- Bird-specific **onomatopoeic words/sentences** might be a **problem** for the pre-trained language models

References

[1] K. Jung, et al., Bird survey data 2012, all 300 eps, Dataset. Published. Version 2, Dataset ID: 24690, 2020.
[2] L. Svensson, et al., Birds of Europe (Princeton Field Guides), 2nd ed. Princeton University Press, Feb. 2010, 448 pages, ISBN: 978-0691143927.
[3] J. A. Tobias, et al., "Avonet: Morphological, ecological and geographical data for all birds," Ecology Letters, vol. 25, no. 3, pp. 581–597, 2022.
[4] L. Storchová and D. Horák, Data from: Life-history characteristics of european birds, https://doi.org/10.5061/dryad.n6k3n, Dryad, Dataset, 2018.
[5] J. Devlin, et al., "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. of the 2019 NAACL-HLT, Volume 1, Minneapolis, Minnesota: ACL, 2019, pp. 4171–4186.
[6] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in Proc. of the 2019 EMNLP, ACL, Nov. 2019.
[7] H. Xie and T. Virtanen, "Zero-shot audio classification via semantic embeddings," IEEE/ACM TASLP, vol. 29, pp. 1233–1242, 2021