Reviewed by Jacob Davidson, 17 May 2023 21:25
The authors develop a machine learning approach for predicting animal trajectories that uses the Transformer network architecture in order to incorporate past information of multiple features into predictions. By fitting to data from many different species using openly available MoveBank data, the authors compare predictive ability for different features and as a function of how much time of previous data is included.

I think the approach is interesting and makes a contribution that others working with movement data can use and build on. I have one main concern about the fitting procedure though: mainly, can the conclusions comparing the species be made, when the model is fit to all data together and the number of data points varies so large between species? The authors also note this (line 468). Does this discrepancy of data affect the conclusions? I could imagine an alternative fitting procedure, where each species is weighted equally, instead of each trajectory point. I feel that this comparison, or else further description justifying why the species comparison is driven by behavioral differences instead of simply different amounts of data, is needed.

We actually do not make any definitive conclusions about the behavioral differences between species precisely because of this issue, which we have highlighted in the paper. While we had thought about an experiment addressing the data imbalance, we realized that this could be tricky. We could oversample less frequent taxa or put a proportionally higher weight on them during training, however this could easily result in an overfitting problem where the model would simply memorize these species' trajectories, possibly further hindering generalization instead of improving it. For this reason, we leave such experiments for future work, and also hope that this issue will be alleviated by increased availability of diverse animal movement data in the future, as we mention.

Minor comments:

Legend text on Fig 1 is too small, and lacks units

We have improved the legend and added a caption to clarify the plot shows the number of location events.

Fig 2 shows PCA results for comparing the species, but does not show the PCA vectors. I'm not familiar with the Wikipedia2Vec data, but for PCA the vector components are normally shown, so that one can see what the embeddings represent. If this is not relevant for showing the Wikipedia2Vec embeddings, then it should at least be mentioned.

Wikipedia2Vec is a distributed representation that does not have meaningful dimensions or units, hence showing the components indeed wouldn't be informative. We now mention this in the paper.

Reviewed by anonymous reviewer, 12 May 2023 12:04
# General comments

In this paper, the authors propose a deep learning (neural network) model for analysing animal movement trajectories, called MoveFormer. The model is step-based, predicting an *animal’s next step based on the environmental context (as in a step-selection function). However, the model learns the entire trajectory before that step, thus incorporating (potentially long) temporal context to make predictions. Being a deep learning approach, we expect that the model is capable of learning complex relationships and having high predictive power. I believe that a similar deep learning approach for analysing trajectories (sequences) is Long Short Term Memory, but the paper uses recent developments such as the Transformer architecture. In my reading, I have not seen a similar approach applied to trajectory data. In general, I think this is a useful contribution to movement ecology: I feel we will (and should) see more approaches like this, which leverage the potential of deep learning methods for incorporating sequence (temporal) information when analysing animal trajectories. Specifically, the incorporation of previous movements (history) is an important advantage of the approach, and the estimation of a ‘context length’ (time window) that is most important for being able to learn the trajectory is a key contribution.*

I have some familiarity with simpler machine learning methods, but not much expertise in deep learning. I cannot comment on many technical aspects of the work, particularly the specific architecture and implementation of this model. Nonetheless, I offer below a few general comments, and a small number of specific comments.

Clearly the model has impressive predictive capability – I wondered whether it’s possible to forecast more than one step ahead, or to predict a whole sequence of the same output length as the input? I understand that this is no longer exactly step selection, then, but I think this would be the kind of application many movement ecologists would be interested in. If we can only predict one step ahead, then a simpler approach may be better (next point)?

A one-step-ahead prediction model can in principle be repurposed for forecasting many steps ahead simply by iteratively predicting one step ahead and hence generating an entire trajectory. However, such a model has to be extremely robust in order to be useful for these kinds of predictions: small prediction errors can quickly accumulate, and even a single bad (out-of-distribution) sample can put the model in an unrecoverable "failure state" (a typical problem in language models even today). Our initial experiments on this did not seem to result in meaningful trajectories, and hence we do not further explore this direction in the paper.

Another possibility would be to simply "ask" the model to predict farther into the future, as our time representation is flexible enough to allow prompting the model with an arbitrary future timestamp. However, the model would also need to be trained to do this kind of prediction, requiring a more complicated training strategy. In particular, it is not clear how to generate suitable candidates for selection, both during training and testing.

I was curious about inference from the model. If the model cannot (or should not) be used to predict more than one step ahead, and inference is limited due to the black box nature of the model, is it better to sticks with more traditional step selection functions if inference is the goal? I guess this is particularly relevant given how much data are clearly required for the present model. Regarding inference, is it possible to look at ‘selection’ of environmental conditions, along with the importance currently shown in the paper?

For (one-step-ahead) predictive purposes, nothing prevents one to use standard step selection function approaches if one thinks they are well suited for the job. However, as argued in the manuscript, it is difficult to integrate information about past context in such models, and being able to do so is one of the strengths of the approach proposed here. We trust this should improve the predictive power much, but a quantified comparison between these approaches would be interesting to conduct. It is however beyond the scope of the study presented here. Also, we note that if the goal is prediction, then the black box nature of the model should not matter.

We interpret the second part of the question as a question about whether it is possible to obtain ‘selection coefficients’ as in standard step-selection-function. The answer is no because the model structures are widely different and the deep-learning model proposed here does not estimate the relationship between use vs available case and environmental variables in the same way as the SSF models fitted by conditional logistic regression or Poisson models do.

I appreciated that some features of the model – I’m thinking particularly of the different time-scale periods in the model – were kept general to maximise the wider (future) application of the model.

Regarding the stated contributions number 2 and 3 -- “Second, the proposed approach is flexible enough to allow each step in the context to be defined not only by the locations of the start and end points, but also by any kind of features that could be relevant, in particular environmental variables. Third, we show how the model can be used to gain insights about the importance of the provided context, both in terms of the extent of the past that it is useful to know, and in terms of what kinds of information are most ecologically relevant to predict an animal’s movement” – it would be really interesting (in future work!) to see how the model responds to variation in the spatial scale and resolution of the environmental context variables.

This is a good point and we now mention it in the discussion.

I found the evaluation of the relevant context length very interesting. In future work it would be interesting to further examine patterns of context length among species, beyond what is presented here, and in different ecological settings.

We agree. We however did not make reference to this in the revised version because we do not have immediate predictions to offer for testing, and we feel that without them it would unnecessarily lengthen the manuscript.

The present study uses GPS data – I would be interested to hear the authors’ thoughts on how to deal with lower accuracy tracking data such as Argos.

We suspect that the main issue with low-accuracy tracking data would be that the environmental variables associated with location would be incorrect as they would correspond to the wrong location. A possible solution would be to take some local neighborhood into account, e.g. to retrieve an entire patch from the raster (centered on the desired location, instead of just the single value at that location) and then either compute some sort of weighted average over it or directly feed it to the model to compute an embedding. This may however require some careful engineering to make it work efficiently, considering that random access to the raster (which is too large to be stored in RAM) can be quite slow and already retrieving a single value for each location became a training performance bottleneck in our experiments.

Regardless, this is in fact a modification that might improve the results even for high-accuracy location data, since information about the neighborhood will likely be useful anyway. We now mention this in the discussion.

An important positive element of the work is the open-source release of the software, although I have not had the opportunity to try it.

Overall, the manuscript is clearly written and neatly presented.

# A few specific comments

L70-73: I don't agree that step selection functions are *the* approach to analyse animal trajectories. My own feeling is that other methods such as Hidden Markov Models or regression of trajectory parameters against environmental covariates are *at least* equally common, if not more so. To avoid this statement (which I think is debatable), consider: "Step-selection function (SSF) models, which compare actual movement steps with realistic candidate ones, are routinely used to infer and quantify the effect of environmental variables, such as land cover or temperature, on animal trajectories".

Thank you for the suggestion. We have updated the paper with your formulation.

L105-107: I assume the time-window is arbitrary, which was one of your criticisms of current methods for incorporating previous context ("familiarity") in step selection functions.

The time-window chosen is indeed arbitrary. However, our ‘criticism’ (in our mind it was more identifying where there is room for improvement) of current methods was not so much on the arbitrariness of the time window (as there will rarely be enough information to make an a priori informed choice that couldn’t be questioned), but rather on the coarseness of the variables used. We had written: ‘familiarity is usually incorporated into SSF models using a previously visited yes/no variable, or a time-spent variable, often calculated over an arbitrary time window’. The benefit of the approach we propose is to let the model find the optimal way to use past trajectory data to improve the movement’s predictions.

Table 1: Add column name abbreviations to the table caption. Describe 'Section' part of the table (training, validation, test) in the caption.

Done.

L150: '408 observations'. Also, translate this approximately to a real duration in days?

Done.

L150-153: How is the split assigned (what proportions)?

We assign 4 % of individuals to the validation section and 5 % to test. We have added a footnote explaining this in detail.

L156-158: What is the reason for doing this?

We have added a footnote explaining this.

L159: I think the taxon vectors will need some further explanation. Is the vector approximating the taxonomic relationships? Okay, I see this information down on L171-172. I suggest moving this information (or something like it) up to the beginning of the paragraph, to immediately give readers the context. Although, I still wonder if the vector embedding is capturing the actual taxonomic relationship, or only the 'semantic similarity' [L164-165] (in other words, how similar given Wikipedia entries are). The PCA figure, for example, shows that the embedding is okay at class level, but not very good at order level (no clear clustering of orders within classes). Also, the Spearman correlation (= 0.68, L 170) is not great. Could there be a different way to embed taxonomy? In understand that there may not be, and this is a minor point.

Wikipedia2Vec embeddings are not based purely on the text of the articles, but also on the link structure of Wikipedia, and for this reason, they are able to capture semantic *relatedness* of concepts. We have improved the wording of the paper on L167–171 to make it clearer that we mean semantic relatedness of entities (extending to taxonomic and other relationships between species) rather than just "superficial" similarity.

While we agree that the correlation is not that high, note that the number of common phylogenetic ancestors was simply chosen as a somewhat crude but simple-to-compute feature to at least partially validate our approach, and not as an ideal target. There are probably other properties that the Wikipedia2Vec vectors encode about the species that might help the model make a better prediction. Also, we use the embeddings simply as one of several input features, without the need for it to be a perfect representation of the species.

L184: 'resample' rather than 'sample'?

We have rephrased this as: "and use linear interpolation when retrieving the values by location".

L189: From this dataset (and Figure 1), I gather that no marine trajectories are included. Were these explicitly excluded using a filter at some point?

No, we include all GPS data that was publicly available under a CC license and fit the criteria that we mention. There are some trajectories that are at least partly over the sea surface; for these, the land cover data will indicate a missing value, which we represent using a special learnable embedding as written on L235–237.

L312. Perhaps start a new paragraph here.

Done.

Figure 7: The labels ‘bioclim’ are not informative.

We have added a table of the bioclimatic variables to the appendix and referenced it from the figure caption.

L502-505: I’m not sure I completely understood this point, but it seems to be quite relevant give our wish to make inference from the model. So, currently you present no inference for features that were relevant in the learned trajectory, only for the next predicted step? How are the two related to one another?

Indeed, we only perform the feature importance analysis on the candidate features, which is equivalent to what has been done traditionally in SSF analysis. Similar experiments could be performed on the conditioning (past trajectory) features, however the trick of shuffling the feature's values among the candidates (whose locations are conditionally independent given the past) does not work in that case, so a different ablation strategy (e.g. shuffling across time or replacement by a constant value) would need to be explored. This could also be very interesting when combined with the context length analysis, e.g. to see to what extent the model is sensitive to long-term context on a feature-by-feature basis.

L537-540: I wonder, also, whether using higher temporal resolution data would reveal a second peak in context length, indicative of nested scale-patterns in the trajectories (e.g., fine temporal context nested within a longer context).

We agree and had thought about this too. We think this might indeed be observed if we had enough high-resolution data, which was not the case here. Given that it seems that it is a question that can arise when reading the paper, we have now included this idea in the discussion.