Erreurs d'annotation en deep learning, et analyse statistique des interactions prédateurs-proies en écologie.

Olivier Gimenez*

Résumé

Les interactions prédateurs-proies façonnent les communautés animales. Le piégeage photographique permet d'étudier ces relations à partir de l'identification des espèces sur les photographies prises en conditions naturelles. Pour annoter automatiquement de grandes quantités de photographies, le deep learning est de plus en plus utilisé. Deux difficultés se posent toutefois. Premièrement, la reconnaissance n'est pas parfaite, et les taux d'erreur ne sont pas nuls. Deuxièmement, le principal language du deep learning, `Python`, est peu connu des écologues. Dans cette communication, on se pose la question de l'effet du taux d'erreur dans l'identification des espèces sur l'inférence des interactions entre elles. L'analyse est faite entièrement sous `R`.
Mots-clefs : Ecologie Statistique – Deep Learning.

Motivation

Nous illustrons la reconnaissance d'espèces animales sur des photos à partir du deep learning, et l'application à l'étude des interactions prédateurs-proies en écologie. Les relations prédateur-proies sont des interactions entre espèces qui façonnent les communautés de grands mammifères. Il s'agit d'un travail en collaboration avec Anna Chaine et Maëlis Kervellec, Vincent Miele, les collègues de l'Office Français de la Biodiversité (OFB) et des fédérations de chasse de l'Ain et du Jura.

Les collègues de l'OFB et des fédérations de chasse collectent des données dans l'habitat naturel des espèces qui nous intéressent, grâce à des pièges photos laissés à des endroits stratégiques. Il s'agit ici du lynx et de ses proies le chevreuil et le chamois. La méthode est non-invasive, autrement dit on n'a pas besoin de capturer physiquement les animaux. La difficulté est que l'on se retrouve avec des grandes quantités de photos auxquelles il faut associer une étiquette espèce.

C'est là qu'entre en jeu le deep learning, de plus en plus utilisé en écologie, voir par exemple Christin, Hervet, and Lecomte (2019). L'idée est de nourrir les algorithmes avec des photos en entrée pour en sortie récupérer l'espèce qui se trouve sur la photo. Nous avons utilisé la librairie fast-ai qui repose sur le language Python et sa librairie Pytorch. Un avantage de cette librairie est qu'elle vient avec un package `R fastai` qui propose plusieurs fonctions pour l'utiliser.

Quels sont les résultats obtenus? Nous avons d'abord fait du transfer learning sur un site d'étude dans le Jura où nous avions des photos déjà étiquetées. Nous avons utilisé un modèle resnet50 déjà pré-entrainé. Nous arrivons à classifier le lynx, et ses proies, le chamoix et le chevreuil, avec un degré de certitude satisfaisant. Ensuite, nous avons utilisé le modèle pour étiqueter automatiquement des photos prises avec des pièges installés sur un autre site, dans l'Ain. Ces photos ont aussi été étiquetées à la main, on connait donc la vérité.

Sur la base du nombre de faux négatifs (une photo sur laquelle on a un lynx mais on prédit une autre espèce) et de faux positifs (une photo sur laquelle on n'a pas de lynx mais on prédit qu'il y en a un), les résultats sont peu satisfaisants. Toutefois, la question est de savoir si le manque de précision nuit à l'inférence des interactions prédateur-proie. Pour ce faire, on a utilisé des modèles statistiques qui permettent d'inférer les co-occurrences entre espèces en tenant compte de la difficulté de les détecter sur le terrain. Ce sont les modèles d'occupancy développés par Rota et al. (2016) et implémentés dans `R` par Fiske and Chandler (2011).

On obtient les probabilités de présence du lynx, conditionnellement à la présence ou absence de ses proies (Figure 1). Il y a un léger biais dans l'estimation de la probabilité de présence du lynx sachant la présence

*CNRS, olivier.gimenez@cefe.cnrs.fr

de ses deux proies favorites quand on se fie à l'étiquetage automatique des photos. Etant donné que les différences ne sont pas énormes, l'écologue pourra décider de les ignorer au regard du temps gagné par rapport à un étiquetage à la main. Maintenant le biais est plus important sur la probabilité de présence du lynx sachant la présence du chevreuil et l'absence du chamois qui elle est sous-estimée.

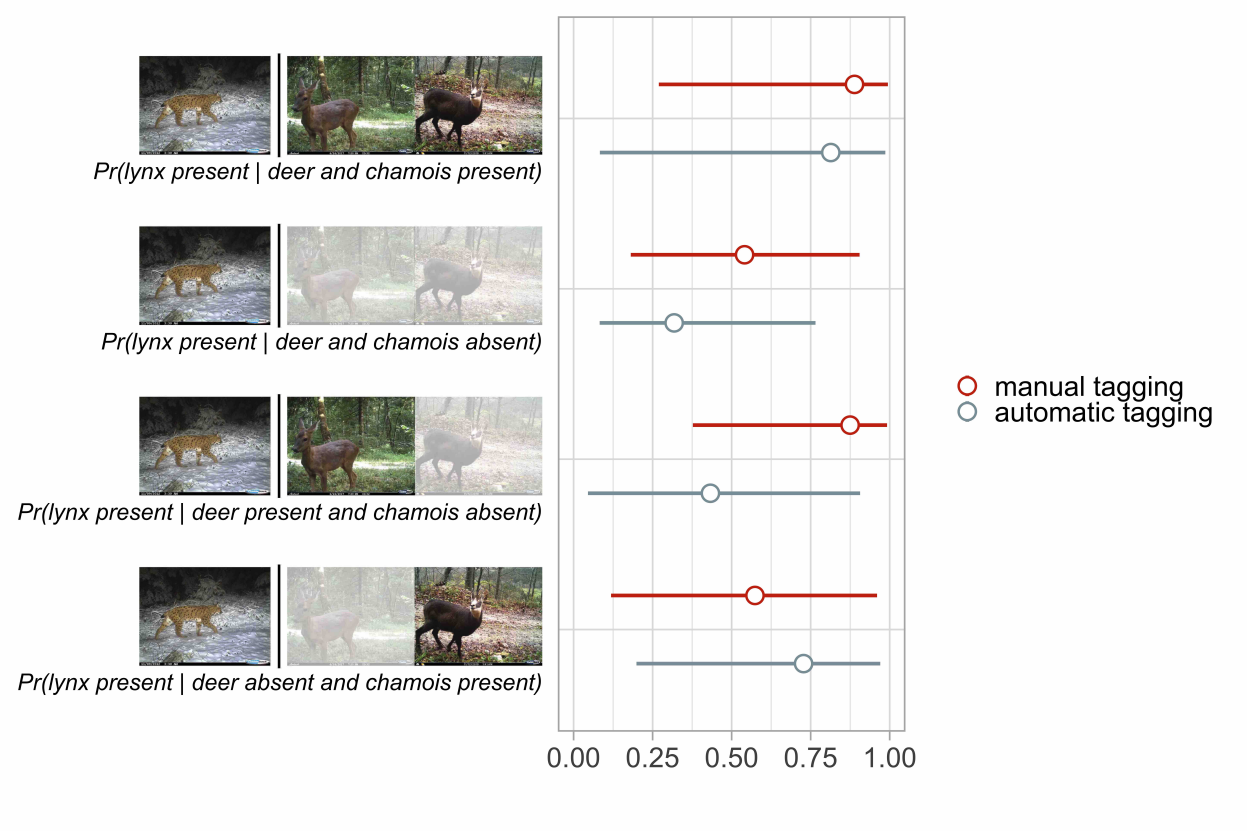

Figure 1: Probabilités de présence du lynx, conditionnellement à la présence ou absence de ses proies. En rouge, avec les photos étiquetées à la main. En gris-bleu, avec les photos étiquetées automatiquement.

En conclusion, l'utilisation d'un modèle entrainé sur un site pour prédire sur un autre site est délicate. Il est facile de se perdre dans les dédales du deep learning, mais il faut garder le cap de la question écologique, et on peut accepter des performances moyennes des algorithmes si le biais engendré sur les indicateurs écologiques est faible. Malgré tout, on peut faire mieux, et nous développons actuellement des modèles de distribution d'espèce qui prendrait à la fois en compte les interactions et les faux positifs et faux négatifs. Pour aller plus loin avec le deep learning et l'analyse d'images, nous renvoyons vers Miele, Dray, and Gimenez (2021).

Références

Christin, Sylvain, Éric Hervet, and Nicolas Lecomte. 2019. "Applications for Deep Learning in Ecology." *Methods in Ecology and Evolution* 10 (10): 1632–44.

Fiske, Ian, and Richard Chandler. 2011. "unmarked: An R Package for Fitting Hierarchical Models of Wildlife Occurrence and Abundance." *Journal of Statistical Software* 43 (10): 1–23.

Miele, Vincent, Stéphane Dray, and Olivier O. Gimenez. 2021. "Images, écologie et deep learning." *Regards sur la biodiversité*, February. https://hal.archives-ouvertes.fr/hal-03142486.

Rota, Christopher T., Marco A. R. Ferreira, Roland W. Kays, Tavis D. Forrester, Elizabeth L. Kalies, William J. McShea, Arielle W. Parsons, and Joshua J. Millspaugh. 2016. "A Multispecies Occupancy Model for Two or More Interacting Species." *Methods in Ecology and Evolution* 7 (10): 1164–73.