Self-supervised learning of deep visual representations

Mathilde Caron [1]

Cet article est le résumé de la thèse de doctorat [2] de l'auteur ayant reçu l'accessit du prix de thèse Gilles Kahn 2022 décerné par la Société informatique de France.

Les humains et de nombreux animaux peuvent percevoir le monde et le comprendre sans effort. L'espoir du domaine de la reconnaissance d'image, une branche de la vision par ordinateur, est de recréer cette capacité de perception visuelle par un ordinateur, c'est-à-dire par une « intelligence artificielle ». Les êtres vivants acquièrent une telle perception du monde visuel de manière autonome, c'est-à-dire sans l'intervention d'un superviseur externe leur disant explicitement où, quoi ou qui doit être vu. Cela suggère que la perception visuelle artificielle pourrait être acquise en l'absence de supervision experte, tout simplement en laissant un système observer de grandes quantités de données visuelles.

De nos jours, l'approche la plus prometteuse et la plus adoptée pour la reconnaissance d'image est l'apprentissage profond, une puissante branche d'algorithmes issue de l'apprentissage automatique (*machine learning*) qui consiste à laisser le programme découvrir (ou apprendre) par lui même comment résoudre un problème à partir d'une grande quantité d'exemples, au lieu de lui dicter en amont une série

1. Université Grenoble Alpes

2. https://www.theses.fr/2021GRALM066.

DOI:10.48556/SIF.1024.21.171

de règles. Le paradigme dominant en apprentissage profond est celui de l'apprentissage supervisé. Selon ce paradigme, chaque image doit venir avec une annotation décrivant son contenu et le programme apprend conjointement à partir des images et de leurs annotations. Les approches supervisées sont les plus adoptées en reconnaissance d'images aujourd'hui mais elles sont problématiques pour plusieurs raisons.

D'abord, comme évoqué plus haut, l'apprentissage supervisé n'est pas calqué sur la manière dont on a l'impression que les humains ou les animaux acquièrent la perception visuelle. Par exemple, les bébés sont capables de percevoir les éléments de leur environnement souvent bien avant de les nommer, ce qui suggère que l'apprentissage se fait plutôt par le biais de l'expérience et l'observation répétée.

Une autre limitation de l'apprentissage supervisé est le biais ou même les erreurs dans les annotations. En effet, l'acte d'annotation de données est ambigu dans le sens ou il y a souvent plusieurs manières d'annoter, sujettes à interprétation. Cela peut introduire des biais dans les systèmes qui utilisent ces annotations.

Finalement, la raison peut-être la plus pragmatique pour inciter à se débarrasser des données supervisées est le fait que les annotations sont souvent compliquées à acquérir. En effet, annoter des données requiert une expertise humaine, ce qui empêche le passage à l'échelle des algorithmes et est donc limitant dans les domaines où l'accès à ces annotations est coûteux, difficile voire impossible.

Dans la thèse, nous présentons différentes contributions au domaine en plein essor de l'apprentissage auto-supervisé de représentations visuelles. Nous commençons par étudier une catégorie prometteuse d'approches auto-supervisées, à savoir le *clustering* profond, qui permet d'entraîner des réseaux de neurones tout en trouvant des groupes d'images visuellement similaires dans une collection de données. Nous identifions ensuite les limites de ces méthodes de *clustering* profond, telles que la difficulté à passer à l'échelle, ou le fait qu'elles sont souvent sujettes à des solutions triviales. En conséquence, nous proposons de nouvelles méthodes auto-supervisées qui surpassent leurs homologues supervisées sur plusieurs *benchmarks* et présentent des propriétés intéressantes. Par exemple, les réseaux auto-supervisés ainsi obtenus contiennent des représentations génériques qui se transposent bien pour résoudre diverses tâches sur d'autres ensembles de données. Ils contiennent également des informations explicites sur la segmentation sémantique d'une image, obtenue de façon complètement non supervisée. Finalement, nous évaluons également nos modèles auto-supervisés sur des données brutes, en les entraînant sur des centaines de millions d'images non supervisées prises aléatoirement sur Internet.