Lecture sur les lèvres : l'IA fait désormais mieux que l'homme

L'IA plus performante que les spécialistes de la lecture sur les lèvres ? C'est en tout cas que laisse entendre une étude menée conjointement par Google DeepMind (une entreprise britannique rachetée par Google en 2014 et à l'origine d'AlphaGo) et l'université d'Oxford. A partir de 5 000 heures de programme de la chaîne britannique BBC, les chercheurs ont créé une application de lecture sur les lèvres qui dépassent les performances des spécialistes humains de cette discipline. Au total, cette masse de vidéos, issues de 4 émissions, renferment 118 000 phrases.

Les chercheurs ont commencé par entraîner leur IA sur les programmes diffusés entre 2010 et 2015. Puis ont testé les performances des algorithmes ainsi préparés sur un lot de programmes diffusés entre mars et septembre 2016. Résultat : l'IA a reconnu **sans erreur 46,8 % des mots** prononcés par les personnes présentes à l'image sur ce jeu de données. Y compris des phrases entières. L'étude signale qu'une bonne partie des erreurs relevées tiennent en réalité à de légères déformations de certains mots (comme l'absence d'un s à la fin d'un mot, un élément très difficile à déceler dans de nombreux cas en anglais).

Deep Learning et jeu de données

Surtout, l'IA fait ici mieux qu'un spécialiste humain du sujet, travaillant pour la BBC. Sur la base d'un jeu de 200 vidéos issues du jeu de données soumis au système WLAS (Watch, Listen, Attend and Spell), cet expert affichant environ 10 années d'expérience parvient à déchiffrer **moins d'un quart des mots prononcés** sans erreur, et ce même s'il avait le loisir de regarder plusieurs fois les vidéos. Un taux en ligne avec d'autres expériences sur le sujet, notent les chercheurs.

Selon les chercheurs, il s'agit là d'une avancée importante par rapport aux autres recherches menées sur le sujet, une percée reposant sur l'emploi de modèles de réseaux neuronaux et l'exploitation d'un jeu de données très étendu (or, les ambiguïtés de la lecture labiale ne peuvent être levées que par une bonne compréhension du contexte). Récemment, un autre système de Deep Learning baptisé LipNet – lui aussi développé par l'université d'Oxford – a également dépassé les performances humaines en matière de lecture sur les lèvres. Mais l'expérience (connue sous le nom de GRID) se limitait à un vocabulaire de 51 mots… là où WLAS s'attaque à un corpus de 17 500 mots ! Qui plus est, les données fournies par la BBC renferment des discours réels prononcés par

différents individus avec des structures de phrase très variées, alors que GRID se basait sur des phrases reproduisant un modèle bien défini.

Son et image synchronisés… par IA

« *Une machine qui peut lire sur les lèvres ouvre la voie à de multiples applications : dictée d'instructions ou de messages à un téléphone dans un environnement bruyant, transcription ou doublage de films sans son, compréhension de discours où plusieurs personnes s'expriment ou, plus généralement, amélioration de la performance de la reconnaissance vocale* », écrivent les chercheurs Joo Son Chung, Andrew Senior, Oriol Vinyals et Andrew Zisserman dans leur étude.

Signalons au passage que, avant de s'attaquer au déchiffrement des phrases prononcées par les personnes filmées, les chercheurs ont également exploité le Machine Learning pour préparer les données. L'enjeu ? Recaler le son et l'image sur certaines vidéos, une étape indispensable pour assurer la phase d'apprentissage de WLAS. Les 5 000 heures de vidéos fournies par la BBC ont ainsi été passées au crible afin de resynchroniser automatiquement les clips qui auraient sinon nui aux résultats de l'expérience.

A lire aussi :

Une IA pour piloter le datacenter du futur ?

Carnegie Mellon se penche (à son tour) sur l'éthique de l'IA

Une IA est capable de concevoir son propre chiffrement