Commencer tout de suite non plus. Ouais, tant que tout le monde se connecte tranquillement. Normalement, pour le la signature, vous êtes tous autonomes. Il y a ce matin et qui ont pas été notés aussi n'hésitez pas pour la signature, il nous faut le code. Si vous le faites, signature code. Je le mets dans le chat ici, ça vous va ? Oui, très bien, merci. P. Devrait tous pouvoir voir mon écran si tout va bien. Je suis désolé, j'utilise pas du tout teams pour faire des visions. Non, je pense que ça devrait fonctionner. Ok alors est-ce que tout le monde est là ?\*\* \*\*\*\*\*\*. Alors ? Qu'est-ce qui était pas là ce matin ? Thomas, Fanny Rami. Ok, normalement vous êtes tous une présents, n'hésitez pas à vérifier. Hop donc j'avais juste. Je veux juste Dylan et Louis des Clopes par leur sens ce matin. Dites schotz où on Recale comme vous voulez. Ok. Tante. C'est rigolo, on est 35YA 39 consigne, c'est pas mal. Vous êtes très fort. Hop. Bon, c'est pas grave, on s'en fiche. Du coup tout le monde est là. Enfin tout le monde est, tout le monde est prêt, on est bien installé. Ouais yes parfait. Bon, on arrive sur la fin du cours. Doucement hein on. On a vu pas mal de choses jusqu'ici, on a commencé avec le. Avec le RL Tabulaire. On a vu les méthodes de base globalement, quand on avait des tables comme ce qu'on a fait ce matin avec les les que table. Ensuite, on a vu que. On a vu qu'on pouvait faire des méthodes de danse quand on avait des modèles plus compliqués. Donc, notamment grâce à la chaleur de le dernier cours, grâce au théorème du policier gardiens. Et donc là globalement on a fait le tour des grands paradigmes qu'on peut trouver en RL. On a fait les liens un peu entre les 2 avec les histoires de les histoires de tiling. Comment est ce qu'on peut traduire un un environnement en environnement continue à en dans un environnement discret et globalement le but de la séance d'aujourd'hui. Ça va être de. De vous montrer que maintenant vous avez toutes les clés pour comprendre les dernières méthodes qui sont utilisées dans en RL, dans dans la vraie vie quoi. Donc aujourd'hui on va avoir plusieurs exemples, on va voir pour ceux qui connaissent donc Google Deep mind je pense que tu me connais mais le leur dernier algo donc quand ils ont battu. C'était assez récemment, c'était 2016, je me trompe pas, on verra dans le cours, mais ils ont battu le champion du monde de Go donc c'était un grand, un grand événement déjà que dit c'était un grand événement. Quand on a battu le champion d'échec Lego c'est encore plus dur, donc quand ils ont réussi à faire ça, c'est pour ça, c'est pour ça fait beaucoup de beau. Je sais pas si ça, si ça parle à certains, mais donc on va voir que en fait maintenant. Vous avez les bases pour comprendre comment ils ont fait pour faire ça, comment ils ont fait pour pour entraîner leur leur algorithme Alpha Go. Et on va voir les les les suivantes Itérations on va, on va voir les itérations suivantes jusqu'à arriver à effet 0 qui est un peu le dernier modèle qu'on appelle le modèle base. Et puis. Et puis, et puis TAC TAC TAC. Ouais, et ensuite, dans la dernière séance, on verra quelques applications concrètes. Qui sont utilisés. Récemment ? Pour l'instant je vous en ai mis 2 mais ça je je suis-je suis ouvert à des suggestions si jamais vous avez envie de plonger un peu plus dans un dans des détails. Mais donc là pour l'instant je vais vous parler de du RHF donc qui a été utilisé par open AI pour entraîner charge et donc c'est ils utilisent une méthode qu'on va voir aujourd'hui, un algorithme aujourd'hui. Mais du coup en dernière séance on verra. Alors concrètement, comment ils ont fait et pourquoi est-ce que c'était intéressant d'utiliser cette méthode là qui s'appelle PO ? Et on va voir un pareil, un petit reportage sur. En en labo de de Zurich qui est super fort et qu'on entraînait des drones pour faire des courses drone Racing et donc ils ont réussi à entraîner des drones à sur des courses assez impressionnantes vont comparer. On verra comment ce qu'ils ont fait et et le but c'est vraiment de vous montrer que maintenant vous avez toutes les clés, en comprendre vraiment en profondeur ce qui a été fait. Voilà euh et donc pour la dernière séance, je suis ouvert aux suggestions. Si jamais vous avez un un cas d'usage, vous savez qui a du R, qui est utilisé sous le tapis mais que vous vous vous avez pas trop compris, vous êtes pas trop pris le temps de de vous intéresser. Je peux éventuellement essayer d'en parler donc n'hésitez pas à me à proposer. Et puis la séance prochaine sera surtout aussi, donc le le cours, ce sera assez court et ce sera surtout à un moment où vous pourrez me poser vos questions pour le projet, pour que je puisse vous aider un peu, vous guider si j'avais, s'il y a des erreurs, et cetera. Voilà pour le planning de fin de cours et donc pour aujourd'hui on fait ce cours là, on a 1TP où on va voir en pratique comment est-ce qu'on fait pour entraîner un modèle de RL, donc en utilisant des outils un peu plus évolués que numpy et et panda, on va voir donc dans des BB, je pense que vous devez connaître, mais voilà comment est-ce qu'on fait et comment en pratique on fait pour, pour monitorer un entraînement, quoi. Et puis je vous présenterai le projet à la fin et vous pourrez. Bah vous pourrez vous lancer dessus. Donc le but ça va être de le rendre pour plus tard. Je vous donnerai les dates. Je vous donnerai les dates, c'est tout clair pour tout le monde. Oui, Monsieur parfait. Super. Et si y A quoi que ce soit qui va pas si vous m'entendez plus bref je vois pas forcément tout ce qui se passe. Donc n'hésitez pas à m'interrompre même si vous avez des questions du coup hein hein, n'hésitez pas à m'interrompre et plutôt en vocal. Du coup je sais pas si je vais voir facilement les réactions normalement j'ai le chat sous le sous les yeux. Eh Ben tiens, madelaine. Guylaine, c'est été présente ce matin, c'est ça ? C'est pour la présence de ce matin. Probablement bon, je pars plus tard, OK, donc. On arrive aux méthodes de Deep run for Smart Learning. Tac. Ce qui va se passer ? Je j'aime beaucoup ce même qui qui dit beaucoup de ce qui s'est passé des dernières années et. Avant on faisait beaucoup de de trucs très très compliqués pour faire du RL ou pour faire des modèles statistiques de manière générale pour ceux qui ont touché un peu en cours de ml avec les dimensions domicile et. Et tout ça là ? Et bah c'est quand même vraiment galère alors qu'aujourd'hui Ben on a des raisons de drones donc on va essayer de faire des trucs avec. Hop, même si on passe un peu pour des clowns, donc là je vous ai mis un petit. Je vous ai mis un petit arbre qui. Qui ? Et merci Léonard de de plus à nous même. J'ai mis un petit arbre qui qui résume un peu tous les tous les algos de de R qu'on peut trouver. Et et notamment donc vous devriez pouvoir reconnaître reconnaître certains qu'on a déjà vu. Donc pour l'instant on a fait on a pas fait de modèle base donc on reste à gauche de l'arbre. Ok on a vu des des algos value base donc V un et ça ça vous parle normalement j'espère. Sinon vous êtes un peu dans la \*\*\*\*\* pour le partiel. Et l'autre côté ensuite on a vu des algans policy base donc qui se basent en général sur les politiques alliance donc on a vu dans le dernier cours renforce ok acteur critique donc atroce et à 3 si et pas TRP aussi je me trompe pas. Donc globalement là-dessus vous avez tous les noms des algos les plus importants qu'on utilise en RL donc c'est assez bien, c'est un c'est un un bon résumé à connaître. Et donc pour aujourd'hui ? On va voir tout cela. Donc au niveau of policy, comment est ce qu'on peut faire en of policy pour apprendre pour apprendre des algos de RL donc des des PG et sac peut-être que y en a qui ont qui ont connaissent déjà on va voir côté on policier. Donc les grands classiques de \*\*\*\* et tiers pivot, ça c'est des des alcools très importants. Donc notamment Tipi, Euh, \*\*\*\* au pardon, euh. Qu'on va qu'on va utiliser là pour entraîner un modèle. Et pipio qui est notamment utilisé par un élève pour le pour le relief. Pardon. Et de l'autre côté on va aussi voir les derniers algos du coup model base à partir de à partir de la phase enfin d'Alpago vous allez voir mais modèle base Pour rappel bah tiens qui est-ce qui pourrait me qui peut me rappeler ce que c'est ? Un modèle, enfin qu'un algorithme modèle base. Sur les pages à vous, on voit, on voit même pas vos têtes là. Soyez pas timide. C'est quoi différence entre modèle base et et modèle free ? Je l'ai redit à certains d'entre vous ce matin. Je suis pas sûr, mais c'est pas en rapport avec l'environnement parce que on défigurerai. Ouais, c'est un rapport. C'est un rapport, Ouais. En rapport avec l'environnement, le fait de modéliser l'environnement. C'est pas aussi avec l'ordinateur d'un ami de programme parce que ce qu'on utilise pas. Non, pas vraiment, donnée programming, c'est dans le dans les premiers algos, là, dans dans samsa éventuellement. Et c'est quand on est en tabulaire. Dynamique programming, c'est vous retenez divisé pour mieux régner, donc on prend un gros problème, pardon c'est vrai, c'est vrai c'est ça ? Donc c'est le dernier problème, c'est quand on sépare la table, donc typiquement le labyrinthe. Là si vous vous rappelez, on va le séparer en plein de petits problèmes, on va propager du gros problème en plein de petits problèmes, donc on prend page d'une case à l'autre pour arriver à résoudre. Donc ici non, en fait c'est une autre référence model base version simulation. Mais c'est bien de rappeler sa fait un petit un petit rappel pour tout le monde. On a un. Le modalité est déjà l'environnement et un autre, il apprend sur le tas. Ouais, y a un peu de ça. Il y a un peu de ça. L'idée derrière, c'est que. De manière générale en RL vous pouvez pas modéliser votre environnement parce qu'il est trop compliqué donc vous pouvez pas typiquement faire un bac propre sur votre environnement. Si l'environnement c'est le monde réel pour un robot bah vous pouvez pas faire monde point bac point backward quoi. Donc ce que vous faites c'est vous pouvez calculer vos gradients ou de manière générale Modélisée uniquement votre votre agent à vous que votre policier OK donc on dit que dans ce cas-là si l'algo il utilise pas de modèle du monde ? Typiquement on va pas chercher à prédire ce qui va suivre dans dans dans notre notre environnement. Bah dans ce cas on modélise pas l'environnement c'est ça en fait nous nous êtres humains, ce qu'on fait un peu dans la vie de tous les jours quand vous conduisez une voiture ou quand vous faites du vélo. Bah votre votre action est dépend pas uniquement de l'État dans lequel vous vous trouvez. Vous allez aussi chercher à prédire ce qui va se passer plus tard, vous voyez le mur devant vous ? Et bah si je décide de freiner c'est parce que je me dis Bah si là dans ma représentation mentale je vais me prendre le mur parce que si je freine pas. Et donc voilà dans ce que je fais, dans enfin ce que je fais quand je quand je réfléchis comme ça, c'est que je modélise mon environnement. Donc je je cherche à prédire ce qui va se passer. Ok donc ça en RL. C'est pas simple, ça veut dire qu'il faut que vous, il faut que vous ayez un modèle du monde qui qui vous permet de faire des prédictions. Et donc modéliser le monde. Bah là on modélisé le monde modèle base R donc on cherche à à modéliser l'environnement là ou en modèle Free. On a juste une mobilisation pour notre agent et il se base uniquement sur des rewards et et pas il va pas chercher à prédire l'environnement. Ok c'est vraiment et ce que vous modélisez l'environnement ou pas. Voilà et donc on va voir des exemples de d'Algol Model-based aujourd'hui. Toc toc. Ce que j'attend. Là dessus ? Tac. Très bien du coup. On va commencer par. On va commencer, ouais, TAC. On va commencer par les algos policiers, OK. Donc c'est ce qu'on a fait jusqu'ici. Enfin, pardon du on est sûr du modèle Free et on va on va passer à droite Là, vous voyez ? Of policier, of policier, ça veut dire quoi ? Ça veut dire que on apprend avec une. On va apprendre avec une certaine policier et on va en déduire la police optimale, mais sans forcément l'utiliser pendant le. Pendant l'entraînement OK, au niveau format du cours, là on va on va moins entrer dans les détails parce que encore une fois le le but c'est de vous montrer que maintenant vous avez toutes les bases pour comprendre un peu tous les paradigmes qu'on va utiliser dans plein d'algorithmes. Et il y a tellement d'algorithmes que en fait ça sert à rien de de plonger dans les dans les détails techniques de chacun. Mais je pense que l'important, c'est que. Ces algorithmes, vous les connaissiez ? Vous sachiez qu'il est qu'ils existent, pourquoi ? Enfin pourquoi est-ce que ils peuvent être utilisés ? Et. Et dans quel cas ? Ce que faudra y penser si jamais vous en aurez besoin. Voilà voilà donc globalement on va passer assez vite sur tous les petits algos que j'ai que j'ai entouré ici avec le but de vous d'essayer de vous faire comprendre. Bah c'est pas l'intuition derrière. Pourquoi ce qui a été développé et vous montrer que que si jamais vous vous deviez le réimplémenter ou ou l'étudier. Bah ce serait pas un problème, voilà. Donc on commence par T 3T 3 pour tout indy Deep déterministe pour les Syriens. Donc police agence c'est ce qu'on a vu au cours dernier donc c'est pour ça que c'est logique de le de le voir maintenant. Mais mais. Il y a beaucoup, il y a, c'est un peu comme toujours, on rajoute plein de plein de de petites astuces dans le modèle pour que. Pourquoi ça fonctionne ? Tout simplement. OK, donc. Comme. Ce qu'il y a de particulier sur t 3 on donc on est acteur critique OK polyvalence souvent on est on est du acteur critique parce que pour avoir une bac propre sur la police, si souvent il nous faut une estimation de la de la fonction valeur ou de la fonction action valeur et donc on va avoir un critique. Ok et. On est policier et on est modèle Free toujours OK donc ça c'est les les buzzwords à connaître pour un peu avoir une idée de ce qui va se passer. OK donc là on est of policier donc à priori of policier on va avoir un replay buffer ou avant on va avoir un moyen d'explorer notre environnement et d'apprendre la policie à posteriori. Ok parce que on est of policy, ça veut dire que a priori on on. On on n'améliore pas notre politique au fur et à mesure de enfin comment dire, n'utilise pas la politique qu'on est en train d'apprendre pendant l'entraînement, c'est comme ça que je le que je voulais formuler. Donc en effet on voit on utilise bien un replay buffer pour donc là en l'occurrence c'est ça. Nous ça nous aide encore plus pour enfin avoir plus d'efficacité, c'est-à-dire si on a un environnement qui coûte cher. Faire tourner ? Bah autant retenir les transitions qu'on a fait pour les réutiliser, c'est ce qu'on a vu ce matin aussi. Ce matin. Ce matin on on a vu que quand on utilisait un replay buffer. Bah le modèle il pouvait apprendre à partir de plus de données et donc on a vu qu'il convenait plus vite. C'est-à-dire que notre notre robot une fois qu'il a compris que euh en allant une fois seulement tout droit il avait une grande récompense et bah le fait de le garder dans le replay buffer ça nous a permis d'apprendre plus vite à davantage utiliser cette action là quand c'était possible, là où si on avait pas de replay. On a peut-être pu l'oublier plus rapidement. Et donc voilà, je vous ai mis les 3 les 3 comment dire. Astuces principales qui sont utilisées dans les étroits qui sont le clip double ce learning donc là c'est le. Comment est-ce que j'expliquerai ça, ce que je peux, vous voyez à droite, là, en dans la, dans la boîte bleue, en 5 ? Vous pouvez voir que vous allez en gros faire un clip sur vos valeurs pour éviter que. De faire des gros, des très grosses lunettes. Ok. Et donc vous voyez votre épilogue là, il il permet de limiter, on va dire le le, la quantité de changement qu'on va faire à chaque étape. Et on va voir où ça on va, on va en reparler juste après. Je vais pas m'attarder beaucoup de suite le Delight Policy Update ça c'est un peu le principe du du replay buffer, c'est à dire que on va pas updater notre tout de suite mais on va attendre de d'avoir un peu de données quoi. On va beaucoup explorer, on garde tout dans le. Le buffer et de temps en temps, PAF. On utilise notre note donnée pour calculer, enfin pour optimiser notre critique et ensuite optimiser notre acteur. Donc notre acteur il y a notre police, OK ? Et donc TARGET polishing pareil, c'est dans la manière de de prendre nos données, on va essayer de faire les moyennes pour éviter pareil de d'avoir une variance trop grosse, ok, avec toujours l'idée derrière du compromis biais variance qui devrait vous parler je l'espère. Quand on utilise des méthodes typiquement TD versus sur Monte Carlo, Ouais, ça vous rappelle des choses ? Alors oui, Monsieur. Ouais, une question. Non du tout. Je dis juste oui, ça rappelle des trucs très bien. Merci OK. Ok OK donc pour t'aider 3 il y a des questions. C'est normal si vous comprenez pas dans tous les détails mais bah l'idée c'est que c'est de comprendre l'idée générale et pourquoi est-ce qu'on a utilisé telle ou telle astuce ? Très bien. On passe donc à sac pour soft actor critique OK donc là on est sur un acteur critique. OK of policy, toujours pareil modèle Free, toujours pareil. Ok donc là celui-là il s'utilise pour des des des actions qui sont continues. C'est pour ça qu'on appelle ça soft acteur critique. Je sais pas si vous. Bah là pour l'instant jusqu'ici on a vu des actions qui étaient plutôt discrètes. Notre robot, là notre notre DQN il a un nombre fini de d'output donc on a un nombre fini d'actions qui sont possibles. Ok Bah dans le cas où notre. Notre réseau il doit être put une valeur continue. On va pouvoir utiliser soft acteur critique. Ok donc donc on combine des facteurs critiques avec un. En principe de maximum de d'entropie maximale, entropie, ça va permettre de de. De quantifier la la diversité des. Cela, comment dire ? De la distribution, pardon, de de nos sorties. Normalement l'entropie pareil. C'est un terme qui devrait vous parler et donc c'est beaucoup utilisé pour faire de la régularisation avec des réseaux de neurones. Donc là on l'utilise dans le cadre du renforcement learning pour avoir une régularisation sur les prédictions de notre acteur et notre critique. Ok donc. On utilise notre technique, on va, on va avoir 2 réseaux de neurones différents pour le critique. Ok. Euh et donc ? Pareil vous pouvez, c'est vraiment le but de garder les ces boulets de ponza en en tête pour se dire OK j'ai un problème qui est dans le dimension bah peut être on va regarder acteur de soft, acteur critique quoi et donc en pratique en effet. Donc vous avez 2 critiques donc qui vont être capturer des des choses différentes dans votre espace et donc c'est en général ça marche bien pour. Font des tâches de contrôle continu. Encore une fois je pense que ce qui est bien c'est de d'associer dans votre tête le soft avec quelque chose de continu là où je sais pas 1DQN ou ou 1T 3. Ça va être des actions plutôt discrètes. Ok. Voilà la ensuite des heuristiques très générales de more more simple Edition and PO donc. Pourquoi ? Parce que, encore une fois, très, très, très heuristique. Ça semble pas le chien, ça veut dire qu'on va apprendre avec moins de données. Et donc là peut être que le fait que ce soit continu et que vous avez 2, vous avez 2 réseaux critiques peut aider à apprendre plus vite. Mais bon, vous voyez un général devant donc c'est pas c'est pas, c'est pas quantitatif comme beaucoup de choses qu'on verra aujourd'hui. Vraiment. L'idée c'est de vous parler globalement des algorithmes et de. Pour vous les connaissiez ou qu'ils aient déjà entendu parler quoi pour éventuellement y penser le jour où vous aurez vraiment besoin. Donc voilà. Donc globalement là on commence à taper dans des algos qui sont State of the art donc c'est les meilleurs algos possibles que vous puissiez utiliser dans le cas de certaines tâches de contrôle continu. OK. Donc pensez y gardez bien en tête encore une fois je pense que il faut garder cette image là en tête. Vous avez vraiment les les algos les plus les plus importants. Enfin peut-être en dessous de l 0. Je rajouterai aussi, on rajoutera tout à l'heure OK. Ensuite, donc, c'est tout pour les offres policiers. On on passe sur les algos, on policier. Avec un enchaînement qu'on va voir de tirs. Pio vert Pipo OK. Ce qu'en gros dans l'idée PO c'est tpo avec quelques améliorations. Donc \*\*\*\* o c'est un peu notre objectif du jour en pour les algos modèle Free en tout cas. Donc on commence par dire Pio donc tiers Pio je l'ai pas si c'est ça non je l'ai pas mis zut on tire pile si vous voulez noter c'est pour trust Regin pour aussi optimization. Donc trust relation ça veut dire on va pas se faire confiance. Faire des trous des trop gros pas sur le côté qu'avec ici encore une image pour montrer que dans le cas où on a des des paysages de l'os qui sont très qui sont très compliqués, bah on n'a pas envie de faire des trop gros. Pas parce que notre gradient il va avoir du sens mais sur une zone très faible autour de notre de notre point. Donc ici là on si on fait un trop gros pas on risque de tomber dans le vide donc on on aimerait pas. Donc tragen ça veut dire qu'on va, ça veut dire que pour notre problème, on sait que. On peut se permettre au maximum de. On peut se permettre au maximum de bouger notre policier d'un d'un epsilon de parce que on va connaître notre espace ou parce que on va faire une de l'épée à 20h00 tuning, pour éviter justement de enfin pour trouver la meilleure meilleure entraînement. Donc on est toujours sur 111 algo. Qui est acteur critique ? Cette fois, on est policier, donc on va apprendre la Policie online. D'accord, et c'est aussi pour ça, pour faire attention, parce que quand on est un policier, globalement, on va explorer notre environnement avec un petit ingrédient ou ce genre de choses. Donc on a quelque chose qui est stable en termes de normes, de de comportement dans le dans l'environnement. Mais là, comme on est, on policy. Ça veut dire que votre robot il va vraiment apprendre avec la politique qu'il est en train d'apprendre. Donc il va se déplacer en utilisant sa sa policier en question. Et donc c'est pour ça qu'on veut éviter de faire des trop grosses updates parce que on est en RL. Donc en fait tout est tout va très vite parce que une petite action peut me faire des vies un petit peu meilleure avec toi. Mais celle d'après peut me faire carrément dévié et me faire tomber ici, vous voyez en violet ? Certes, on a fait une un grand pas dans la bonne direction. Par contre, en fait, on s'est décalé un tout petit peu de la trajectoire optimale, mais ce tout petit peu a suffi pour nous faire tomber en vie. Bref. Donc il est déter bio, c'est avoir une frustration. Donc frustration policier, optimisation ou on va, on va limiter les étapes qu'on va faire. Dans notre. Dans ton entrainement, OK. Et donc pour ceux qui connaissent laquelle ? Divergence ? Qui connaît laquelle ? Divergence ? On l'a fait. Ouais. Laquelle on se rapidement dans ouais, très bien. Laquelle divergence ? Ça mesure. C'est une mesure pour pour quantifier la différence entre 2 distributions OK. Et donc ce ces ces cette métrique là qu'on va utiliser pour éviter de trop changer la distribution au niveau de la police. Ok, c'est un policier, elle est prend en input tous les États possibles, enfin un État possible, elle nous sort une distribution d'actions possible, enfin de distribution d'action, c'est à dire Tu vas faire telle action avec 0, 9% de probabilité et telle autre, et cetera. Et donc on utilise une Call divergence pour mesurer entre 2 étapes enfin entre 2 policiers différentes. Donc si je fais mon update de policier et Ben je vais dire Ah bah là elle est trop différente de la politique précédente donc on est plus dans ma restriction donc je je quitte je vais équiper l update pour que ça ça bouge pas trop quoi OK donc laquelle ? Divergence est notre mesure de changement de notre policier. Ça devrait être 40. Donc voilà l'idée principale de tertio, c'est vraiment cette cette question de trust légion. Pour ça je les mets en en gras et c'est une méthode qui est globalement hyper importante, même en en Deep de manière générale fait beaucoup de Gradin, clipping. Je sais pas si il y en a beaucoup. Enfin, normalement, vous devez tous dû entraîner un réseau de neurones un moment donné de votre vie. Bah parmi les méthodes qu'on peut utiliser pour éviter que notre entraînement parte en cacahuète, c'est mettre un gradient de clipping. Vous allez utiliser par exemple Adam comme tout le monde. Adam lui, il va augmenter le temps de votre quand de la direction dans laquelle vous allez être fait comme le modèle égal beaucoup en performance. Adam il va très vite augmenter beaucoup le. La norme de votre update et c'est pour ça que que Adam converge aussi vite est aussi efficace. Mais en revanche, ça veut dire que parfois vous allez, vous allez vraiment déconner quoi ? Et donc vous allez pouvoir dans certains cas mettre du gradation clipping pour éviter justement d'avoir un des objets qui sont trop grosses. Donc enfin dans mon cas, ouais vas-y, vous pouvez réexpliquer un clipping juste. Parce que c'est exactement. Ouais bah donc le gradient clipping quand vous êtes quand vous entraîner un réseau de neurones c'est je fais mon inférence. OK j'ai y égal à FDX OK avec sur mon réseau neurone et ensuite je fais. Elle est égale l'os de YOK. Enfin y type de un je calcule malos je fais un backward OK donc là à partir de ce moment-là vous avez les gradients sur tous les paramètres de votre modèle et c'est à partir de ce moment-là qu'on fait une descente de gradient on va dire les nouveaux, les nouveaux paramètres c'est les anciens paramètres moins. Alpha learning right, les fois les gradients de bon modèle. Attendez, d'accord. Ça, c'est le gradient descent de base. Oui, voilà ce que j'ai que j'ai temps. Donc Adam lui fait un truc plus compliqué où grosso modo, il va regarder, il va, il va manipuler la norme, il va regarder pour chaque, pour chaque paramètre, il va regarder à quel point ce que le paramètre il a déjà bougé à l'étape d'avant. Il va calculer en quelque sorte des des des moments. Sur la modification de gradient et dans ce qui fait qu'à la fin c'est pas exactement le Gradient, c'est. C'est tout un ensemble donc le gradient Clipping ça veut, ça veut dire que dans certains cas votre gradient il va être très fort dans certaines directions. Et combiné. J'ai gradué en très fort avec des Optimiseurs comme Adam par exemple qui vont être parfois assez violents dans leurs updates. Fait que vous allez vous allez être installé et donc vous pouvez choisir de faire du gradient Clip PIN ou vous allez vous allez regarder pour toutes vos paramètres, c'est quoi le Gradient ? Et vous allez lui dire OK Là il est trop il est trop fort. Donc je mets typiquement une valeur type je sais pas ça va être 5 quoi. Vous voulez pas que votre gradient il dépasse 5 en orme du coup OK ? Ok donc en fait c'est juste dire en gros quand je converge, mettre une limite à la convergence genre comme si on bridait Adam pour dire en gros Bah t'as pas de dépasser OK OK OK moi c'est clair. Voilà. Donc c'est vraiment l'idée que vous voyez là sur l'image à l'écran on a la direction mais on clippe le gradient pour dire Bah en fait non on fera pas plus de 50 cm parce que. Là si je tombe, je meurs quoi donc euh, je vais y aller petit à petit, pas par petit pas, et j'ai pas envie de faire des étapes trop longues, même si on on serait allé beaucoup plus vite au sommet, c'est sûr. OK donc gradient clipping. Donc dans le cas du de entraînement de réseau de neurones, ici on parle de de trust building pour le cas du du RL, mais l'idée l'idée est vraiment la même, juste là on utilise le. Laquelle divergence pour. Pour contrôler le le, la quantité de changement qu'on va faire dans notre dans notre réseau hop. Ok, et donc pipio c'est un peu le le Boss final de de tout ça OK ? On est toujours sur un algo. Acteur critique OK acteur critique c'est un peu le mieux qu'on puisse faire en modèle Free parce qu'on peut pas modéliser l'environnement donc on on estime la la qualité de l'action et de l'État dans lequel on se trouve donc très bien on utilise du Gradient Clipping donc donc comme TPO quoi la fameuse stress Regen qui est. Qui qui évite d'avoir des des changements catastrophiques qui vont se faire complètement changer et donc perdre peut-être son appris juste avant. On utilise aussi les idées de de atroce advantage acteur critique où là c'est vraiment des idées de de des. Comment dire des notions de performance pure. C'est à dire qu'on va avoir plusieurs, plusieurs agents en parallèle qui vont. Qui vont chacun utiliser la policier, vont chacun faire des erreurs de l'autre côté, comme ça vous avez un batch entier. Vous allez pouvoir faire votre bec un peu de la même manière que on pourrait faire du. On préfère du de du Replay offer ou on va apprendre plein de choses. Sauf que là vous êtes en parallèle quoi. La plein d'agents vous faites-vous faites vos actions, vous récupérez vos reward et vous battez ça pour faire votre update de vos réseaux de neurones, donc de votre acteur et de vos critiques. Ok, on utilise un bonus d'entropie pour l'exploration. Ok donc là on est en gros c'est une méthode un peu plus intelligente que épisode crédit. Où on va maximiser on va, on va s'assurer qu'on a une entropie minimale pour pour la prédiction de notre politique. Ok, donc rien de compliqué, juste c'est c'est un peu plus intelligent et plus grandit, c'est tout. Donc donc voilà. Tertio. Toc, toc, Toc donc en en gras, c'est un peu le CE qui est à retenir quoi. On reste à un modèle on policy donc on est moins simple les fish ou que des des algos of policiers comme sac par exemple ou ou des tunnels ou des 3. Parce que parce qu'on est des policiers, donc on retient pas ce qu'on ce qu'on a fait précédemment avec une autre policier. Et donc ça veut dire que, à chaque fois que vous voudriez updater votre policier, il faut faire de nouveaux. Il faut à nouveau comment dire bouger dans l'environnement quoi. Donc là ça dépend. Enfin ensuite, si ça peut être intéressant dans certains cas et moins intéressant dans d'autres en fonction de l'environnement, sûr que vous avez sous les bras. Ok, donc ce qu'on dit c'est que visuellement on est on est moins efficace par sample. Je sais pas comment je traduirais ça simple. Eh Ben c'est très bien aussi. Que des algues policiers donc comme on a déjà vu des un sac étroit avec sacs sac a utilisé en priorité comme on a vu que c'était un plutôt cette tête audiart pour des environnements avec des actions continues, OK ? Et cetera. Je sais, il y a beaucoup de choses et c'est pour ça que je vous ai vu en moins, on a une slide par une slide par algorithme, vous avez vraiment les les les points clés quoi. Parce que ça sert à rien qu'on replonge à chaque fois dans chaque dans chaque algorithme. Et je vous ai mis à la fin les liens vers des tutos et cetera pour bien comprendre en profondeur chaque groupe. Mais de toute façon on aurait pas eu le cas. Je pense à vous aurait pas intéressé de plonger plonger au fond de tous les de tous ces algos là. Parce que si je peux décrire facilement les idées générales, je pense de l'idée d'une entropie, et cetera, c'est assez simple à à à comprendre en. En substance, après, en pratique, les maths, ça peut être plus compliqué. Mais c'est vraiment plus important on off policy, on est acteur critique, on utilise une trace de religion ou du gradient flipping OK, on parallélise sur plein d'environnement pour avoir plein d updates. Et là vous avez une idée générale de comment fonctionne pipe OK pipe qui est vraiment enfin un des algos qui marche le mieux dans du dans le cas du modèle Free. Donc il faut un des premiers qu'il va falloir enfin qu'il faudra tester pour vous, en tout cas pour pour les problèmes que vous aurez sous la main. Ok, il est 59 ce que j'ai. Ok donc donc c'est fini, c'est fini pour les Algos of policy. Félicitations à tous. On on passe maintenant pardon sur les algos modèle Free. excusez-moi. On va vraiment maintenant avoir une. Une. Comment dire, je vais vous présenter en gros le le le la trajectoire qu'on va avoir pour obtenir le dernier algo de RL qui s'appelle efficient Hero. Bon, qui a permis de faire des. Enfin qui qui a des performances assez incroyables sur les dernières, sur les sur certains jeux vidéos, et cetera. Et donc tout ça c'est un peu l'histoire du. C'est un peu l'histoire du. Durée de force learning model base OK avec cette idée fondamentale que dans votre tête faut vous dire que dans tous les cas modèle base ça marchera mieux parce que on on est sur des sur des algos qui essayent de prédire l'avenir par eux-mêmes et donc intuitivement je trouve que on peut vite comprendre que enfin se rendre compte que. On va être plus efficace si on arrive à bien prédire l'environnement, si on arrive à prédire que on va se prendre le mur, si on freine pas en vélo. Bah c'est mieux que d'apprendre que en fait si j'ai une détection j'en sais rien, moi mon capteur et cetera, alors il faut juste que je je le frein pour pas avoir une mauvaise réponse. C'est quand même plus intelligent entre guillemets et plus efficace. De comprendre que en fait, si je dois freiner c'est parce que si je freine pas je vais, je vais me prendre le mur quoi. Je sais pas si la la différence est est est très claire mais. Mais c'est, c'est vraiment cette idée de. Dans la vraie vie, avec des environnements complexes comme comme vous en aurez beaucoup, vous ? Vous avez grand intérêt de manière générale à modéliser votre virement. Bien entendu c'est pas toujours possible et c'est pour ça que c'est pour ça que toutes les algos modèles Free ont été développés. Mais quand quand vous avez la possibilité de notre désir environnement et les ressources parce que ça aussi ça coûte plus cher quoi. Et Ben c'est toujours c'est toujours très intéressant. Donc tout ça pour dire. Qu'on va on va voir de des algos principaux de modèles baisse qui ont été développés dans les années dans les années récentes là. Donc là je vous ai mis un résumé de ce que ce que je viens, ce que je viens de vous expliquer, hein ? On on est plus malins parce que en fait, on arrive à anticiper ce qui va se passer, que ce soit dans n'importe quel jeu vidéo ou n'importe quel cas de de avec un agent dans environnement, c'est c'est toujours intéressant à priori de de de pouvoir prédire ce qui va se passer. Euh mais c'est difficile. Un problème difficile ? Pas. T'as pas t'as pas l'oublier OK. Et donc pour présenter un peu ce ces algos là je vous présente le la la ligne directrice qui s'est passée entre Alpago Etienne 0 donc qui est le dernier, le dernier Algo en date donc qui est de l'art. Euh, parce que ils sont assez, ils sont assez liés, on va le voir. Ok, tout le monde connaît les Jeux Atari. Oui. Alors déjà Atari c'est c'est rigolo mais c'est dur quoi. On a, on a des pixels dans l'écran et on va, on doit-on doit apprendre à. On doit apprendre à manipuler nos nos commandes pour pour gagner. En gros c'est un dataset. Enfin c'est comment dire. C'est une tâche qui est ultra connue en IRL parce que souvent c'est des tâches qui sont faites super dures pour des machines à apprendre toutes seules. Quand vous pensez là vous pensez le le jeu tout en haut à gauche là ou faut faut envoyer la la BIT dans faut envoyer la bille. Pour casser les briques, là bah la méthode optimale pour le faire c'est arriver à casser un petit trou sur le côté. Et et à et à laisser la bille enfin à bloquer la bille dans dans le trou en haut pour qu'elle fasse plein d'alertes trop rapidement et qu'elle fasse plein de briques. Vous avez des des jeux hyper durs ou faut explorer l'environnement ? Que là je pense aux Jeux que. Bah 2e ligne avec les échelles. Et bah votre votre agent, il doit apprendre tout seul que il doit-il doit explorer, il doit aller chercher des il doit aller chercher des objets dans l'environnement, et cetera, c'est super dur. Et si vous êtes pas capable de de de bah de modifier votre numéro suffisamment bien en fait vous y arrivez pas ? Donc techniquement des modèles des des réseaux de RL comme Pipiou et cetera, Bah en fait ça marche pas. Ce que l'environnement est trop grand et vous pouvez pas juste apprendre à partir de à partir de votre État et de votre action quoi ? Je sais pas si vous vous rendez compte mais dans le cas d'une voiture par exemple, c'est pas c'est pas indécent parce que la voiture est toujours dans le même environnement. Typiquement là si on prend le si on prend le jeu qui est qui est ici là je pense que vous voyez ma souris là un pupille ça pourrait bien marcher parce qu'en fait le le l'État est relativement simple à encoder. On a la position de notre voiture sur la route et on doit juste faire droite ou gauche en voyant un peu en avance. Si j'ai des obsèques devant, vous voyez donc là un \*\*\*\* ou un sac, ça ça peut bien marcher, ça c'est c'est, c'est ça pourrait pas mal quoi. Par contre ici et ici, c'est un il est celui-là, il est connu je crois en plus. En R, pour être particulièrement dur parce que on demande à à l'agent d'être d'être curieux d'aller chercher. Donc voyez, il faut qu'il aille chercher les petites clés. Faut reconnaître. Et il faut qu'il fasse la différence entre une clé qui est intéressante et des monstres qu'il faut surtout pas toucher par. Donc là avec Pipo c'est c'est très dur, parce que comment est-ce que j'encode mon environnement là ? Et comment est-ce que j'arrive à voir où est-ce que ça m'intéresse quoi hein, qu'est-ce qui qui est vraiment intéressant ? Bref donc les jeux Atari c'est un problème super dur en avril OK ne vous inquiétez pas, je vous demanderai pas de coder un algo de R pour ça, pour votre projet final. Parce que c'est vraiment des problèmes auxquels s'attaquent uniquement les Deep mind. Non, pas du tout. Enfin quoi. Maintenant maintenant on arrive bien parce qu'il a des infos qui sont sortis étaient efficaces, mais. Mais voilà, c'est très difficile. Donc pourquoi est ce que je vais vous parler de Fish and 0 ? Et Bah pour ça, globalement ça c'est la figure qui était dans leur papier. Donc un papier de 2021 qui sortira de RIB pour que connaisse une une des plus grosses conférences de dia au monde. Et donc globalement bah leur modèle il a écrasé tous les autres quoi. Mastering Amsterdam, vous les mettez de l'État donc c'est Etienne 0 en 2021 et ils ont que juste ils avaient des performances incroyables sur donc les 100 jeux enfin les le benchmark Atari quoi donc ce qu'il faut, ce qu'il faut retenir c'est que c'est particulièrement une performance intéressante parce que. On est sur un benchmark où il y a des données limitées. Et c'est ça qui est intéressant, non ? J'en parlais encore ce 12h00 avec votre prof. En c'est c'est ce qui est beaucoup dit EN y a pour la raison pour laquelle on atteint pas encore l'IA général, c'est que c'est Diane Lecoq qui dit ça tout le temps aujourd'hui un un être humain il apprend à conduire en 10h00 alors qu'une IA donc vous voyez les les les ressources comme Tesla notamment pour apprendre à le véhicule à. À circuler mais on est sur des 1000000, des milliards d'heures de conduite enregistrées avec tout pour essayer d'avoir un algo qui marche pas. D'où d'où pardon le le l'intérêt de ce benchmark Atari sanka qui limite en gros la donnée. Enfin le temps d'entraînement que le modèle a à jouer au jeu avant de de d'apprendre. OK. Donc la performance avec une une donnée non limitée et moins impressionnante mais donc là dans le cas de données limitées, le efficient 0 est vraiment vr ? Comment ça, vraiment très impressionnant ? Ok, donc il est-il est meilleur que l'être humain en. Avec avec un une quantité de données limitées donc meilleur être humain au jeu Atari hein. Donc nous on est quand même censé être plutôt bon à ce genre de jeu. Mais là Edition 0 est est meilleur que nous en avec le nombre de données limité donc. La question ça va être bah comment ils ont fait quoi ? Et et là vraiment, ce que j'aimerais vous faire comprendre, c'est que avec le cours que vous avez suivi, vous avez toutes les clés pour comprendre vraiment comment ils ont fait quoi. Enfin concrètement, comment est-ce qu'ils ont implémenté leur truc ? Après bien entendu beaucoup d'ingénierie pour pour y arriver, mais conceptuellement, vous avez maintenant les clés pour. Pour comprendre ça. Ok donc pour arriver à l'effet 0, il faut repartir en quatre-vingt-seize avec la le succès de Deep Blue. Qui a qui a battu Kasparov aux échecs ? Ok là y avait pas d'i a à ce moment-là, c'était vraiment un truc un peu brute force parce que les échecs c'est suffisamment petits comme jeu pour qu'on puisse un peu calculer. Tous les États possibles. Ok. Et du coup après Deep Blue, on a dit, Bon bah très bien, les ordis sont forts parce qu'ils sont capables de tout calculer comme des gros bourrins. Mais on a encore des Jeux, il y a tellement d'États possibles et ce sera jamais possible pour y arriver. Donc notamment le le jeu de Go, je pense que tout le monde le connaît et qui a beaucoup plus d'États possibles que le jeu d'échec. Est-ce que quelqu'un a une idée du nombre d'États possibles pour le jeu de go ? Du nombre de positions possibles pour un jeu d'échec, je Go jeu. Des chèques en a pas tant que ça. On est de l'ordre de je pense de quelques 1000000 de positions possibles, donc quelques 1000000 c'est faisable par un ordinateur. Et là pour le judo ? Je sais pas, peut-être un truc genre. Les puissances 12, 10, puissance 15. Encore je au pif 10 puissance 25 qu'il faut mettre. Puissance 170 on est sur un un, c'est c'est complètement indécent. Euh. En gros, c'est impossible pour un ordi de de calculer tous les États possibles. Et c'est pour ça que tout le monde disait, tout le monde disait. Que que les y arriveraient jamais à ce jeu là. Parce que en fait c'est c'est juste impossible d'explorer tout là et que il y avait en fait l'humain qui a l'intuition et la la qui va avoir une stratégie vraiment pour, parce que il a une bonne capacité à évaluer la qualité du jeu, et cetera. Mais du coup ça ça veut dire que pour le jeu de Go. On pourra jamais avoir un ordinateur qui est sûr de battre un humain à tous les coups comme les échecs. Exactement. Ouais, sur et certain évidemment c'est pas possible. OK c'est c'est très, c'est très malin ce que tu dis, parce que en effet, en échec, comme t'es capable de d'explorer tous les États possibles, tu peux être certain que quoi que l'humain fasse, il soit, il va toujours être battu. Alors que là en effet, on n'est pas capable. Dans tous les cas, mathématiquement c'est impossible de. Et physiquement c'est impossible de tout calculer. Donc on peut-on peut au mieux avoir une stratégie qui bat tous les humains à tous les coups parce que les humains sont trop nuls. Mais on aura jamais de garanties que on pourrait y aller. Donc c'est c'est une une différence qui est qui est très, qui est très pertinente. Merci d'être là. Donc ouais, l'idée c'est un peu comme ce qu'on s'était dit avec l'histoire de Fortnite versus le labyrinthe, on n'est plus capable de faire du tabulaire avec ce genre de ce genre de modèle et donc on va être obligé de de développer des algos qui vont qui vont. Qui vont explorer efficacement le le bon. Euh OK, donc le jeu de go ? On n'est pas possible. On n'est pas capable de de d'appliquer du du R tabulaire Dessus. On va avoir besoin de du renforcement. On parce que justement, c'est ce qu'on a vu dans le cours, quand on n'est plus capable de mettre tout dans une grille. Eh Ben. On va, on va avoir des méthodes pour. Pour rendre ça continue d'une certaine manière et ne pas tout explorer. Donc typiquement un si j'utilise un algo style acteur critique et Ben il suffirait de regarder la grille et de me dire à quel point est-ce que cet état là il est intéressant pour moi et avec un réseau de neurones je suis capable de le faire parce que j'ai pas tout retenu. Je je prends juste en entrée l'État. Virtuel et je. Et je en sortie, je le dis à quel point ce que cet État est intéressant. Donc à quel point est-ce que l'action de poser la Pierre, ma Pierre à tel endroit là va être intéressante ou pas ? Ok, dans l'idée c'est ce que c'est un peu ce que va faire alpago, mais. Il va le faire un peu mieux parce que on est sur des algos, vous rappelez des algos modèle paste, donc on va chercher à modéliser qui va être possible OK ? Ce que je remette. Et donc pour ça ? On va utiliser un algo qui s'appelle Monte-Carlo Tree Search, normalement vous vous devriez en avoir pas entendu parler aussi. Donc là c'est une slide sur le Monte-Carlo en entier. Je c'est vraiment au cœur de ce qui se passe dans alpago donc donc je le mets ici. Et de manière générale c'est un algo qui est beaucoup utilisé en en modèle base. Parce que l'intuition de Monte-Carlo twitch c'est que on. On va chercher à. On va chercher à prévoir ce qui va se passer dans dans dans l'environnement et ça ça se présente un peu sous forme de graphe. On l'avait déjà, on a déjà exprimé les enfin écrit par des problèmes de R comme ça je vous rappelez avec les États, avec les flèches ou avec les transitions et et les actions. Je sais pas si vous rappelez les graphes avec les grandes, les grandes sphères blanches pour les États, les petites sphères noires pour les actions. Et donc entre chaque, on calculait des actions value. Enfin oui ça des des valeurs de de State action value. Pour pour arriver à trouver c'est quoi la meilleure action à faire quand je suis dans tel État, donc l'idée de monter car maîtresse c'est un peu comme c'est un peu du du même style où je vais. À partir de du point où je suis-je vais envisager plusieurs chemins possibles. Donc là ici bon là vous en avez 3 à chaque fois, enfin 3 au début. Donc je suis dans telle position. Bah le joueur en face il va pouvoir faire tel ou tel ou telle action. Et comme ça, je vais étendre mon arbre. Pour pour trouver, enfin pour pour envisager plusieurs scénarios quoi. Et donc là la clé c'est que on sait que on sera pas capable d'étendre notre arbre jusqu'au bout parce que il y a 10 puissance plus de 100. Possibilité donc, c'est impossible, mais si j'arrive à choisir intelligemment les chemins que je vais explorer ? Trouver les chemins les plus intéressants pour le pour, pour le suivant. Vous vous doutez bien. Dans le cas du Go, bah ça va pas être intéressant forcément de placer une Pierre au milieu de nulle part. Donc j'ai pas forcément explorer cette cette cette trajectoire là. Et c'est un peu je pense la manière dont on va pouvoir réfléchir pour ceux qui jouent aux échecs, c'est un peu comme ça qu'on fonctionne on va se dire. Bah en fait il va pas mettre en échec son roi tout seul, il va juste pas faire ça. Donc je vais pas considérer cette possibilité là par contre Ah bah il y a tel truc ça M Emmerderait bien donc je vais explorer cette direction-là OK, si il fait telle action, moi je ferais quoi derrière ? OK, est-ce que ça me met dans la \*\*\*\*\* ou pas trop, ou et cetera. Donc c'est l'idée de monter Carlo Tree Search ou on est sur du Monte Carlo, donc on fait quand même du sampling, donc on va sélectionner de manière un peu avec un peu de de stochasticité les les branches qu'on va explorer. Typiquement avec épisode crédit OK ensuite donc ça c'est la sélection. C'est quelles ? Quelles actions je vais considérer pour la l'étape d'après ? J'étends, OK, donc je je pars de cette action là, je fais OK, qu'est-ce qui va se passer ensuite ? Je modélise mon environnement, ce qu'on a toujours du modèle base. On donc on l'étend. Et on va voir, on va estimer. Donc on a la simulation, donc on va estimer ce qui va se passer si on fait, si on suit tel ou tel chemin, OK. Je. Et donc grâce à ça en fait, ça nous permet de d'accéder à des trajectoires possibles et de détecter peut-être à l'avance que. Eh Ben là je suis en vélo, et bah en fait dans 2 trajectoires. Enfin dans dans 3 Timestamp je me prends le mur et donc là j'ai une information qui a beaucoup de qui a beaucoup de de valeur parce que je sais que je vais pouvoir peut-être me mettre à freiner parce que. Enfin pour éviter justement de de me taper le mur d'après. OK donc ça s'appelle la propagation. C'est à pas confondre avec la back propagation d'un d'un réseau de neurone. Mais c'est l'idée de je prédis sans vraiment avoir fait le. Sans vraiment avoir comment dire sans vraiment avoir suivi ces trajectoires là je la prédis et J update la valeur de mon état maintenant en fonction de cette prédiction là OK. Donc ça nous permet en quelque sorte. D'explorer l'environnement de manière virtuelle avec l'estimation qu'on a de le, de, de, de comment on modélise l'environnement. Donc voilà ensuite Ben c'est quelque chose qui coûte cher parce que en fonction des hyper paramètres que vous prenez, bah certes on va pas explorer les 10 puissance 170 sans possibilités. Mais on va en estimer beaucoup et ça bah ça coûte, ça coûte de 10 ans de calcul OK ? Est-ce que c'est à peu près clair ? Pour ce que c'est, montecarlo Tricia Rech. Ouais, à peu près Monsieur. Ok. Merci. Of. Ah bah je sais pas. Attends. J'avais changé ce slide. Juste une petite question Monsieur, j'ai déjà vu en ligne vite fait les vidéos qui expliquaient comment on dit Blue fonctionnait et c'est à peu près de cette manière-là. Ouais, je crois exactement. Ouais, ouais, c'est pour les échecs du coup. Exactement. Parce que en fait, à à l'époque, à l'époque c'est pareil, on pouvait pas tout calculer parce que les ordis étaient moins puissants aujourd'hui. Mais dans l'idée, si aujourd'hui tu voudrais faire la méthode brute-force parce que nos ordis ont des GPU énormes, et cetera, on pourrait le faire là où Deep Blue pouvait pas vraiment le faire. Donc ils ont un peu fait, ils ont un peu fait pareil. Ils ont un peu fait pareil que d'alpago dans ce dans ce sens-là. Mais alpago, il y a d'autres choses en plus hein. C'est pas juste Monte-Carlo mais en effet il y a beaucoup de y a beaucoup de possibilités et en général, enfin, quand vous avez des idées échecs dans vos, dans vos téléphones, et cetera. Parce que bah c'est certes c'est possible, mais ça reste quand même assez cher de calculer toutes les possibilités. Et Ben on utilise quand même des des algos du type montée à autres Search dans le cas de dans le cas des échecs. Ouais du coup Alpha Go on utilise le mts donc la date à retenir c'est mars 2016 où on a où il a battu lycée. Je sais pas si vous vous rappelez de vous vous rappelez de de des News à ce moment-là, mais ça avait quand même fait vraiment couler pas mal d'encre. Et donc en plus de du Monte-Carlo Tree Search. On utilise ? Ils l'ont entraîné avec de l'imitation learning et du safe play, OK ? Imitation learning c'est on l'a pas trop vu, mais l'idée c'est que vous donnez à votre à votre agent. Euh, des trajectoires humaines, des trajectoires supposées optimales pour pour aider le modèle à apprendre, c'est à dire c'est un peu comme tout à l'heure dans le TP de ce matin où vous vous pouviez cliquer sur une action en particulier pour forcer le modèle à faire une action. Bah là ce que vous faisiez c'est que vous donniez des exemples au modèle pour qu'il vous imite. Je sais plus avec quel groupe je l'ai fait. Mais aussi y avait un mode supervise learning où vous pouviez même y avait pas de RL du tout juste vous lui donnez une trajectoire donc vous cliquez plusieurs fois sur des actions et le modèle a essayé de reproduire votre comportement. Et on voit que ça marchait assez bien d'aller tout droit et parfois il tournait, et cetera. Donc ça c'est imitation learning, ça veut dire que on va. Va partir d'une d'une politique qui est Random. On part d'une politique. On va essayer de fitter sur un joueur humain. Il serait un bon joueur humain. C'est un peu une un.de départ. Et plus malin que juste faire une policière Random par exemple. Donc voilà après je trouve que imitation learning c'est assez clair quand même comme comme nomenclature. Mais c'est un terme, c'est un terme à bien connaître. Enfin, c'est c'est un terme assez important en RL de manière générale et self play. C'est pareil, c'est assez transparent, c'est l'idée que une fois qu'on a un modèle qui est bon. Eh Ben on va le faire jouer contre lui-même. Pour qu'il apprenne de ses propres erreurs, c'est dire qu'il va voir ses propres faiblesses. Donc à la fois il va il va pas propager pour anticiper que ça c'est une faiblesse. Donc il faut pas arriver dans cet État là. Et en même temps il va apprendre que. L'adversaire est dans cet État là. Et bah là justement il a il a une carte grise pour pour aller pour être encore plus fort. OK le safe play c'est l'idée que un modèle peut apprendre contre lui-même. Donc là c'est beaucoup le cas dans des Jeux. Typiquement des jeux à 2 joueurs, des Jeux de plateau. Ok. C'est le le ciel plaît. Est ce que on peut faire des paroles avec les gars ou les réseaux adversaires ? Pourrait. On pourrait, sauf que dans un Gan t'as. La discriminateur est exactement mais du coup t'as bien un modèle qui joue contre un autre mais les 2 sont différents et du coup l'arbitrage y fait Comment ça va partir dans cette 2 règles déjà donné ? Bah là du coup on est sur du Go donc oui les règles sont connues, oui OK, mais du coup si on voulait implémenter du siège là pour d'autres cas de figure ? Bah l'ensemble de règles. Ouais exactement Bah t'as tout c'est c'est comme toujours en R, t'as ton, t'as ton environnement. Et donc là le Safe place ça veut dire que. On va dupliquer notre modèle, on va avoir 2 agents avec les mêmes poids qui vont jouer l'incompris, qui vont apprendre l'un et l'autre de leurs erreurs. Donc la différence, la différence avec un Gan, c'est que donc pour ceux qui connaissent, pas un Gan pour generated adversaire Networks. On va donc typiquement pour générer des images, on va entraîner un modèle à générer une image. Et on va entraîner un discriminateur qui va dire ça, c'est une bonne image, ça c'est pas une bonne image. Et donc j'ai un réseau qui est entraîné à produire des bonnes images et le et donc le réseau générateur, il est entraîné à faire en sorte que le discriminateur se rende pas compte que c'est une fausse image. Et donc en général, ça marche super bien, vous avez ce que vous avez sur le site Internet. Ça, c'est rigolo ça. Personne je sais pas si vous avez un on a fait chez les gars de on les voit en même temps en générative ifor comme Peter ça ça m'étonne pas. Pour ceux qui connaissent pas, hop je vous mets le lien je et donc si vous cliquez sur le site, là à chaque fois que vous rentrez la page vous avez une photo de quelqu'un. Qui a été généré par un gag ? Donc c'est pas un modèle de diffusion, c'est vraiment un gag là. Donc on est pas sûr. On n'est pas sur le même genre de modèle génératif mais donc c'est pour vous dire ça marche super bien. Donc on a un un modèle qui génère une image, une autre qui va discriminer. Ok, donc c'est quand même différent comme Jeff play parce que. Parce que le sage place est le même modèle à droite à gauche qui vont qui vont jouer, alors que pour un an on a un modèle qui est entraîné à générer et l'autre est entraîné à discriminer, c'est pas du tout pareil. On entraîne 2 modèles et dans le cas du du Gan d'ailleurs, le discriminateur à la fin il sert à rien, on l'entraîne mais dans le vide, on ne gardera pas ses poids, on va garder uniquement le générateur. Donc voilà. Mais donc l'idée du chef play pareil je trouve ça assez assez transparent. Du coup j'ai une grosse tête qui apparaît sur mon écran. Hop je. C'est assez transparent. On a on fait jouer le modèle contre lui-même pour qu'il s'améliore. Ok donc ça c'est pour Alpha Go avec 3 erreurs 3. 3 concepts fondamentaux qui sont Monte Carlo, imitation learning, safe play. Ok je l'ai remis après. Ah non, camino, j'ai changé. Bref, pardon, j'avais changé l'ordre. Ok, donc le problème d'alpago, c'est que. On est sur du de limitation learning, ça veut dire qu'on a besoin de trajectoires humaines pour apprendre au début, d'accord ? Et ça en pratique. Bah on a pas trop envie. Parce que si vous Découvrez un nouveau jeu, un nouveau jeu, vous avez pas envie de demander de payer des gens pour vous faire de la donnée pour que votre modèle il il apprenne dans un premier temps à faire comme eux pour ensuite être meilleur quoi. L'idéal ce serait d'avoir un agent qui apprend vraiment de 0 justement de de pouvoir vraiment partir d'un policier Random et d'apprendre. D'apprendre par lui-même quoi ? Euh donc. Oula octobre 2027 dit Donc c'est c'est très fort ça, je suis devant. Donc donc un an après, vous aurez deviné que c'était 1017 hein, un an après alpago. Deep mind, sorte Alpha 0 ou en fait. Euh. Ils ont enlevé la dépendance en limitation le. Donc y a plus besoin d'avoir des des exemples humains pour apprendre. Ok, et sinon ça, ça fonctionne un peu par. Ouais, c'est pour ça que on vous allez voir à chaque étape, on gagne, on gagne encore en capacité et en performance. Euh. Mais donc donc toujours dans le cas du Go, OK donc Alpha 0 était bien meilleur que Alpha Go donc il va. Enfin il était encore meilleur quoi donc c'est. Autant vous dire que aucun de l'être humain peut battre Alpha 0 au Togo en tout cas. Mais il y a encore un truc qui pose problème, c'est que dans ces 2 modèles, là encore une fois, on est basé sur mon cahier d'Autriche Arch. Implique que pour faire notre expansion en fait on a besoin de connaître les règles du jeu pour faire mon expansion, je sais que c'est pas de placer ma Pierre n'importe quel endroit. Donc les règles d'Hugo étaient toujours hard, codées dans le modèle à ce moment-là. Ok, donc. Et donc ça. C'est aussi un problème parce que dans beaucoup de cas donc typiquement dans dans un jeu Atari ou si vous voulez jouer à fortnite avec un modèle de RL, Ben en fait vous avez pas accès au code source, vous pouvez pas dire Vous savez pas exactement ce qui va se passer. Vous connaissez pas exactement la la physique du jeu quoi. Euh ouais. Et donc c'est une limitation un peu de ces modèles là. Ah c'est tout. Faut que j'aille voir après. C'est une imitation de ces modèles là parce que on est obligé de coder à l'intérieur du modèle. Quels sont les quels sont les règles du jeu ? Ok donc encore meilleur, on passe à 1000 0 cette fois. Et bah vous vous en doutez, on. On s'extrait du des règles hardcore. Ok, encore et toujours du mind hein. Ils étaient sur une bonne lancée cela. Je sais plus si ils étaient déjà rachetés par Google ou pas. Mais bon c'est pas grave. Euh. Donc là vous avez un vous avez un petit résumé, vous avez un petit résumé pour pour News 0 mais au début vous voyez Alpha Go c'était uniquement le Go. Et on avait Human data, domaine de la page et non non rouge Alpha Go 0. On avait encore les hard codés de on. On avait on a enlevé le domaine New AGE et domaine data mais on avait encore les règles du jeu étaient encodées. Alpha 0 cette fois. Euh il connaissait. Euh pardon oui à Chicago, oui il y a eu y a une, j'ai sauté une étape y Alpha Go 0 et Alpha 0 où en gros Alpha 0 c'était l'adaptation. D'alpha Go 0. À d'autres jeux que Lego, et donc dans l'idée, on a toujours, on a toujours les règles de du des jeux en question qui sont implémentées. Et donc ensuite on passe à mew 0. Où cette fois lui n'enlève les les règles du jeu, c'est à dire que il fait que observer le jeu de manière un peu brutale, que on observe les pixels, on observe l'État, mais on ne sait pas quels sont les. Quelles sont les les prochaines étapes possibles ? On a pas d'idée de quels sont les mouvements possibles ? Au début au moins en tout cas. Euh et donc dans ce dans dans mieux 0 il va falloir un. Il va falloir un moyen de modéliser ce qui va pouvoir se passer, OK ? Euh et donc dans le cas de numéro, on va modéliser plusieurs choses ? D'alors la value. La police, ça vous connaissez ? Ça s'appelle un modèle acteur critique. Donc le le critique c'est celui qui va qui va estimer la valeur de mon état actuel. L'acteur, c'est la policier. Ok ? Quelle action est-ce que je fais à partir de mon État ? Et cette fois-ci ? Et c'est ça qui se fait qu'on devient un modèle modèle base enfin, malgré ce modèle base, on va modéliser la reward ? Ça veut dire quelle est la la ? Enfin, on mobilise la reward ? Et donc qu'est-ce qui va se passer au moment où je vais faire que l'action ? Donc c'est un peu comme si on on apprenait nos propres règles. Ok. C'est à dire je sais plus. Enfin dans le cas de du goût, bah on me dit à la fin t'as une surface de temps donc tu as gagné donc tu connais ta surface donc tu connais ton tard au fur à mesure du temps-là c'est lui qui va apprendre tout seul comme un grand. Bah en fait d'après mon État et Ben Marie Wars ça devrait être ça. Et donc je vais pouvoir prédire là Marie, ça va être ça du coup après je vais faire telle action, et cetera. Donc on arrive sur 0. Hop. Là, je mets une petite une petite animation sur qui est disponible sur. A été faite par Deep mind au moment où ils ont sorti. Où ils ont sorti mieux 0 justement. Donc vous voyez que on a notre État, on va estimer notre policier, enfin notre on a nos policiers et on va estimer notre nouvelle action. Avec avec ce. Avec ces cette idée de on on va créer un on va créer un un arbre de possibilités pour mieux estimer notre value function, donc notre politique, et cetera, donc là on estime l'état d'après et cetera. Ok. Est-ce que c'est clair sur comment ? Sur l'idée générale de 0 c'est normal de pas comprendre tout en détail mais mais ce qu'il faut c'est vraiment l'idée générale. Ça pour un oui. Hop. Enfin, on arrive sur Edition 0, donc la dernière, la dernière itération est celle. Celle qui avait des des performances tiers impressionnantes sur les Jeux Atari. Hop tout, tout truc. Et donc epsilon 0. Bah on part de 0 et on ajoute encore de nouveaux des nouvelles implémentations des nouvelles optimisations. La force de certaines 0 d'ailleurs, c'est que c'est plus par Deep mind, c'est et ça a été fait par une une équipe d'universités américaines dont j'ai oublié le nom. Euh. Donc c'est vraiment ils ont pris mieux. 0 ils ont fait. Bah je pense qu'on peut en faire encore mieux et donc on a plusieurs tricks en plus. On a une ça, je peux pas constance kilos qui nous donne en quelque sorte. Une leçon en plus de juste pour dire comment dire. On va pas uniquement chercher à optimiser les la prédiction. Donc notre là ce sera pas juste la différence entre ma prédiction et ce que j'obtiens réellement. On va avoir en plus une sorte de régularisation, donc de consistance pour faire en sorte que en gros que ma prédiction reste un peu un minimum. Donc on va pas uniquement avoir comme objectif de prédire la vérité, mais en même temps on va voir comme objectif d'une certaine mesure de prédire quelque chose de consistant, donc de plus de plus d'une certaine manière. Je. On fait en quelque sorte du strapping OK. Donc c'est comme ça qu'ils l'ont rédigé dans leur bloc post. Mais en gros c'est du Bootstrapping, ça veut dire que n'attend pas la fin de l'épisode pour tout. Bah propager mais on essaie de faire des des updates vraiment à chaque à chaque comme steppe OK. Tac TAC, donc ça ça revient. Fameux value petic value P 6 et une offre politique collection. Là pareil je vous ai mis le lien du je vous ai mis le lien du blog post qui franchement explique bien. Un peu ce qui se passe. C'est pourquoi est-ce que ça fonctionne ? Bien sûr, je ouais, je pense que. Ouais, ça fait déjà 01h30, je pense que c'est bien, on va pas aller beaucoup plus loin. Et donc là je. Un petit graphe sur toutes les idées qui ont été utilisées pour développer l'action 0 à la fin. Avec ici la trajectoire qu'on a qu'on a décrit tout de suite, donc avec Alpha Go, Alpha 0 et 2 0. Et en plus bah les idées qu'on vient de voir donc le Tracking et cetera. Donc quelques autres papiers que enfin il faudrait citer. En tout cas, si vous étiez sur un papier de recherche pour pour voir un peu, bah les idées ont été récupérées par par l'équipe des fichiers 0 en tout cas. Ouais, voilà toujours la même idée que. On a compris maintenant le principe de base d'URL est donc dans ce cas-là de du modèle de modèle DSL. Mais ensuite c'est toujours une histoire de je vais rajouter des tricks, je vais optimiser la telle ou telle ou telle manière de de de modéliser. De modéliser le l'environnement, la reward et cetera. Je ouais ce que vous avez des questions sur section 0, sur ou sur la ce qu'on vient de faire au niveau du modèle base des règles de manière plus générale. Bon bah très bien toc OK donc on va faire une pause, on va faire une pause de 15 Min. Et là je je vous ai Claude les je vous applaudis les slides et et le TP on voit le TP juste ensuite juste après donc rendez-vous qui est super dans 15 Min à tout de suite. Monsieur. Ouais. Alors non je vous ai mis le j'ai mis bien du TP sur sur moodle. Vous avez. Vous avez accès au notebook ? Je vous présente rapidement et puis ensuite vous êtes parti, hein ? Vous avez l'habitude, je. Vu que tout le monde est là, j'en profite aussi, je vais vous présenter donc là il va être 15 h. Je vous laisser 01h30 de TP. Et je vous présente le projet à 16h30. Si ça vous va comme ça, vous savez, une demi-heure pour enfin pour pour discuter un peu. Je vous présente le truc et je vous donne les dates. Enfin on vous inquiétez pas, il y a pas, vous aurez le temps vrai, mais comme ça je m'assume que je m'assure que vous êtes. Ça vous avez bien compris ce que je vous demande et. Et on pourra lancer le truc. Donc ouais là tout de suite je vous présente le TP vu qu'on est en distance c'est compliqué de vous aider un par un mais vous pouvez me vous pouvez m'envoyer des messages et on peut essayer de faire des réunions en 1V un ses besoins et. Et puis et puis voilà je vous enverrai la solution à la limite à la fin et donc je vais mettre en plus gros je peux tu penses à plus petit non ? Ben non. C'est bon ? Donc. Pour mettre en application ce qu'on vient de voir, on va utiliser les dpg. Que je vous ai que je vous ai présenté aujourd'hui tout à l'heure on se rappelle des dpg, on a dit quoi ? Hop il est ouf il est non on l'a pas présenté, pardon on a, on a juste parlé, on a juste parlé rapidement donc c'est oui justement c'est bien comme ça, on a présenté les autres et là je vous en présente un nouveau. Qui comme sac est un un algo qui est fait pour pour des tâches qui sont continuées OK. Et on utilise ce qu'on appelle le lavage Howard et enfin ça va être impliqué là-dedans, mais dans l'idée on va se mettre dans le même environnement que la dernière fois. Je sais pas si je vous rappelle la dernière fois on avait vu qu'on avait été obligé de le de le discrétiser avec du timing pour pouvoir résoudre le problème et donc là. On va le résoudre en compte, c'est-à-dire que on va pouvoir avoir une output qui est continue entre donc entre le fameux moins 2+2. Pour, pour apprendre, pour apprendre à à garder le le bon. Enfin à prendre la bonne police. Ok, donc. Et ouais, on réutilise un comme un replay de fort. Ok, vous avez l'implémenter. Donc encore une fois, c'est comme la dernière fois. On stocke un ensemble d'États actions reward extrait OK. Et grâce à ça on va pouvoir réentraîner notre modèle, notre modèle acteur et modèle critique. Ok. Euh quoi de plus ? Je pense que c'est assez bien expliqué ensuite donc le mieux c'est que vous lanciez dessus donc on va vous allez implémenter le replay buffer et ensuite vous allez implémenter la fonction de update comme d'habitude avec votre y qui va être enfin. Pour pour fêter votre votre, votre critique, on va réutiliser pareil, exactement comme le DQN, on va, on va réutiliser notre évalue qu'on connaît déjà et la valeur qu'on aurait dû prédire. Tac, TAC, TAC et la loi pour pour l'acteur qui qui maximise les les actions pour lesquelles. La, le critique donne de bons résultats, c'est-à-dire le critique va dire telle action, elle est bonne dans ce cas-là et donc dans ce cas j'optimise mon acteur pour qu'il le fasse enfin pour la pour qu'il a plus de grosses pénalité pour cette action. Là voilà normalement est-ce que ça s'éclaire jusqu'au pour l'idée générale des dpg ou s'il y a des questions plus spécifiques ? Et sinon juste je vous invite à. Ah vous lancez dessus et vous hésitez pas à me à me pinger si vous avez besoin d'aide ? Ça va à tout le monde. Ah bon, c'est ça, ça va ? Okay très bien et donc pour pour les autres. On se retrouve à 16h30 pour pour présenter le. Représenter les projets. Voilà. Si vous avez des questions sur le cours ou sur le TP, c'est maintenant. Bon Ben j'en ai vu que tout est clair comme 2 broches alors hop Ouais n'hésitez pas à m'envoyer un message ou aux limites à m'appeler directement. Je sais pas comment ça marche mais je pense vous pouvez m'appeler en direct. Donc à tout à l'heure pour les autres. P. Est-ce que tout le monde voit mon écran ? Et on voit Monsieur. Ok, parfait. Donc. Comme vous le savez, je vous l'ai déjà dit, la fin du cours sera évaluée sur un projet. Un projet qui sera fait en groupe de 2 ou 3, éventuellement en fonction de l'impartialité. Mais le mieux, c'est par 2 parce que. Après on se retrouve avec quelqu'un qui enfin des gens qui font rien donc donc faites le par 2. Normalement, vous devez voir la section sur Moodle. Et donc vous devriez avoir accès au Notebook OK Notebook que vous voyez ici. Euh. Et donc je vous propose de d'entrer. Un agent sur un environnement plus compliqué que ce qu'on a fait jusqu'ici en environnement qui s'appelle le Highway House, OK, c'est un un environnement qui sert à entraîner des véhicules autonomes en 2D. Donc là je vous ai mis le le Repo Github. Donc c'est un c'est un environnement qui a été qui a été développé par-dessus hockey IG donc c'est un c'est le même interface que ce que vous avez utilisé jusqu'ici. Seulement c'est un environnement qui a qui a été implémenté en plus. Donc vous l'avez pas par défaut sur sur open mail. Donc sur. Sur ce répertoire, il propose plusieurs catégories possibles. Ok, il y a. L'autoroute, on va dire de base ou l'agent va apprendre. À changer de voie, accélérer et freiner en fonction des véhicules tout bas. Et il obtient une récompense en fonction de la distance parcourue au cours de l'épisode. Ok donc ça c'est la l'environnement de base que je vous propose de d'attaquer, d'accord, mais je vous donne la possibilité de s'entraîner aussi en bonus. D'essayer de de trouver des environnements plus compliqués. Donc typiquement, Highway of propose d'entraîner des véhicules avec des espaces d'action beaucoup plus complexes où on va avoir le contrôle sur le sur les roues du véhicule, là où dans le la. La version de base, c'est simplement, je choisis de changer de voie. Il n'y a pas de on maîtrise pas. Les rouges sur ce plan-là on va juste choisir. Hop je change une voix, je change une vie où j'accélère où je vous allez voir que c'est déjà pas si évident que ça d'entraîner un modèle qui fonctionne bien sur cet environnement là. Donc les règles du jeu sont plutôt simples. Je vous donne un autre look de base avec quelques implémentations que vous avez pas besoin de toucher, ni donc notamment le fait de pouvoir. Enregistrer une vidéo. Une fonction d'évaluation qui va prendre votre modèle et l'évaluer sur le Highway of et me sortir moi une métrique pour que je puisse voir à quel point votre modèle est bon ou pas. Et donc vous voyez que je fais une moyenne sur 30 épisodes de base que je verrai. Je verrai comment je fais ensuite pour pour l'évaluation finale. Mais donc dans l'idée je vais je vais faire 30 épisodes avec votre modèle et je récupère donc votre partie de votre note. Ce sera. La performance que j'obtiens en moyenne sur les 30 épisodes que j'aurais effectué avec votre agent, d'accord, rien de compliqué. Normalement, si vous si vous exécutez le notebook, vous avez déjà de quoi tester en un un mlp enfin. Un agent avec un LP de base en pour voir à quoi ça ressemble quoi, euh. Et ensuite, vous aurez tout ce qu'il faut pour pour. Pour continuer le projet, j'ai mis moins de questions que ce que vous avez des. D'habitude dans les TP parce que le but c'est que ce soit un peu ouvert, vous êtes pas obligé de le commencer maintenant sachant que on aura encore un dernier TP. Encore un dernier TP la semaine prochaine ou notamment je vous expliquerai comment est-ce qu'on fait pour s'intercaler avec avec Open I et STB Baseline, donc c'est le le fameux TPRL factice. Ou notamment, on va essayer de voir comment est-ce qu'on peut utiliser Tensor Board pour pour monitorer les expériences. Mais je préfère vous le vous lancer comme ça. Enfin, vous lancez sur le projet quand même, notamment pour vous puissiez constituer les groupes. Ok, donc notre prochaine séance elle est en assez longtemps parce que les vacances de février je crois. Pour moi la prochaine séance c'est le 27 février. Donc dans 3 semaines. Je j'aimerai. Vous commenciez à me donner les groupes déjà cette semaine ? Pendant c'est pas très loin, vraiment. Donc je vous donne jusqu'à jusqu'au mercredi prochain pour me donner les groupes comme ça moi je les note. Vous pouvez commencer à regarder tranquillement. Sachant que sachant que du coup dans 3 semaines on va faire 1TP qui va qui va vous aider pas mal à résoudre ce problème là là donc on va avoir 1TP où on va apprendre à utiliser les senseurs board et cetera, donc pas obligé de de de le commencer dès maintenant. Mais je veux les les groupes rapidement, comme ça on est fixés et y a pas de vous Retrouvez pas bloqué ce que j'aimerais qu'on discute maintenant là c'est. Bah déjà si le format vous paraît bien donc un format vous me rendez un code +1 rapport oui je parle du rapport mais le but du rapport donc c'est c'est marqué ici mais c'est de montrer que vous avez pas juste utilisé charge gpt et que vous avez compris ce que vous avez fait pour pour votre étude. Donc je vous demande de tester au moins 2 algos différents que vous allez pouvoir tester. Donc il y a peut être POYA peut-être ça que y a plein de choses que vous pouvez tester. Vous avez vu pleins d'algorithmes, le but c'est que vous en testiez plusieurs et que vous puissiez un peu sentir la différence entre les 2. On a on a beaucoup dit que pippo était +1 peu l'espion par exemple que sac. Mais que ça qui était peut être meilleur dans les environnements continus ou ce genre de choses, bah le but c'est que vous l'expérimentiez un pouvoir vous même. Vous avez observé des logs ? Vous allez voir que votre entraînement va être un peu long, vous allez voir que vous apprenez des choses que vous voulez ou pas. Et donc le but c'est que vous notiez ça dans un rapport pour m'expliquer. Bah on a testé tel algorithme, on a tel résultat, on pense que c'est dû à tel ou telle propriété de mon algorithme, et cetera. Voilà bon je pense que c'est assez clair dans le Notebook donc ça si vous avez des questions n'hésitez pas. Mais la consigne c'est ça, c'est que vous me rendez un notebook, potentiellement même un un repo github si vous voulez comme ça, vous pouvez inclure les poids que vous avez entraînés. Typiquement si votre. Si votre comment dire votre entraînement est très très long parce que vous l'avez fait en local et que vous êtes chauffé à à faire tourner pendant pendant toute une nuit par exemple pour refaire un grand traitement et que vous voulez me donner les poids, bah ce sera peut-être mieux de faire un github. Donc là vous me donnez le github, j'ai juste à faire git clone et je run je run votre notebook OK ? Euh qu'est-ce que je voulais dire ? Ouais je vais pas perdre de temps à à fixer, enfin essayer de fixe des problèmes d'environnement ou quoi ? Donc si votre truc il marche pas, bah il marche pas OK donc je vais voir. Je pense que je vais mettre en place un un environnement exactement celui qui est ici là, celui qui marche sur collab. Ok et en fonction si vous me donnez un collab. Du coup j'ai juste à cliquer sur sur run. Soit vous me donnez un un repo et dans ce cas vous me donnez les poids et moi je pourrais copier les poids dans mon environnement à moi. Mais donc dans l'idée vous gardez si vous faites une install en local vous gardez ces spécifications là. Et comme ça il y a pas de il y a pas de surprise de mon côté quand j'essaye de. Quand j'essaie de de faire tourner votre modèle quoi. Donc vous ne prenez pas une version de Py TORCH archaïque ou un truc comme ça ? Vous me faites une installation propre avec un environnement qui est compatible avec ce que vous avez sur collab ou vous restez sur collab à vous voir. Je vais rajouter là TAC TAC TAC je. Je vais vous remettre là. Ok mais encore une fois dites-moi, dites vous que je vais avoir 20 projets à tester donc je vais pas perdre de temps à à faire marcher votre truc si ça marche pas je voilà. Non. Est-ce que vous avez des questions ? Est-ce que vous trouvez le le projet est super bien ou super nul ? Est-ce que vous avez des suggestions ? Voilà, non, franchement c'est c'est bien Monsieur, c'est clair. Très bien. Ça va le voir. Bah très bien de toute façon encore une fois la prochaine séance donc dans 3 semaines on va avoir un cours très rapide. Un cours très rapide sur les applications modernes, un TP et le reste ce sera ce sera alloué au projet donc vous pourrez me poser vos questions en live et donc je reste bien entendu disponible par mail ou sur teams si vous avez des questions plus spécifiques. Ouais. Du coup Monsieur juste le TP de la séance prochaine c'est un TP noté ou pas ? Non non là c'est la note, la note de de code que vous avez, ce sera la note sur ce projet là. Très bien et le mais le TP de la semaine de de de la prochaine séance permettra de enfin on apprend unitairement à Ouais on verra. Typiquement ce qu'on peut faire pour entraîner 1PO sur sur un autre environnement, mais ce sera applicable à à ce problème là. Ok super c'est très clair nickel si vous avez d'autres questions sur pour oui j'avais juste une dernière question, la date du DE c'est quand déjà ? J'ai mis sur moodle c'est le 5 mars je me trompe pas si certains peuvent du coup le c'est quoi c'est une séance ou c'est possible ? On est en amphi et puis vous filez les sujets depuis le de c'est c'est l'examen c'est ça ? Ouais c'est ça examen écrit de devoir écrit. Ouais alors de ce que j'ai compris, moi je dois envoyer un sujet d'ici. Tout le temps il sera imprimé, vous serez convié dans une salle, il y aura un surveillant avec classique super c'était pour savoir je vous ai mis attends je te parfois certains d'eux ça ils sont sur un PC parce que il faut coder, donc c'était juste pour savoir si t'es sûr ? Non non non. Du coup le. Ce cours là est vraiment séparé en une note de code où vraiment vous êtes sur PC. Un truc et un examen vraiment écrit où justement bah le problème c'est que quand vous quand vous avez votre PC vous pouvez tout faire avec un PC donc bien sûr. Donc là vraiment écrit sans document ni PC faut voir si ça vous arrange. Vous voulez imprimer des trucs ? Pourquoi pas ? Mais par principe j'aime pas trop parce qu'en général du coup les étudiants ils impriment toutes les slides et c'est pas où je pourrais quand même. Et en plus j'ai pas envie de vous vous demander de de. Enfin bref je un bon donc a priori ce sera ce sera sans documents ni PC, juste sur papier. Mais donc ça veut aussi dire que je vous demande pas des choses hyper précises là-dessus donc mais je suis ouvert à la suggestion si jamais vous avez d'autres. Et dernière question aussi, c'est pour le rendu sous forme de rapport plus code, est-ce que si on vous fait un github +1 markdown un rythme markdown pour le rapport ça vous va ? Vous voulez vraiment un PDF propre ? Je veux que je puisse non pardon c'est bien. En plus dans le jargon vous pouvez mettre des des gifs donc. Des mêmes, OK ça marche. Bon très clairement des points bonus. Ah c'est bon ça marche donc c'est super. Non mais juste faites un truc propre quoi, c'est pas. Bon, on m'a regarde si vous voulez, mais ça reste un rapport donc je dois pouvoir imprimer et avoir un document, un un document clair quoi, ça marche ? Super. Ce que j'aimerais définir avec vous, c'est la Deadline pour le rapport et le. Pour la mettre à l'envers. Donc est-ce que vous préférez que ce soit avant l'exécution ? Le plus loin possible ? Parce que là on est, on vient de finir une Project, on a encore le rapport à faire, on a d'autres trucs, d'autres projets et d'autres machins. Là on est vraiment sous le travail. Du coup si on peut le mettre le plus loin du coup c'est possible pour eux, ce serait top. Franchement, c'est que tu dois avoir des. C'est genre 25 mars ou 27 mars, soit le pour dataset, une matière qu'on a. Si ça pouvait être dans ces eaux là ce serait pas mal. Mais avant peut-être. Ouais. De toute façon votre examen c'est le 5 mars. Donc j'aurai vos copies à corriger j'imagine dans la semaine du 10. Donc fin mars c'est complètement gérable. Si ça vous a super que vous préférez que ce soit plutôt comme ça, c'est réglé. Mais je vais faire un sondage dans le groupe class et je vous redis, d'accord. Ok donc donc pour l'instant donc sans sans modification de votre part, je fixe au 25 mars je enfin 25 c'était avant quand même. Attends ouais moi ça me change rien, je je fixe 25, je sais pas s'il y a des contraintes côté fraises en fait donc. Je note, je note 25 pour le moment. Je mette 25. Et je confirme pour. Je confirme avec les frais pour la date parce qu'il y a j'imagine qu'il y a une date à laquelle je dois rentrer les notes. Je confirme avec eux et vous continuez de. Okay, parfait et puis et puis voilà. Après je vous fais confiance hein, c'est pas moi, ça me change pas grand chose de le faire un peu plus tard. Ok super bah parfait nickel ce qu'il y a d'autres questions. Non, n'importe quoi d'autre. Ah oui, et pour la semaine prochaine ? Je pour la semaine prochaine. Alors attendez hop ce que je dis ça. Hop. Donc comme je vous ai dit, on va parler de pour l'instant, je vous ai mis du Dr Ying et. Et le RL chef de poney ? OK donc le but c'est encore une fois c'est de vous montrer la, c'est de vous montrer que enfin vous montrer les l'application vraiment récentes du RL. Vous pouvez si vous voulez, si vous avez. Un problème de RL qui qui qui vous intéresse en ce moment ? M'envoyer. En suggestion. Autre chose qui vous intéresserait ? Pas pour l'instant. Donc j'ai, j'ai juste ces 2 trucs là et j'en ai un autre. Je crois en rab. Pour en discuter quoi ? Le but c'est de montrer que ça ressemble. Mais si vous avez des une envie particulière de de sujets, de sujets, n'hésitez pas à m'envoyer. Ok, on va m'envoyer ça et pour Noé, possible d'avoir un sujet de 2 type non ? Non, parce que c'est beaucoup de travail pour moi. Euh, mais je vais essayer de vous envoyer un. Un peu une liste des choses à savoir. Ouais super. Enfin un peu. Oui un descriptif. Je ferai pas un sujet de de de d'entraînement parce que ça me demande trop de travail, mais je vous, je vous fais un petit descriptif de OK, ça va ressembler à ça globalement. Moi ce que j'ai en tête pour le moment c'est il va y avoir une partie question de cours basique avec vraiment ce que vous devez connaître, ce que je vous ai dit qu'il y avait pas mal de choses à connaître. Enfin voilà, connaître par cœur, ce sera pas des choses très compliquées, mais voilà, ensuite y aura 1QCM. QCM, pareil pour vous rappeler un peu de 2 3 de. 3 notions pour voir si vous vous rappelez de de quelques subtilités qu'on a vu pendant le cours et probablement ensuite du de la réflexion avec du pseudo-code. Donc je vous donnerai un problème et vous allez devoir réfléchir à une architecture, à une architecture en fonction du problème que je vous donne. Donc je sais pas, je vous dis je suis dans le cas du.\*\*\*\*\*\* labyrinthe. Bah dans le cas d'un labyrinthe, c'est quoi l'algorithme qui ira le plus vite et ce que je peux avoir par exemple une preuve de de convergence pour tel ou tel algorithme. Construire des choses avec le but. Le but c'est enfin on n'est pas en licence. Le but c'est que vous montriez enfin que je. Que je. Que je comme alors que je sois convaincu que vous avez compris le CE que vous avez fait pendant le cours. Donc c'est pas c'est pas une histoire d'avoir de de connaître par cœur tout mais. Il y a des choses à connaître parce que en fait si vous ne connaissez pas ce que c'est qu'une value function vous pouvez juste rien comprendre de ce qui va se passer quoi. Mais c'est pas le but de connaître par cœur c'est quoi l'implémentation de pépito ? Quoi. Voilà. Je doute, c'était encore pour cacher. Normalement vous devriez voir le la section projet final sur le moodle mais mais voilà OK parfait. Merci beaucoup Monsieur. Ouais pas d'autres questions. Y a pas que Léonard qui a qui a le droit de poser des questions. Même si j'aime beaucoup les questions de Léonard. Hop. Oui. J'ai entendu quelqu'un, mais ça s'est arrêté. Non, pas d'autres questions. Non ça a l'air de paraître clair hein. Les questions déjà été majoritairement posées par les néo. Ben écoutez Ben merci à tous, je reste là c'est encore 10 Min de façon parce que il pourrait pas finir mais ceux qui ceux qui sont au clair vous pouvez y aller je vous souhaite des bonnes vacances du coup je sais pas si vous avez des vacances mais en tout cas ils vous ont mis des vacances visiblement. Donc. Et puis pour la prochaine fois, vous hésitez pas ? Bah préparez vos questions parce que ce sera aussi la dernière fois. Vous pouvez me poser vos questions avant l'examen. Euh me demander si vous avez un cas d'usage qui vous intéresse en particulier ou des questions n'importe lesquelles ? Voilà voilà, dans ce cas je vous dis à dans 3 semaines et ceux qui veulent encore poser des questions, je reste là encore 10 Min. Merci. Oui Monsieur soirée, bonne soirée. Je vais vous demander un truc, ouais, juste pour outiller pour les implémentations qui pourraient être sympa. Moi j'aimerais bien savoir. Je fais à côté de de pour tout ce qui est juste loisir. J'aime bien faire de la génération d'images ou des choses comme ça. Vous savez avec les modalités diffusion si vous connaissez un peu. Bien sûr. Ouais bah comme fille Why tout ça automatique Eleven et des trucs comme ça. Et je me demandais, est-ce que il y a des implémentations en fait de RL juste dans la création d'images ? Moi en fait c'est juste ça m'intéresserait de savoir. Ce qui est réel pour la création d'images ? Ouais, je sais pas. Est ce que il y a par exemple pour ce qu'il y a du renforcement learning, par exemple pour les modèles récents genre One 1000000 pour les vidéos en français pour passer de l'image à la vidéo ou des choses comme ça ? Est-ce qu'il y a des implémentations derrière elle en. Alors là comme ça ça me dit rien du tout, c'est quand même un problème différent. En le R, t'as vraiment la notion de j'ai un agent qui interagit avec l'environnement quand tu génères quelque chose et ensuite on va, on va te donner un feedback. C'est un peu ce qu'on sait dont on discutait tout à l'heure, ça se rapproche un peu. Typiquement un Gan, ça se rapproche un peu d'un l'acteur critique, mais pour la vidéo réelle, c'est un gamin quoi. C'est génératif pour passer par exemple quand on veut générer de la vidéo. Je sais que la base c'était on prend une vidéo, puis on va essayer de prédire comment ça pourrait évoluer et ainsi de suite avec des avec des trucs comme ça, des gens. Enfin ce qu'on disait quoi. Et je me demandais comment pourquoi aujourd'hui les modèles de vidéos ils sont bien meilleurs ? Et est-ce que si ça pourrait pas être dû ? R, enfin je sais pas si ça fait sens. Je pose cette question vraiment en étant le flou mais mais voilà ça m'intéressait juste si il y a un lien à faire entre les 2 qui existe. Bah en fait ça m'intéresserait de le connaître. Voilà surtout c'est plus ça m'a que Ben écoute. Je suis pas expert non plus de. Parce que les générations de vidéos je. Moi, j'essaie de regarder un truc. Mais. A priori comme ça, je dirais que ce qui a fait que ça s'est amélioré, c'est c'est le skate là. Ouais, il y a pas de y a pas de débat juste. Mais plus de compute et ça marche très bien en réalité, il y a des tricks dont on y a les les derniers trucs de. Flo, matching donc notamment pour les vidéos, il y a des notions de il y a des notions de de cohérence temporelle qui sont pas du tout évidentes à gérer. Ton je sais pas, t'es un bonhomme et t'as pas envie qu'il se déforme au cours du temps quoi, même si il aurait généré un peu différemment. Donc c'est gérer une vidéo, c'est pas vraiment générer plein d'images à côté. Bien sûr il faut la cohérence. Il y a eu beaucoup de progrès là-dedans notamment. Donc où justement t'as il calcule un un flot de entre les pixels et donc il vérifie en gros que. Je sais pas que ton que ton bonhomme il se conserve dans le temps quoi. Ouais le temps que je vois l'idée et ça vous même si c'est pas trop dur citoyennes en parler enfin moi ça moi c'est après sinon je regarderai moi, mais le matin je connaissais pas et je connais pas les techniques. Mais voilà si je sais pas si j'en sais rien, il y a un peu de maths à connaître dessus juste pour m'orienter la prochaine fois ça ce serait cool si vous me disiez un peu. Regardez, je suis pas. Pas, c'est pas un, c'est pas inclus dans le cours donc je le mettrais pas dans les slides, mais on peut-on peut en discuter sans problème. Ouais bon ça marche. Moi c'est juste que je vais vous dire s'il y avait un lien. Voilà et merci beaucoup m'a répondu. Super bon bah moi je vous dis à dans 3 semaines et puis et puis bah Voilà ouais bah écoute allez bonne soirée bonne soirée. Il y a d'autres questions Parmi les autres. Ça me demande. Oui et du coup ça sera sur modèle ses indications ? D'accord ? D'accord ou soit informé Léonard vu que c'est lui le chef de Promo. D'accord ? Merci.