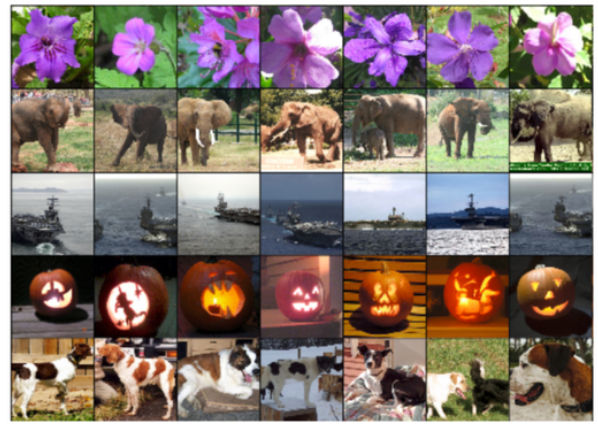
BraInside:
Algorithms for Simplifying Neural Networks

Laurent Viennot et Emanuele Natale

Innovation Defense Lab
10 Janvier 2020

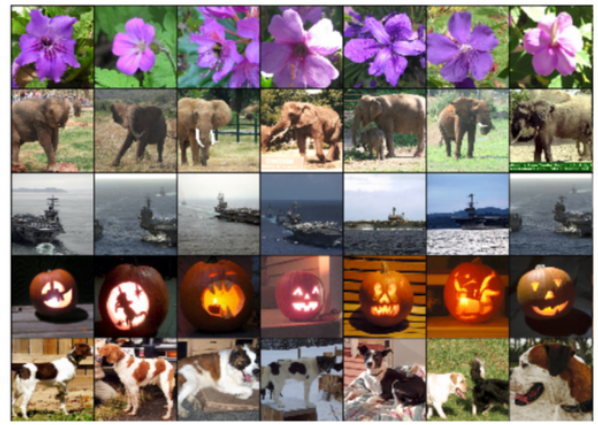
Nécessité de Comprimer les Réseaux de Neurones

Le Deep Learning est (re-)devenu populaire en 2012 quand **AlexNet** a remporté le concours *LSVRC 2012*.

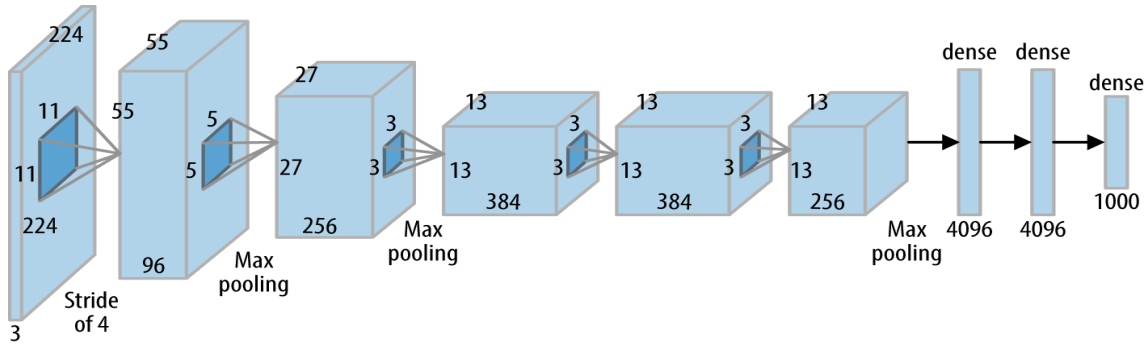


Nécessité de Comprimer les Réseaux de Neurones

Le Deep Learning est (re-)devenu populaire en 2012 quand **AlexNet** a remporté le concours *LSVRC 2012*.

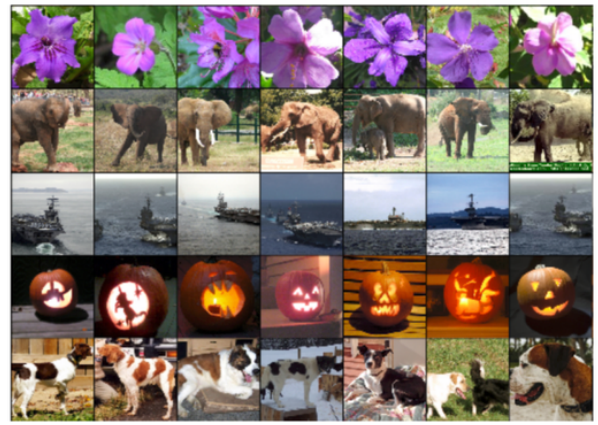


AlexNet a 60 millions de poids (environ 3 Go)

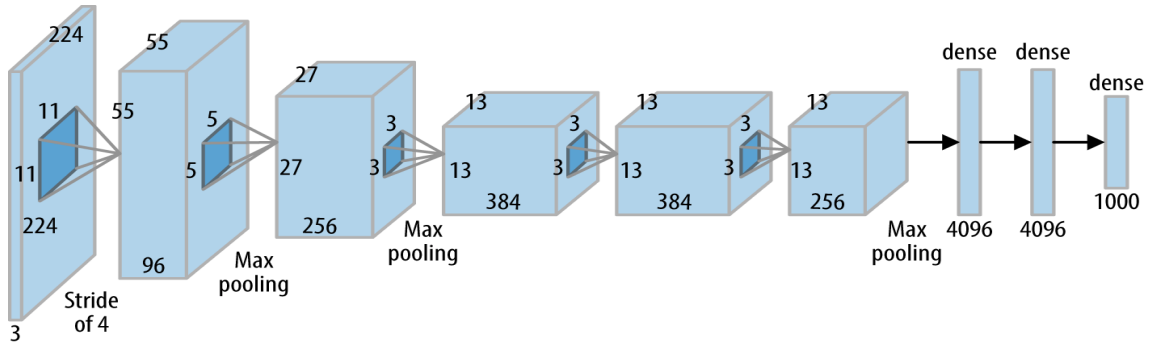


Nécessité de Comprimer les Réseaux de Neurones

Le Deep Learning est (re-)devenu populaire en 2012 quand **AlexNet** a remporté le concours *LSVRC 2012*.



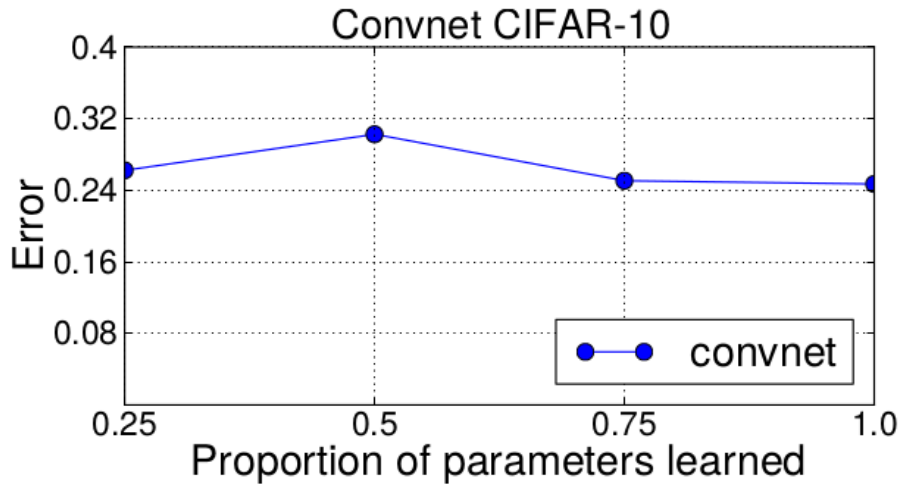
AlexNet a 60 millions de poids (environ 3 Go)



Les architectures de pointe sont “lourdes” pour certains systèmes embarqués

Comment Comprimer les Réseaux de Neurones ?

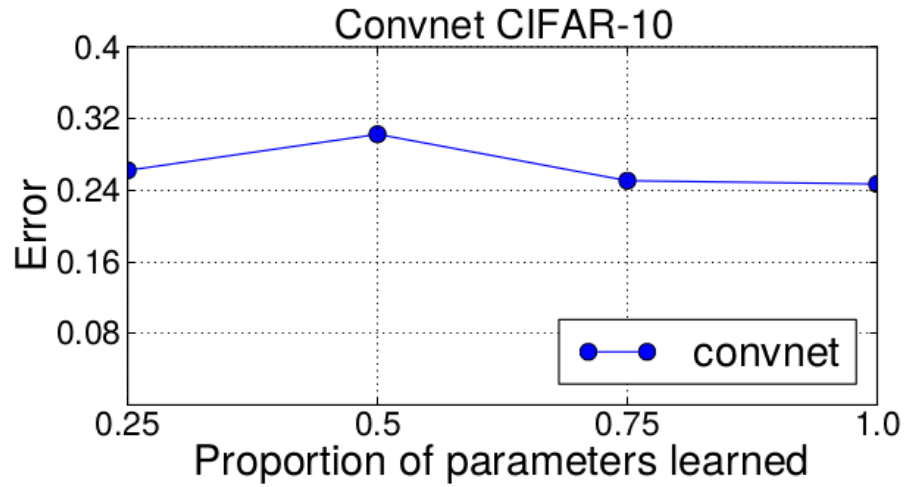
Les réseaux de neurones ont tendance à être très "compressibles"
[Denil et al. '14]



Réalisation d'un convnet sur CIFAR-10.
L'apprentissage de seulement 25% des paramètres a un effet négligeable sur la précision prédictive.

Comment Comprimer les Réseaux de Neurones ?

Les réseaux de neurones ont tendance à être très "compressibles"
[Denil et al. '14]



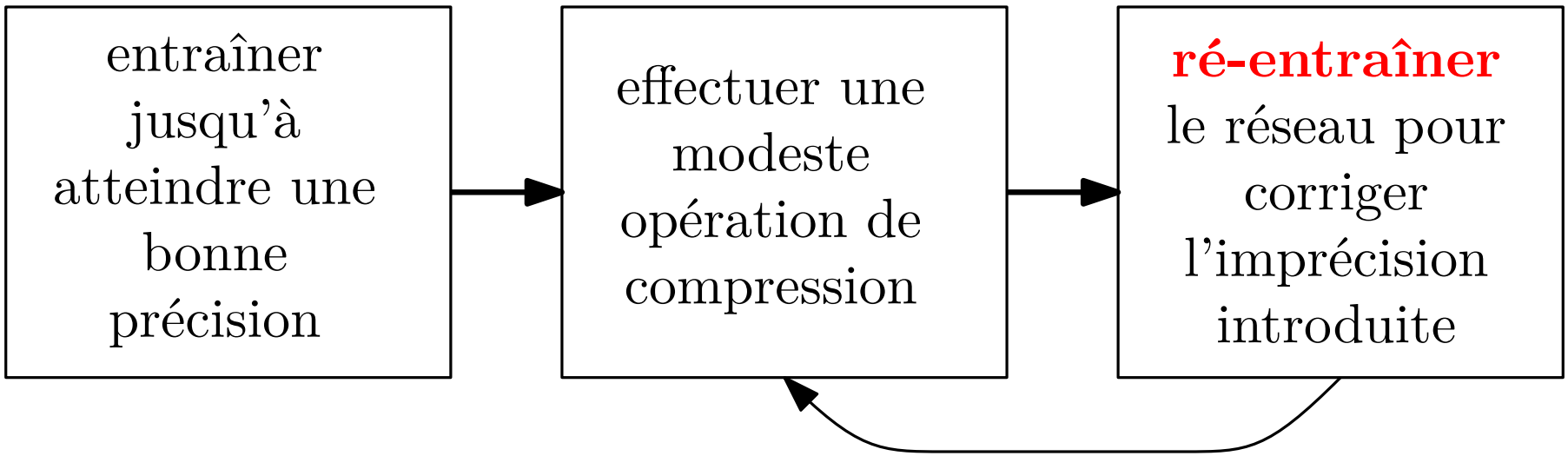
Réalisation d'un convnet sur CIFAR-10.
L'apprentissage de seulement 25% des paramètres a un effet négligeable sur la précision prédictive.

Il n'y a toujours pas de standard pour la compression des réseaux de neurones.

Le groupe MPEG travaille actuellement sur la normalisation de la compression dans ce cadre.

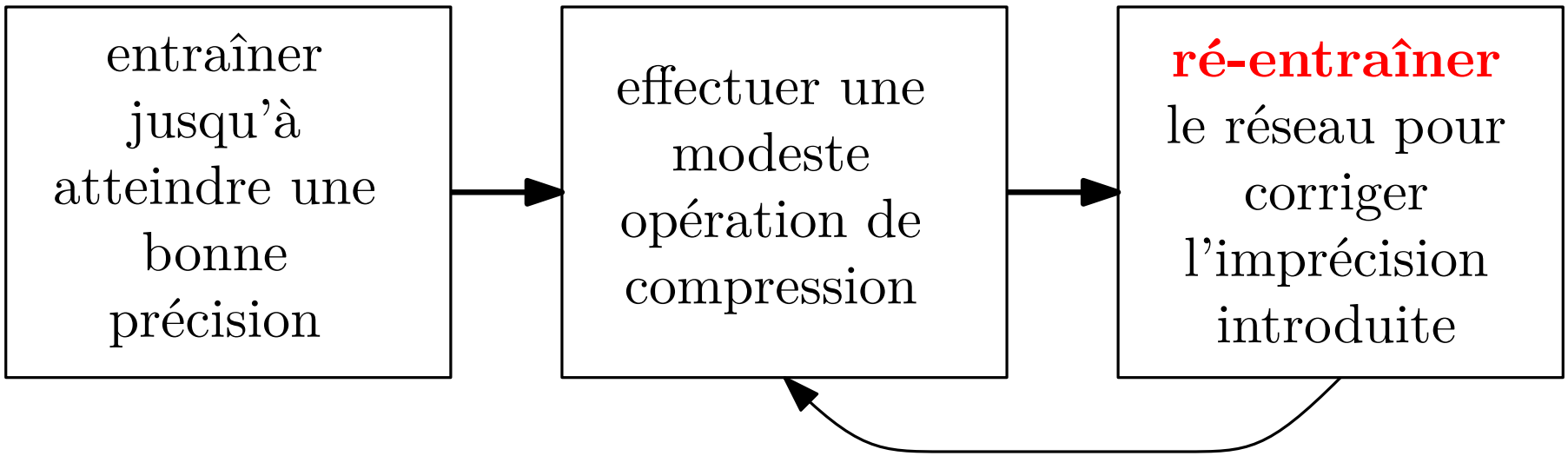
Une Méthode Classique : Élagage par Magnitude

Compression **itérative** pendant l'entraînement

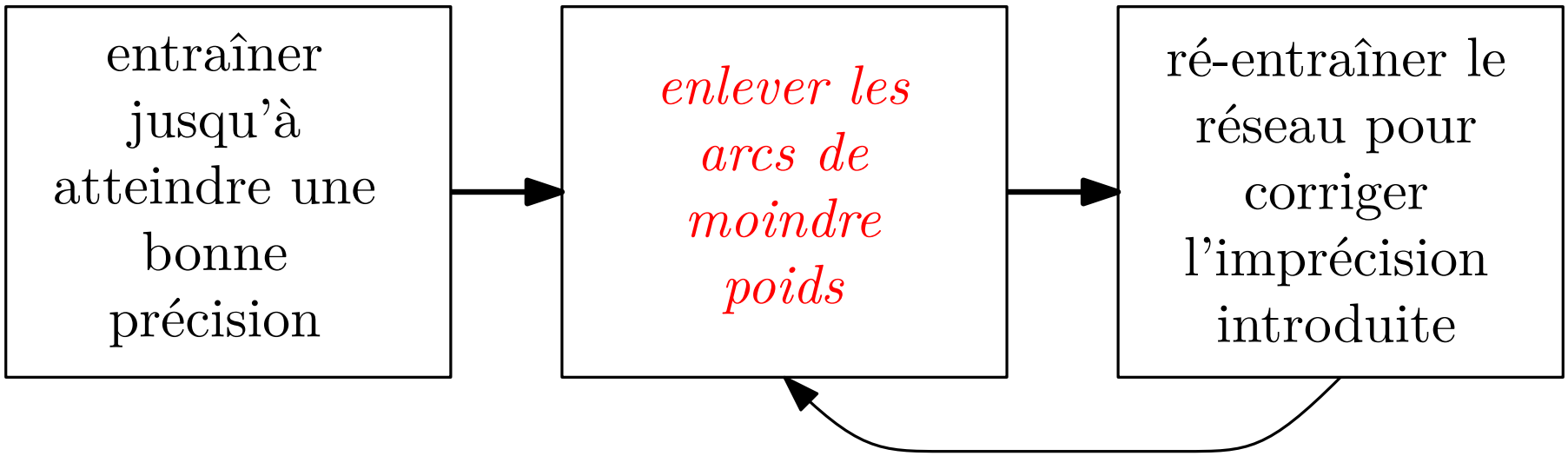


Une Méthode Classique : Élagage par Magnitude

Compression **itérative** pendant l'entraînement



Elagage itératif basée sur la *magnitude*



Autre Méthode: Dommage Cérébral Optimal

[Le Cun et al. 1989]: Enlever les arcs qui causent la moindre perte de précision, par approximation de la fonction objectif

$$\delta E = \sum_i g_i \delta u_i + \frac{1}{2} \sum_i h_{ii} \delta u_i^2 + \frac{1}{2} \sum_{i \neq j} h_{ij} \delta u_i \delta u_j + O(||\delta U||^3)$$

δu_i : composants de δU
g_i : composantes du gradient de E par rapport à U
h_{ij} : éléments de la matrice hessienne par rapport à U

approximation diagonale : l'élimination des arcs affecte δE linéairement ($h_{ij} = 0$ si $i \neq j$)
approximation extrêmale : les éliminations se produisent lorsque l'entraînement a convergé ($\delta u_i = 0$)

Autre Méthode: Dommage Cérébral Optimal

[Le Cun et al. 1989]: Enlever les arcs qui causent la moindre perte de précision, par approximation de la fonction objectif

$$\delta E = \sum_i g_i \delta u_i + \frac{1}{2} \sum_i h_{ii} \delta u_i^2 + \frac{1}{2} \sum_{i \neq j} h_{ij} \delta u_i \delta u_j + O(||\delta U||^3)$$

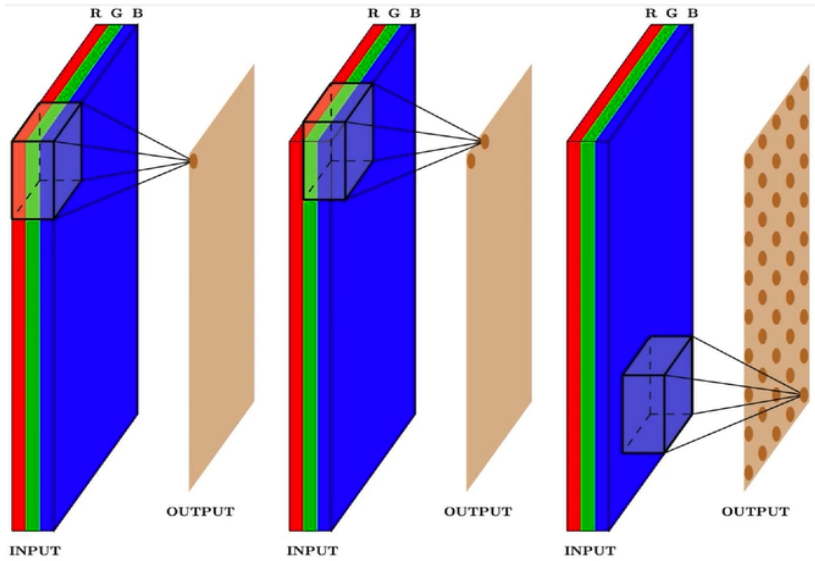
δu_i : composants de δU
g_i : composantes du gradient de E par rapport à U
h_{ij} : éléments de la matrice hessienne par rapport à U

approximation diagonale : l'élimination des arcs affecte δE linéairement ($h_{ij} = 0$ si $i \neq j$)
approximation extrêmale : les éliminations se produisent lorsque l'entraînement a convergé ($\delta u_i = 0$)

Heuristique naturelle pour élaguer,
mais plus coûteuse en termes de calcul.

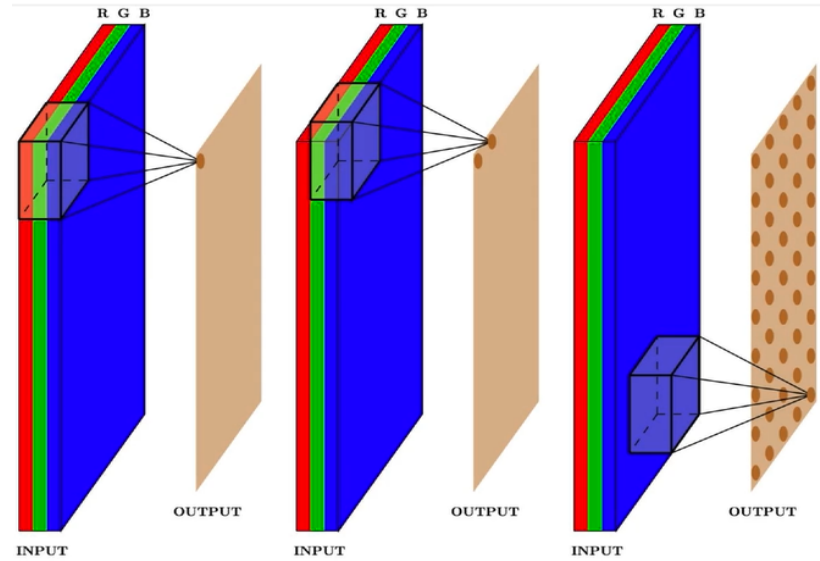
Méthodes de Partage de Poids: HashNet

L'architecture convolutive peut être considérée comme une méthode de partage de poids.

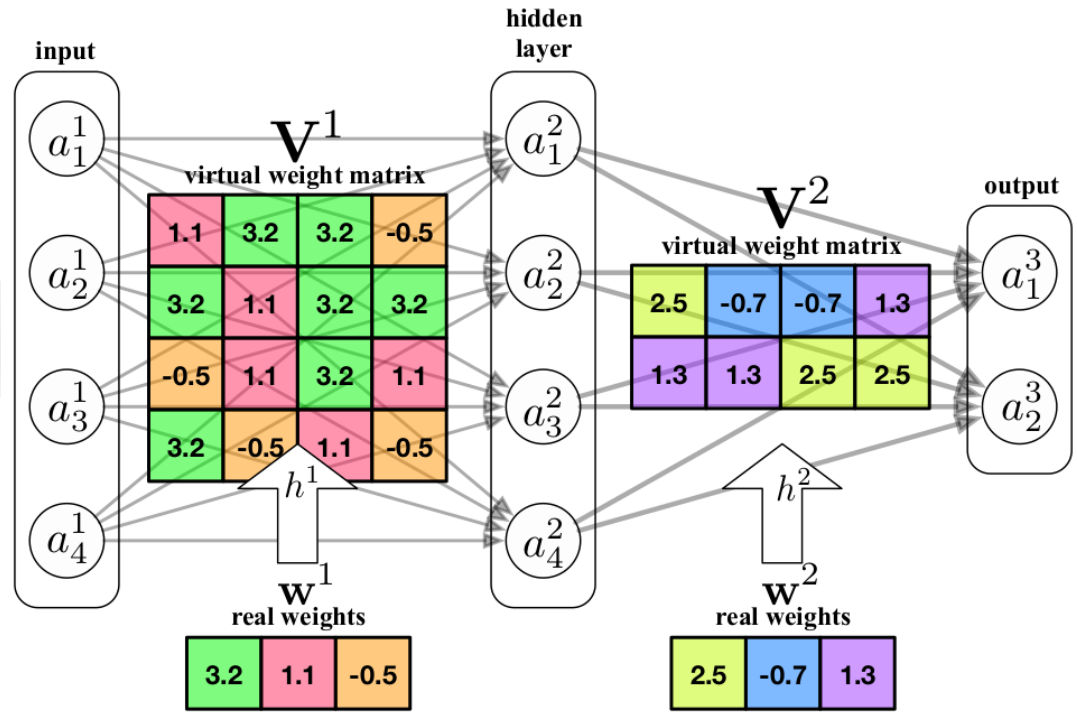


Méthodes de Partage de Poids: HashNet

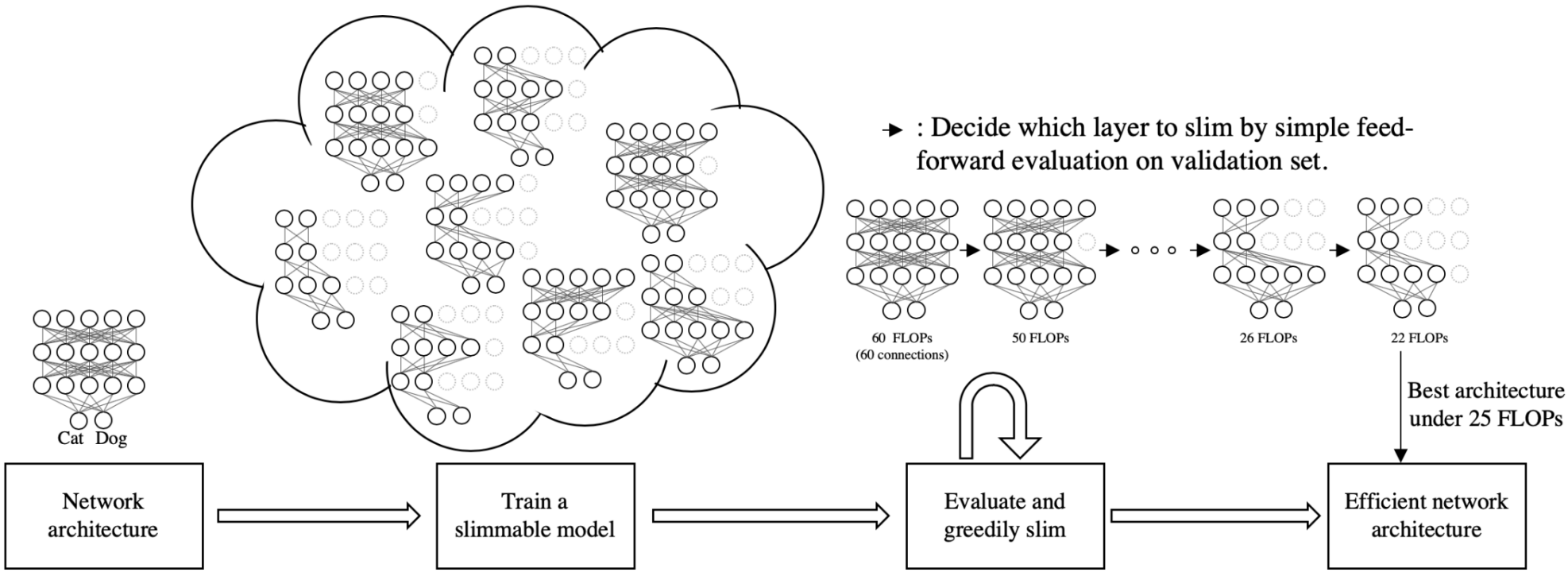
L'architecture convolutive peut être considérée comme une méthode de partage de poids.



HashedNet
[Chen et al.'15]
Forcer plusieurs arcs à *avoir le même poids*.



Une Méthode à la Pointe: AutoSlim

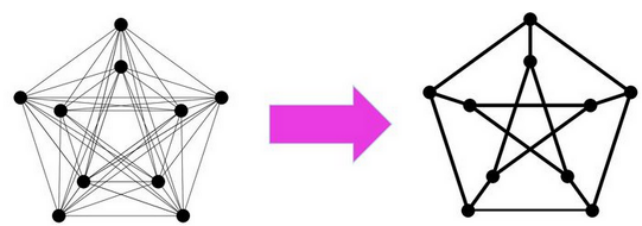


AutoSlim atteint un compromis entre vitesse et précision par une approche unifiée de la recherche de l'architecture de réseau pour différents nombres de canaux.

Méthodes Algorithmiques Connexes

WebGraph et autres frameworks. Plusieurs propositions ont été publiées sur les différentes techniques qui permettent de stocker en mémoire le graphe du Web dans un espace limité, en exploitant les redondances internes du réseau.

Spectral sparsification algorithms. On trouve un sous-graphe du réseau qui approxime les propriétés spectrales de la matrice du réseau



Distance labeling. Prétraiter le graphe en stockant des informations pour chaque sommet afin que les requêtes de plus court chemin puissent être traitées rapidement.

. . . . . .

Déroulement du Projet

Les travaux sur la compression sont souvent fait sur de grands réseaux à la pointe du domaine, à partir desquels il n'est pas facile de comprendre les principes qui sous-tendent l'efficacité de la méthode.

Première Phase : étude comparative sur des *architectures simples*.

Exemple: Baktash H., Natale E., and Viennot L. 2019. "A Comparative Study of Neural Network Compression." https://hal.inria.fr/hal-02321581.

Déroulement du Projet

Les travaux sur la compression sont souvent fait sur de grands réseaux à la pointe du domaine, à partir desquels il n'est pas facile de comprendre les principes qui sous-tendent l'efficacité de la méthode.

Première Phase : étude comparative sur des *architectures simples*.

Exemple: Baktash H., Natale E., and Viennot L. 2019. "A Comparative Study of Neural Network Compression." https://hal.inria.fr/hal-02321581.

Deuxième Phase : Adaptation des méthodes provenant de la communauté algorithmique.

Merci pour votre attention!