Distillation de connaissances dans un réseau de neurones

Une méthode de compression

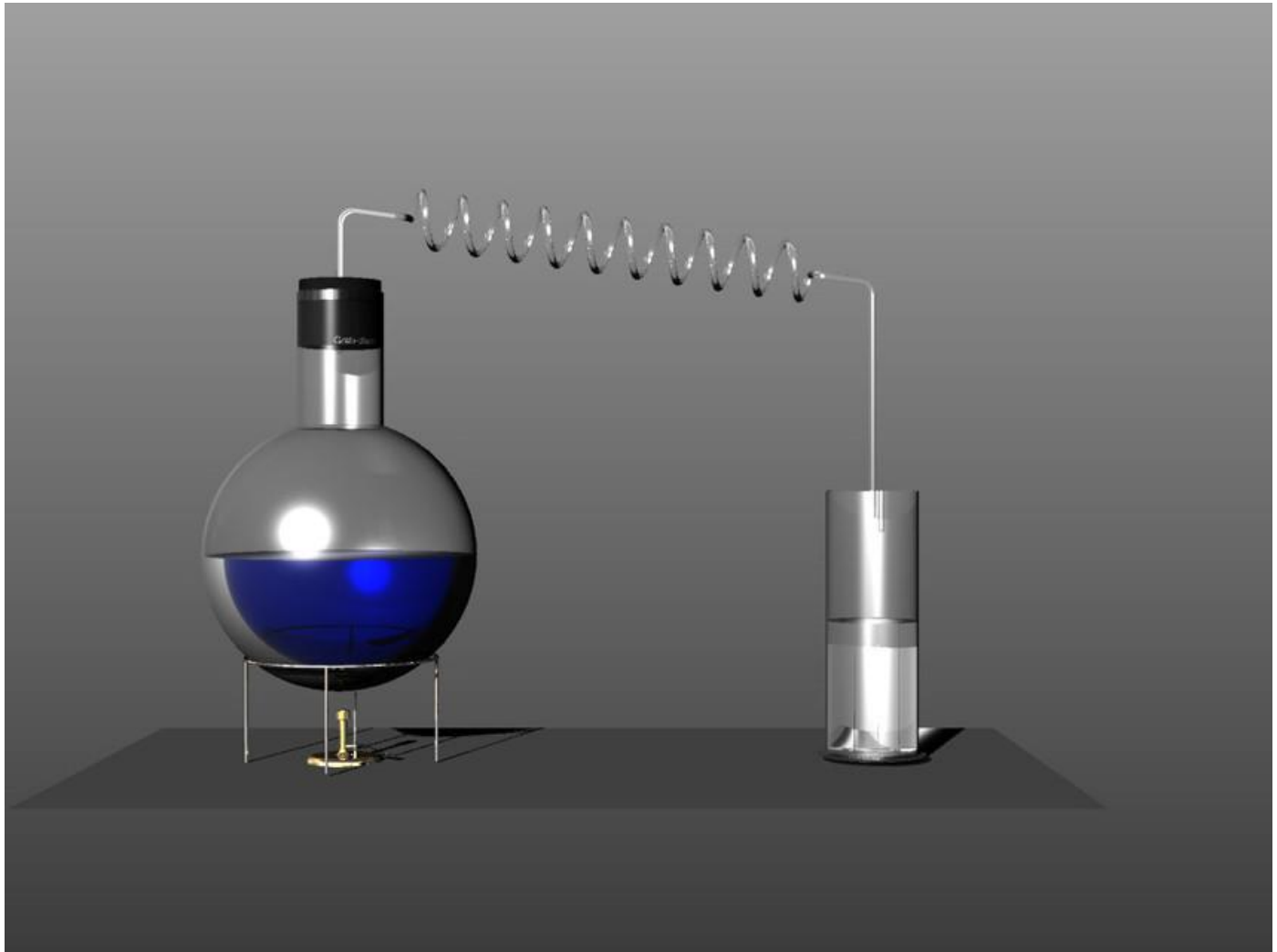
Basé sur l'article : ***Distilling the Knowledge in a Neural Network*** *(2015)* *Geoffrey Hinton* *Oriol Vinyals* *Jeff Dean*

Problématique

- Les modèles les plus puissant sont trop gros pour être utilisé efficacement :
  - Temps d’inférence trop long
  - Trop couteux
  - Trop lourd pour être déployé sur une machine standard

Comment transformer la chenille en papillon ?

DISTILLATION



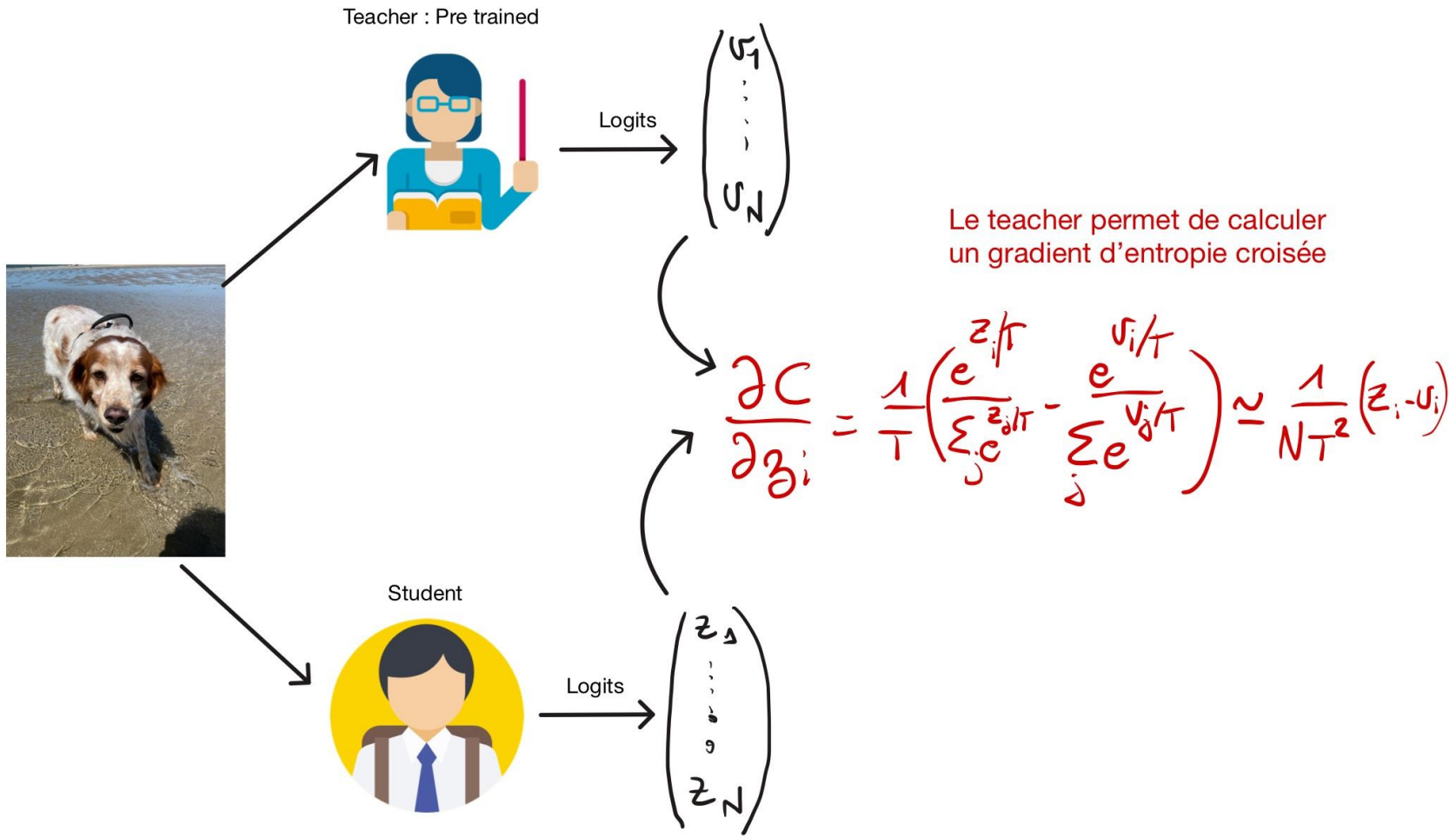
Le principe

Distiller la connaissance d'un model « teacher » vers un model « student »

- Teacher : Model très lourd, ayant d'ingéré beaucoup de connaissances
- Student : Model léger qui synthétise les connaissances du teacher.

Le principe

Hyper paramètre : la température T



Comportement asymptotique

- $\frac{\partial C}{\partial z_i} = \frac{1}{T}\left(\frac{e^{\frac{z_i}{T}}}{\sum_j e^{\frac{z_j}{T}}} - \frac{e^{\frac{v_i}{T}}}{\sum_j e^{\frac{v_j}{T}}}\right)$

A température élevée (par rapport aux logits; $T \gg z_j, v_j$)

- $\frac{\partial C}{\partial z_i} \approx \frac{1}{T}\left(\frac{1+\frac{z_i}{T}}{N+\sum_j \frac{z_j}{T}} - \frac{1+\frac{v_i}{T}}{N+\sum_j \frac{v_j}{T}}\right)$; ddl de l'exponentiel

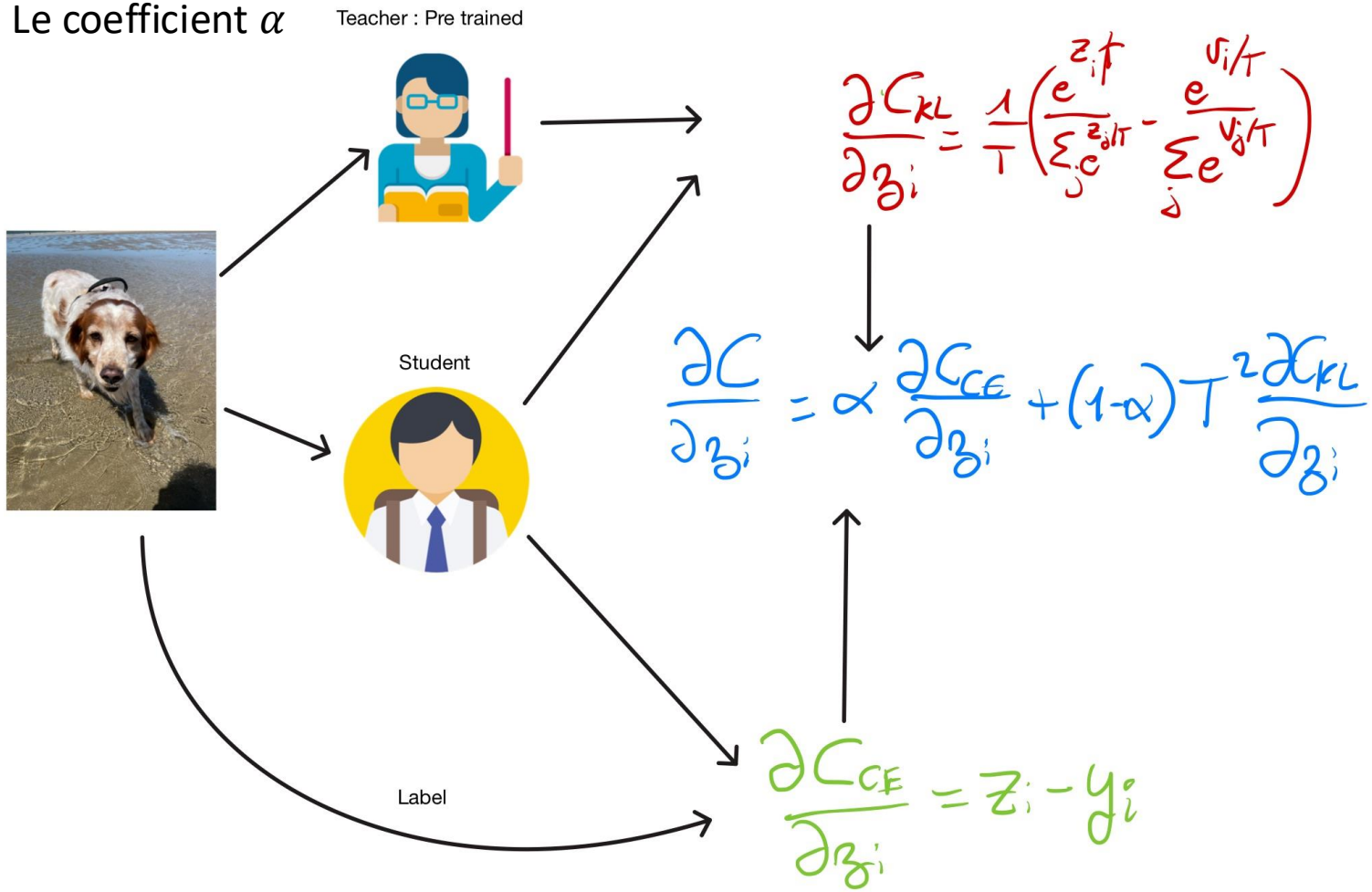
En supposant que les logits sont centrées ($\sum_j z_j = 0$)

- $\frac{\partial C}{\partial z_i} \approx \frac{1}{NT^2}\left(z_i - v_i\right)$

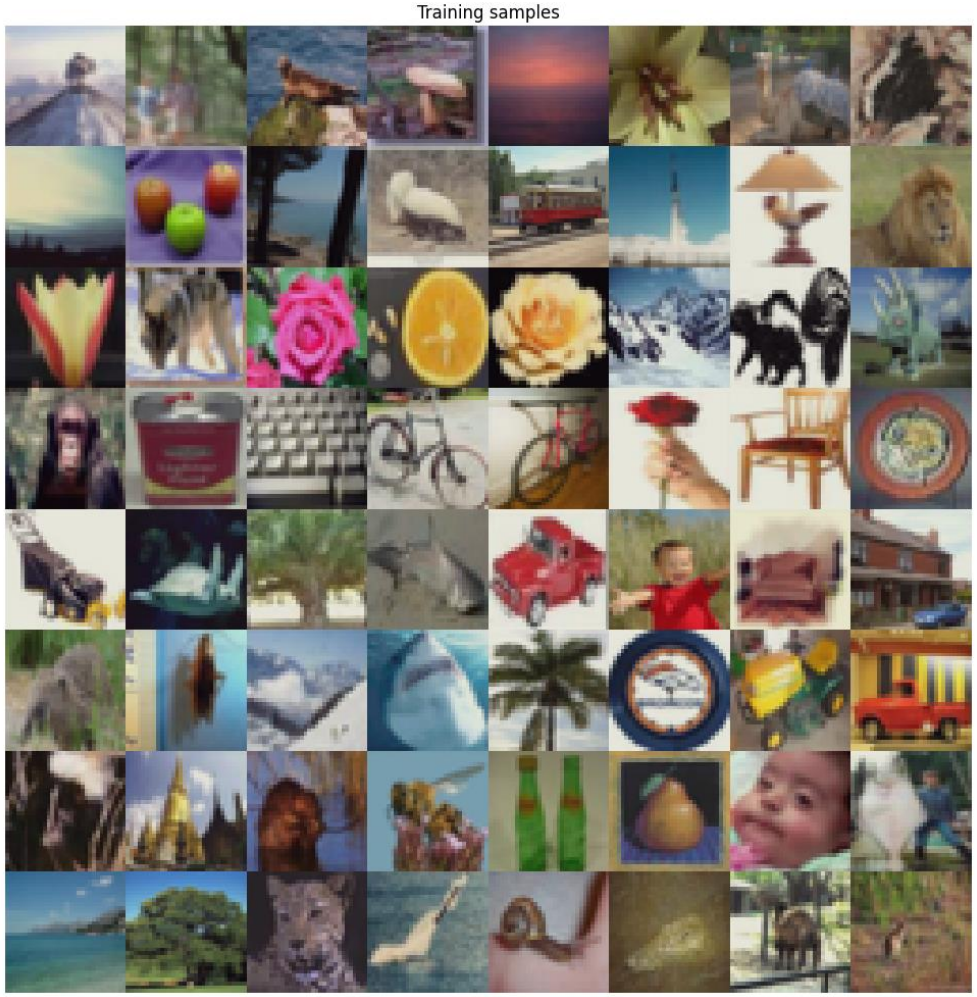
Calcul de la Loss

Deux hyperparamètres :

- La température T
- Le coefficient α



CIFAR100 et RESNET18

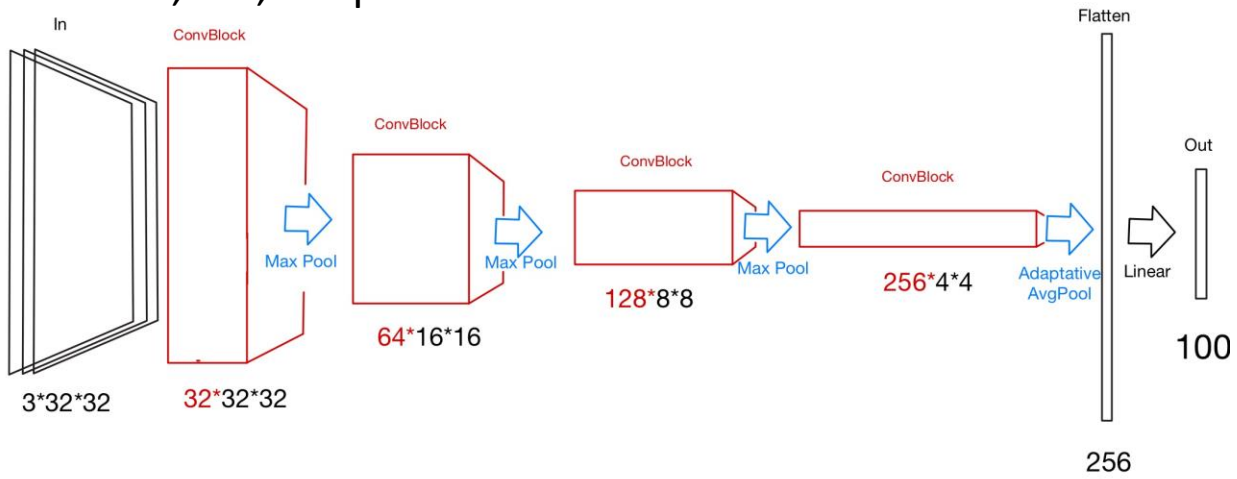


Teacher : ResNet18

- 11,220,132 paramètres
- 80% accuracy, téléchargé depuis la librairie timm

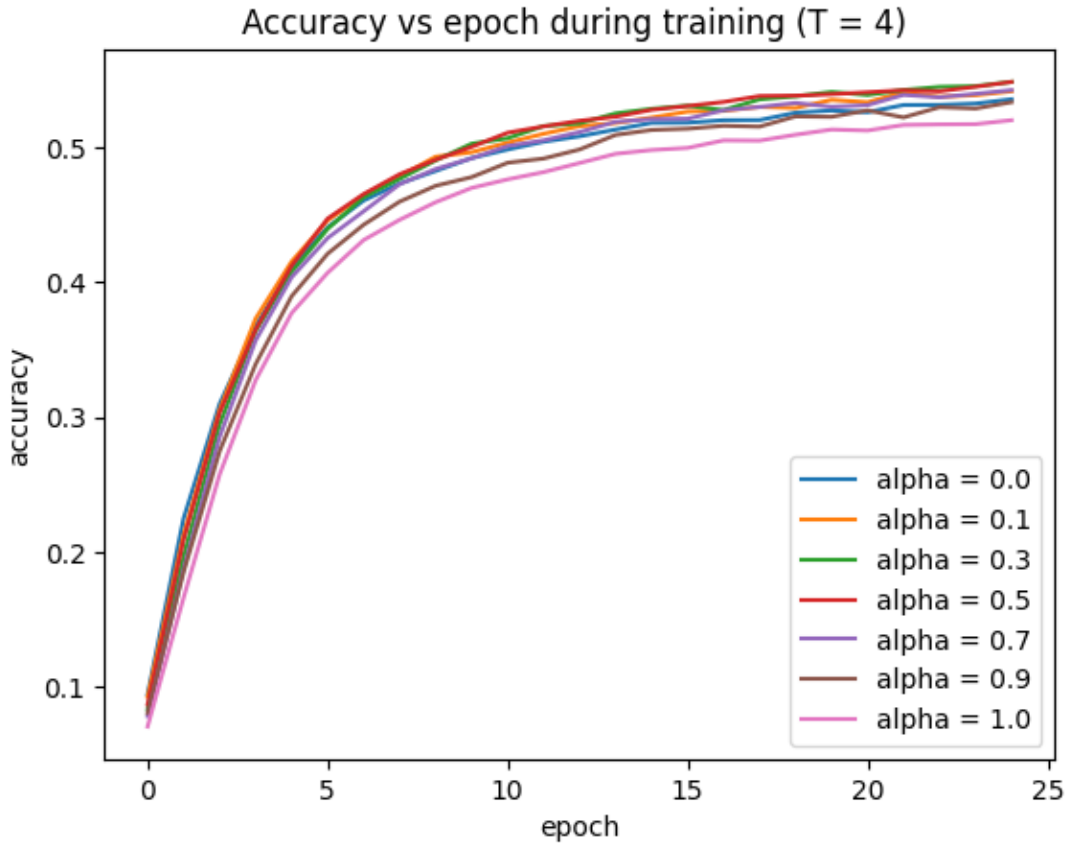
Student : Un « petit » réseau de convolution

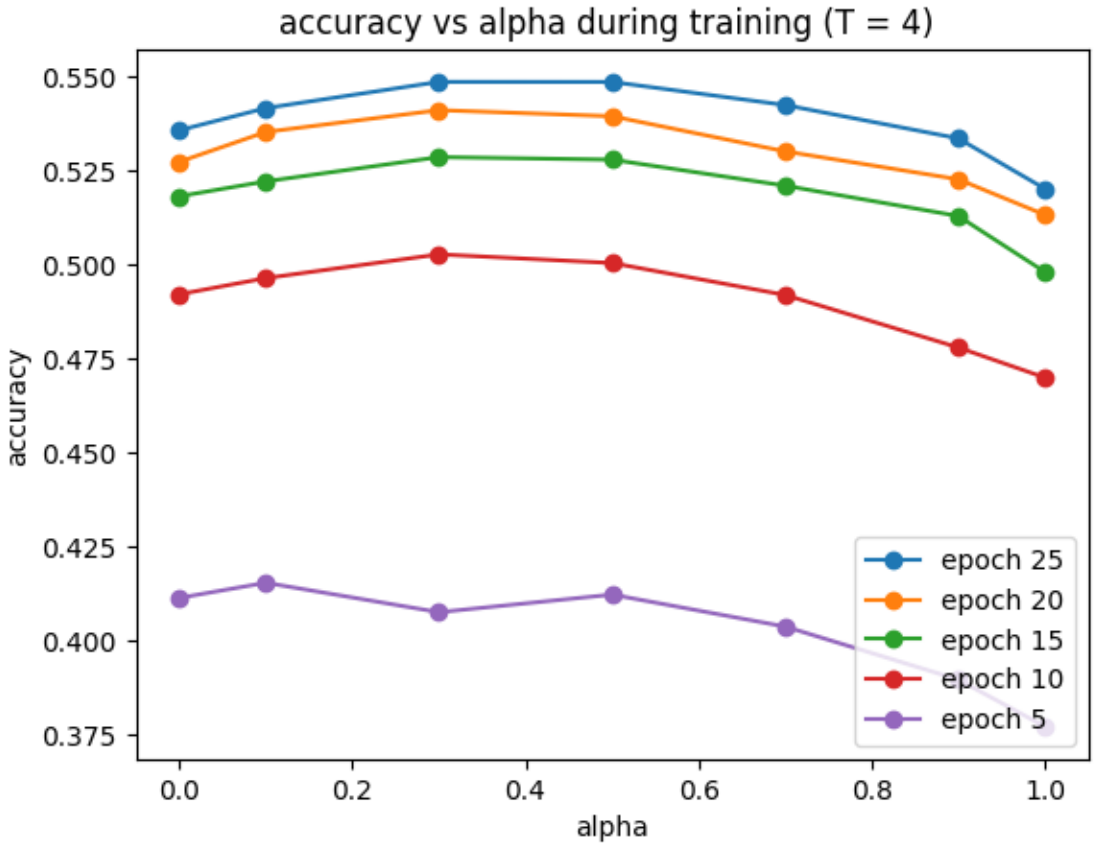
- 1,199,876 paramètres



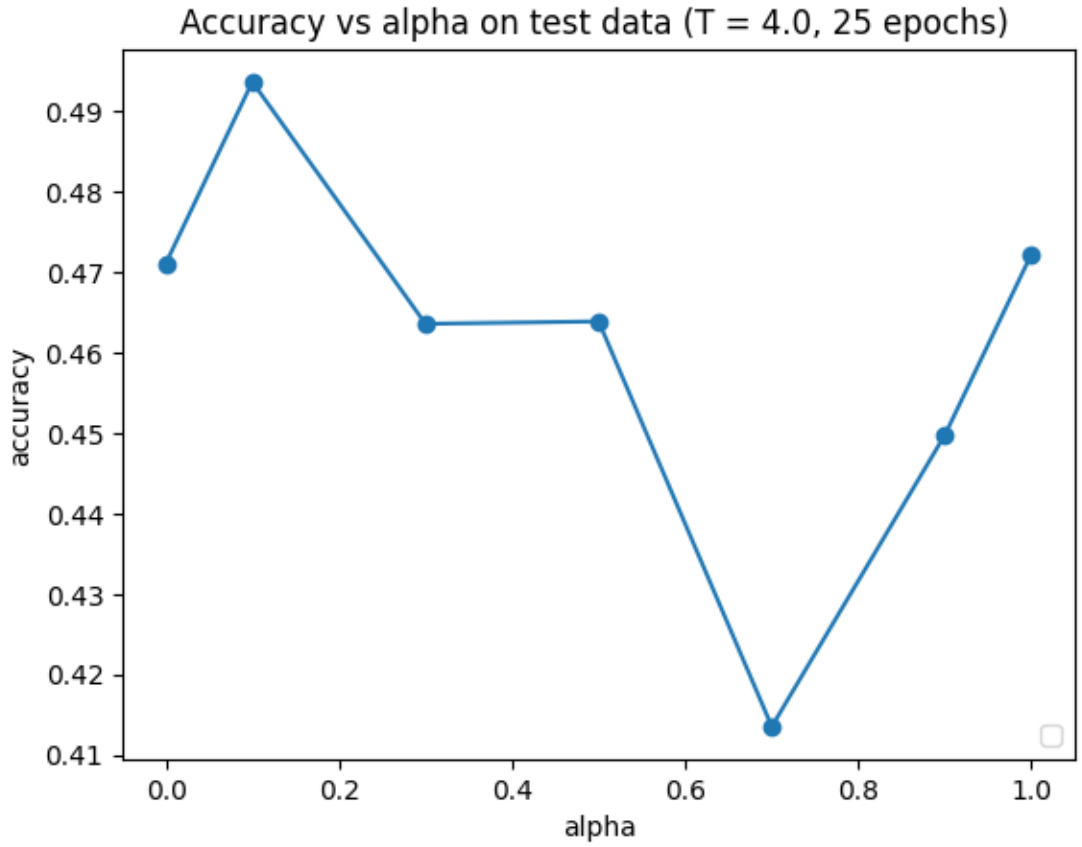
Convblock : convolution -> batchNorm -> Relu ->convolution -> batchNorm -> Relu

CIFAR100 : Effet de α

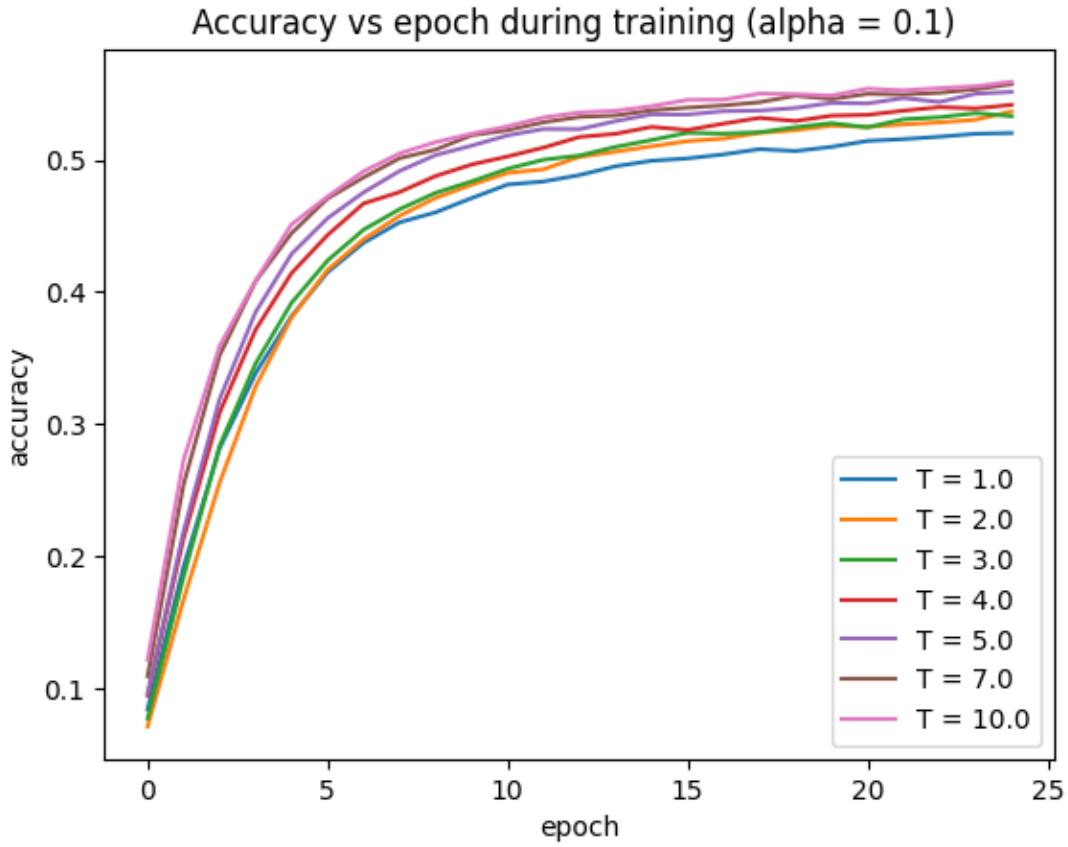


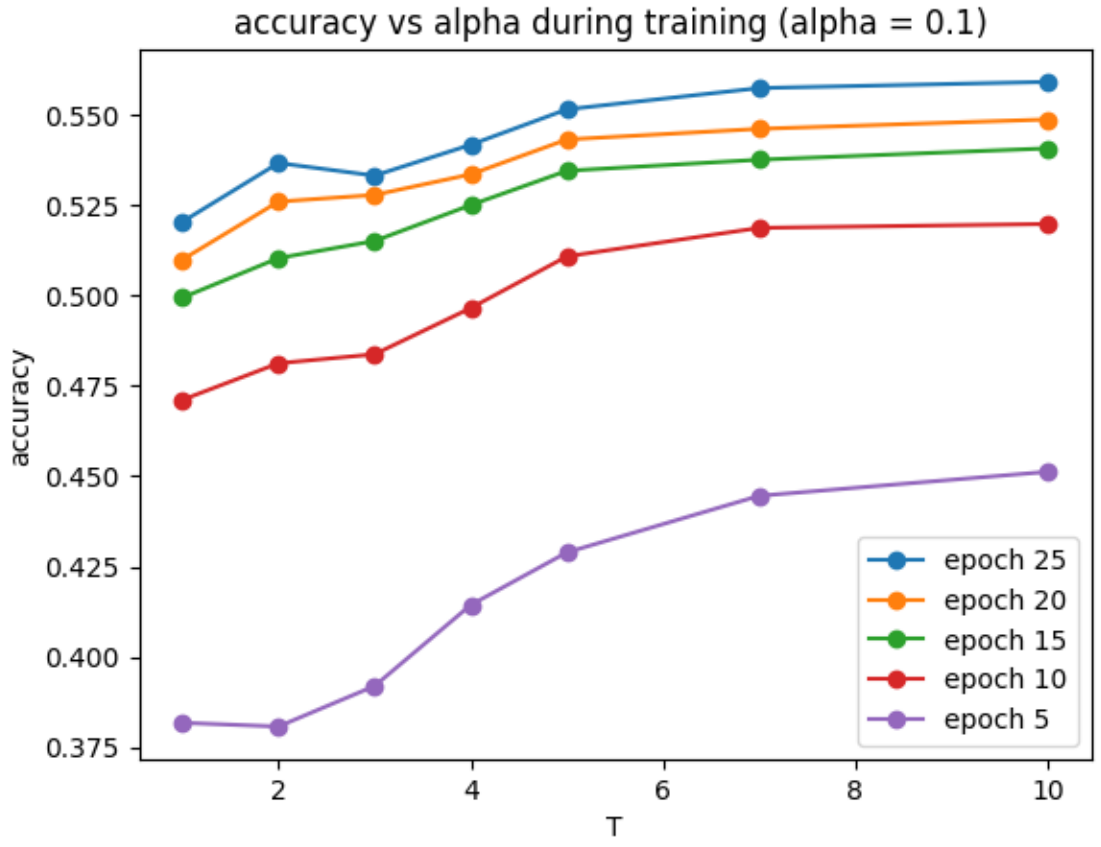


CIFAR100 : Effet de α

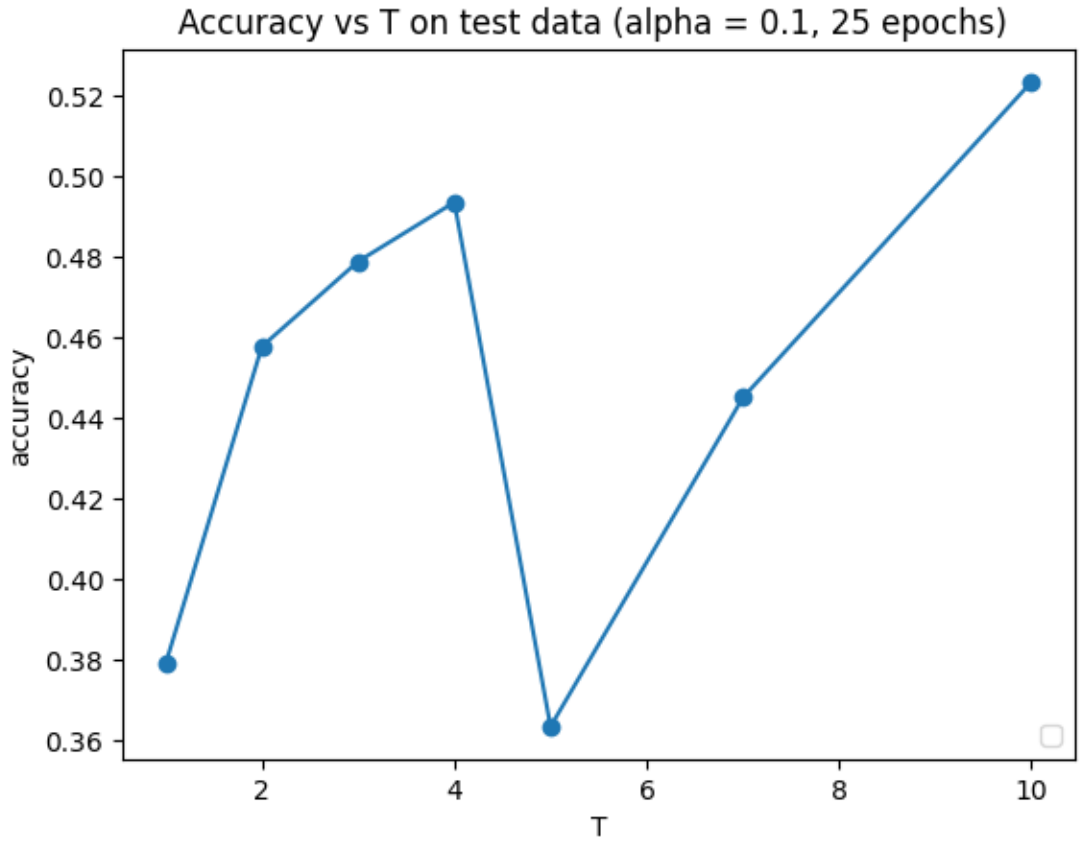


CIFAR100 : Effet de T





CIFAR100 : Effet de T



Pour aller plus loin

- Trois types de connaissances peuvent être distillées :
  - Basées sur les réponses -> *Loss sur les sorties*
  - basées sur les caractéristiques -> *Loss sur les sorties des couches cachées*
  - basées sur les relations -> *Loss sur les relations inter-couches*

Pour aller plus loin

- Schémats de distillations
  - Hors ligne -> *enseignant fixe*
  - En ligne -> *enseignant s'entraine en même temps que l'élève*
  - Auto-distillation -> *Le modèle possède des couches profondes qui enseigne d'autre couches*

Conclusion

- La distillation permet à un réseau :
  - d’apprende plus efficacement.
  - copier un autre réseau
- Ça n’augmente pas la capacité d’un réseau à emmagasiner de la connaissance
- Sensible au réglage des hypers paramètres, leur valeur optimal peut varier en fonction du cas d’utilisation.

Sources

- ***Distilling the Knowledge in a Neural Network*** *(2015) Geoffrey Hinton, Oriol Vinyals, Jeff Dean*
- ***Analysis and Tuning of Knowledge Distillation for efficient Collaborative Learning*** *(2025) Norah Alballa, Ahmed M. Abdeloniem, Marco Canini*
- *Ibm : Qu'est-ce que la distillation de connaissance ?*

FIN