Simple and Fast CNN for Vision

Présenté par Simay CELIK, Aylin SOYKOK, Sarah ENG

ICLR 2025

INTRODUCTION

- Les CNN utilisaient originellement de petits kernels (3x3), mais les informations qu'ils pouvaient capturer ne se limitaient qu'au voisinage.
- Inspirée par la capacité des ViT à capturer les dépendances visuelles à longue portée, certains CNN récents utilisent de larges kernels. Cependant, cette approche est peu compatible avec le matériel et est exigeant en ressources computationnelles.
- Ce papier propose :
  - un nouveau modèle Simple and Fast Convolutional Neural Network (SFCNN), qui utilise des 3x3 kernels, tout en ayant des performances à la hauteur de CNN et ViT de l'état de l'art.
  - une nouvelle fonction d'activation Global Sigmoid Linear Unit (GSiLU) qui capture des informations spatiales globales à grande échelle.
  - plutôt que d'utiliser de grands kernels pour augmenter le champ réceptif, c'est l'augmentation de la récursion qui permet de l'élargir de manière plus efficace.

SFCNN VERSIONS

Le papier propose plusieurs versions de leur modèle afin de trouver la combinaison des paramètres optimale pour la performance.

Model	C	Block Numbers	Expand Ratio
SFCNN-P (Pico)	32	{3,4,12,3}	4
SFCNN-N (Nano)	40	{3,6,17,3}	4
SFCNN-T (Tiny)	48	{4,8,20,4}	4
SFCNN-S (Small)	64	{6,12,28,6}	3
SFCNN-B (Base)	80	{8,15,35,8}	3

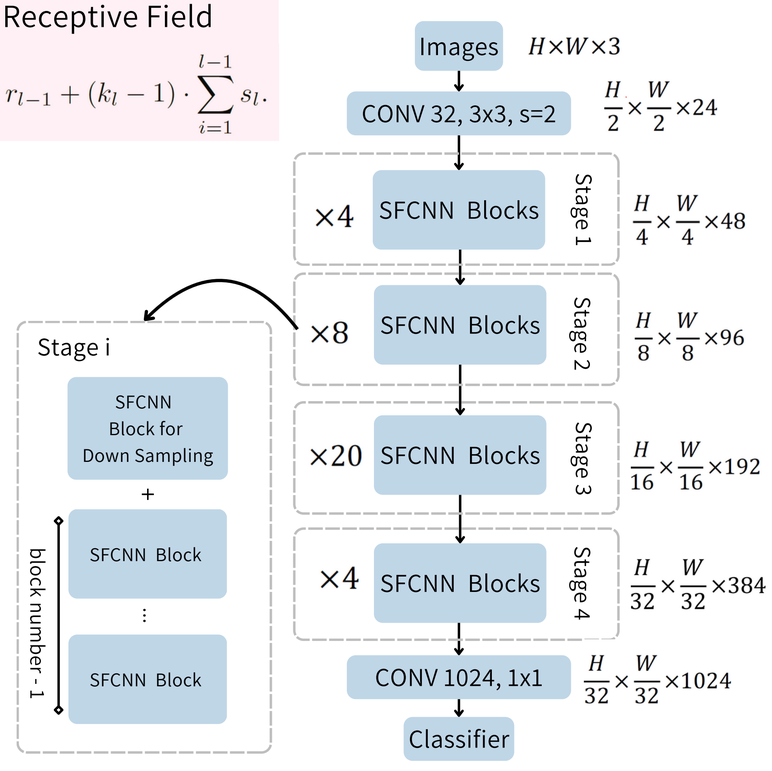
MODÈLES

- SFCNN-T & SFCNN-P avec expansion
- SFCNN-T & SFCNN-P & SFCNN-B sans expansion
- Baseline_CNN: CNN simple avec deux couches convolutionelles (3x3 kernels) et deux couches fully connected.
- ConvNeXt: grands kernels (7x7) et LayerNorm, référence dans le domaine
- ViT: basé sur les Transformers

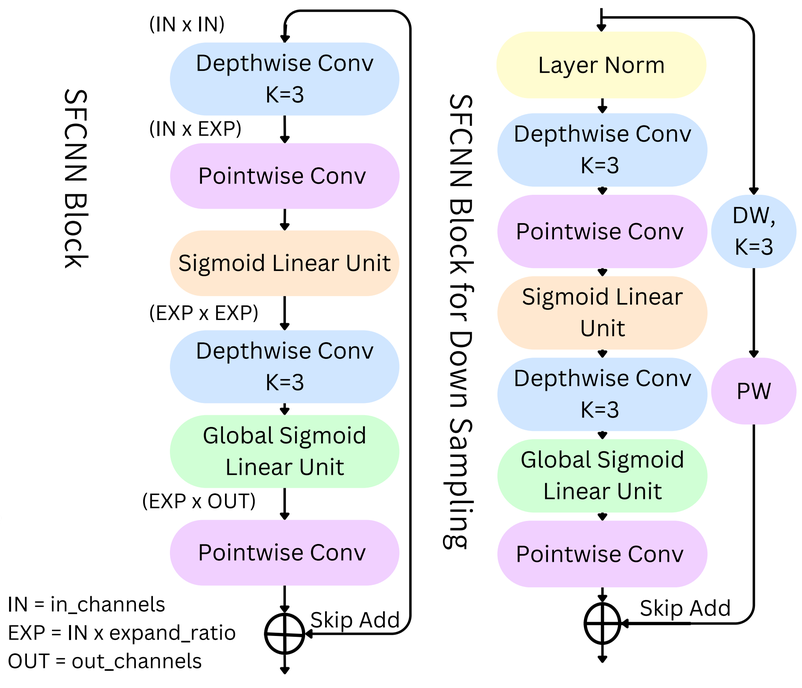
ARCHITECTURE

Receptive Field

$$r_l = r_{l-1} + (k_l - 1) \cdot \sum_{i=1}^{l-1} s_l.$$



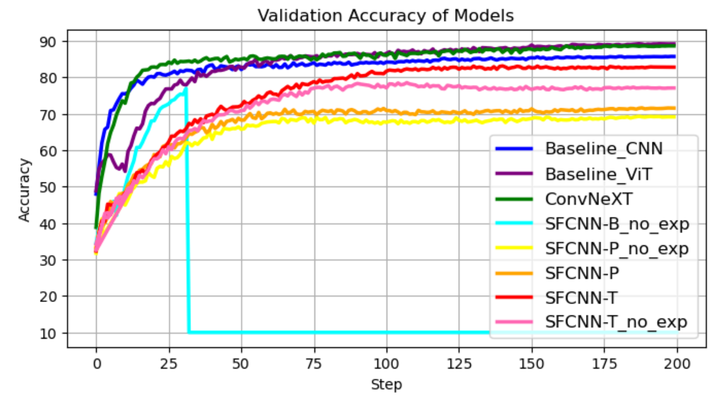
GSiLU (x) = x × σ(GAP (x))



PROTOCOLE EXPÉRIMENTAL

- **Dataset :** Test sur une tâche de classification d'images avec CIFAR-10.
- **Performance :** Comparaison de différentes variantes de SFCNN avec deux baselines et ConvNeXT-Tiny.
  - Entrainement sur la base de donnée CIFAR-10.
  - Évaluation par rapport à l'accuracy, FLOP et le nombre de paramètres.
- **Effets de GSiLU :** Comparaison de performance de SFCNN-P avec et sans GSiLU.

RÉSULTATS



Model	FLOPs	Params	Top-1 Cifar-10(%)
CNN	1.1G	102.8M	85.74
ConvNeXT Tiny	4.5G	29.0M	88.64
ViT	31.1G	9.4M	**89.18**
SFCNN-P_noexp	0.7G	7.7M	-
SFCNN-T_noexp	2.5G	14.9M	79.53
SFCNN-B_noexp	11.3G	50.9M	-
SFCNN-P_exp	2.1G	13.8M	-
SFCNN-T_exp	7.9G	37.6M	82.80
SFCNN-B_exp	39.5G	165.5M	-
SFCNN-P_paper	0.7G	7.7M	-
SFCNN-T_paper	2.4G	16M	-
SFCNN-B_paper	8.7G	49M	-

Table 3: Model FLOPs and Parameters for (224x224) and Top-1 Test Accuracy for Cifar-10 (32x32)

- La différence de performance entre l'utilisation de SiLU seule (69%) et l'ajout de GSiLU (71%) était faible.

CONCLUSION

- Description de l'architecture éparpillée à travers plusieurs sections: difficulté de compréhension.
- Certains éléments techniques (Dropout, entrainement) présents dans le code mais pas dans l'article.
- Moins performant pour un dataset plus simple.