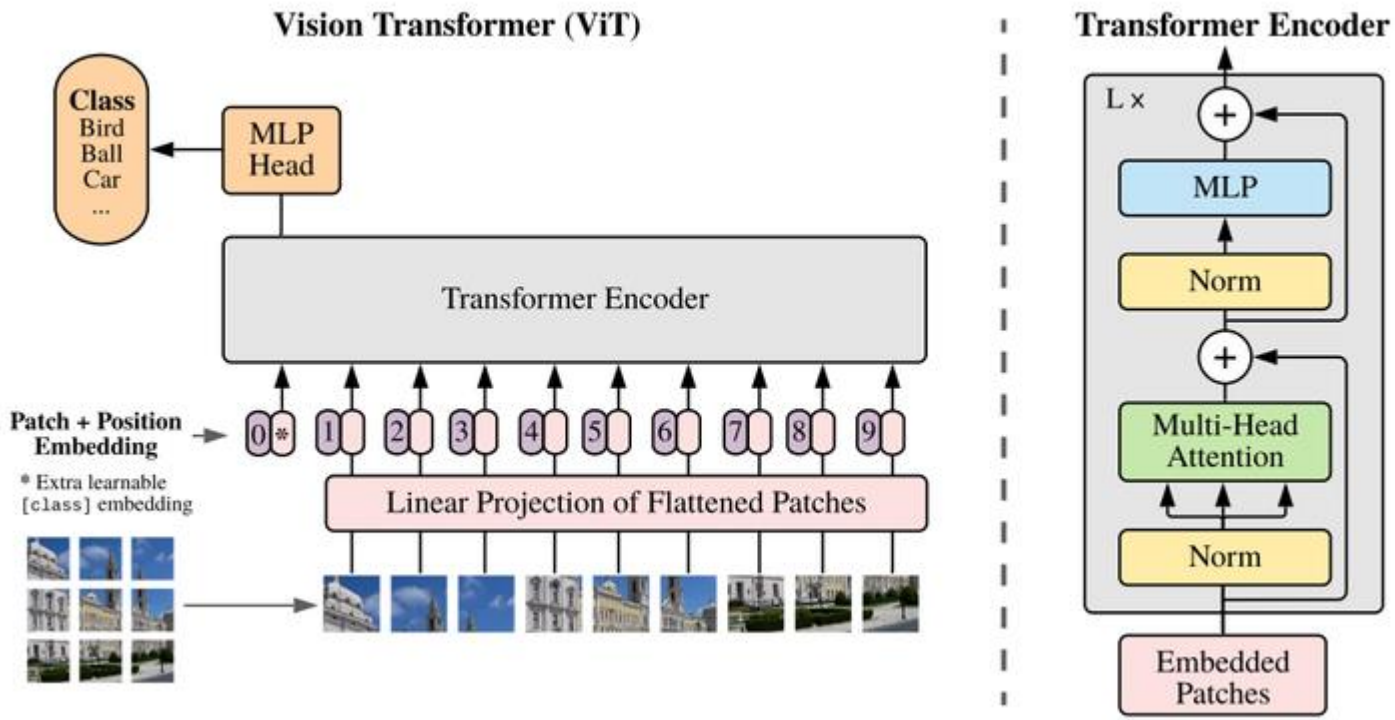
Vision Transformer and attention modules

Antoine DIEZ LATTEUR

V2

ViT Principe de fonctionnement



Self-attention in ViT

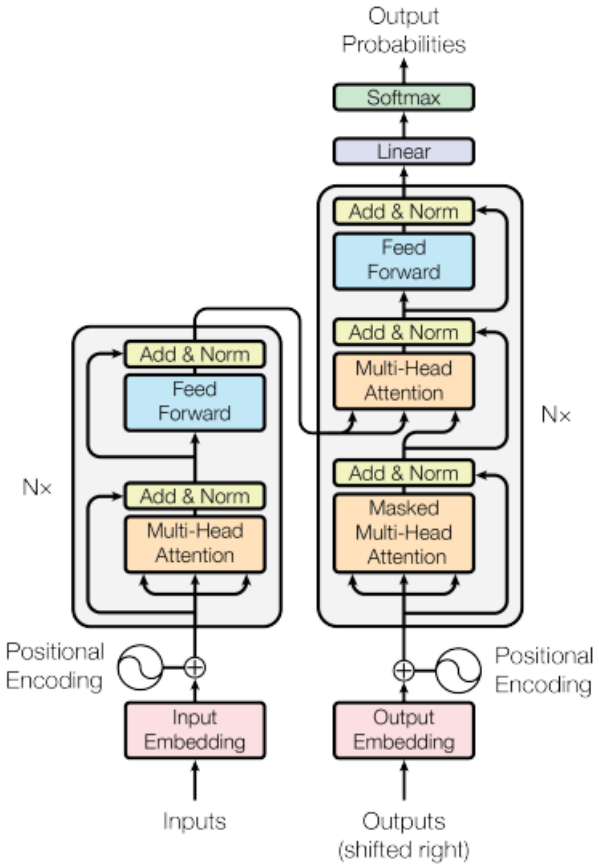
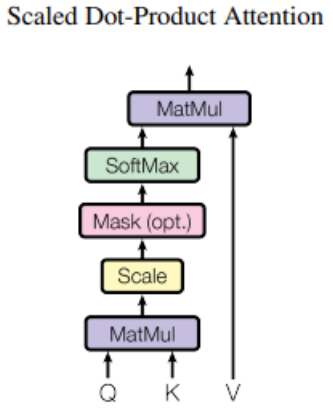
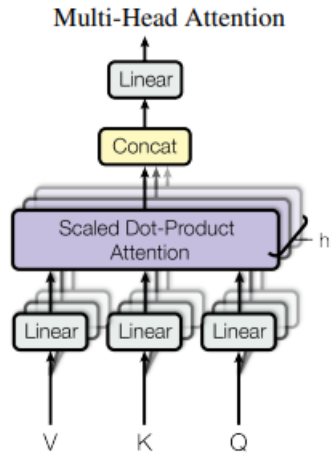


Figure 1: The Transformer - model architecture.





Dataset MSTAR

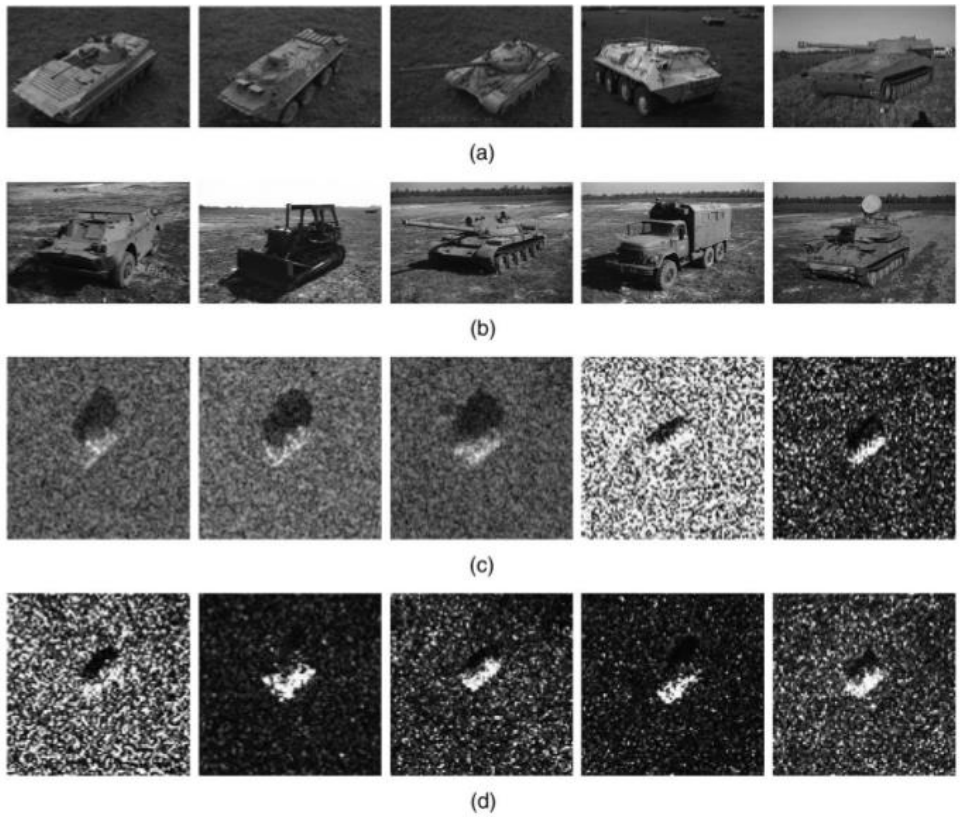


Fig. 2 MSTAR database. (a) and (b) Visible light images for BMP2, BTR70, T72, BTR60, 2S1, BRDM2, D7, T62, ZIL131, and ZSU23/4. (c) and (d) Corresponding SAR images for 10 targets measured at azimuth angle of 45 deg.

Nombre d'images : 2740

Nombre de classes : 10

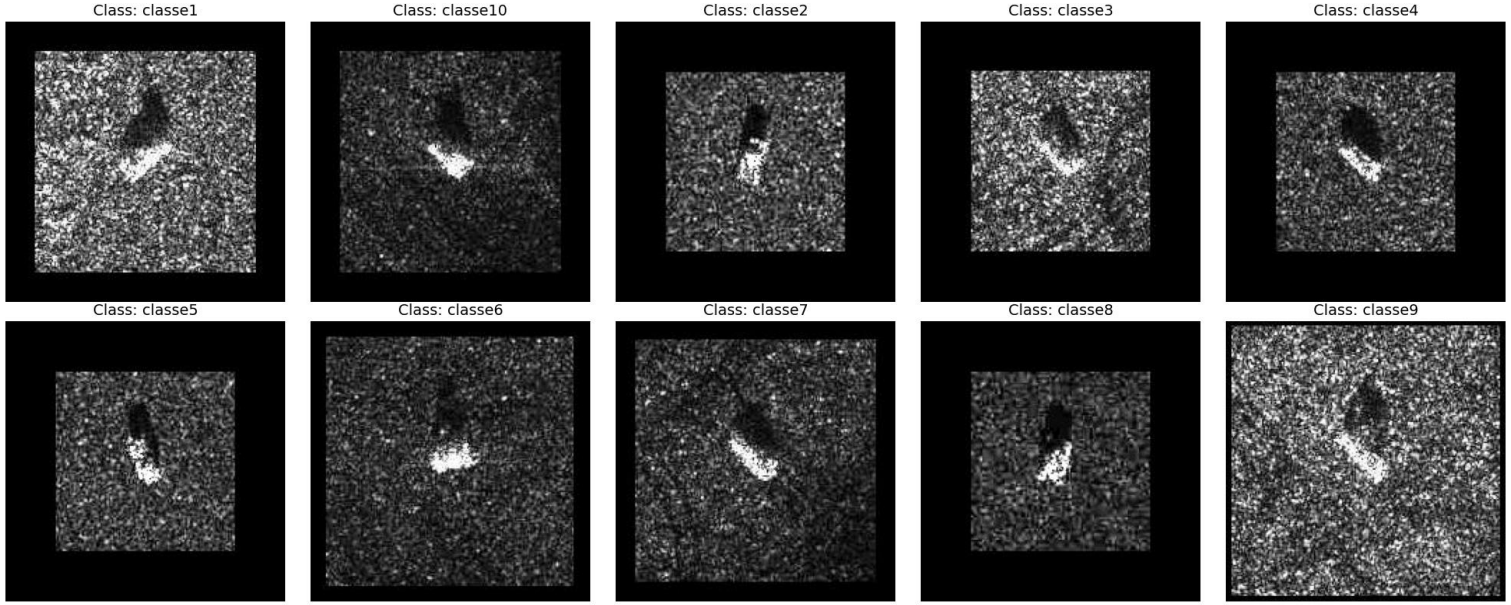
Tailles des images : 200x200 (après pré-traitements)

Images par classes	
Classe 1	292
Classe 2	233
Classe 3	298
Classe 4	256
Classe 5	233
Classe 6	299
Classe 7	299
Classe 8	232
Classe 9	299
Classe 10	299

Pre traitement

Transforms :

- Grayscale
- Resize 224x224
- Horizontal Flip
- Normalize



Our ViT model

86M paramètres

Têtes d’attention : 12

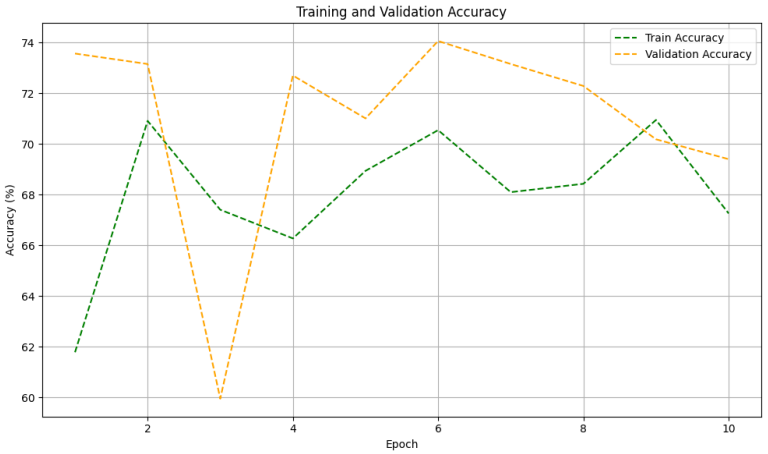
Blocs transformers : 12

Patchs : 16x16

Bloc	Description	Dimensions
Input	Image d'entrée : 224x224x3	224x224x3
Patch Embedding	Divise l'image en patches de 16x16 et les projette en un espace de dimension 768	14x14x768 (patches)
Position Embedding	Ajoute un encodage de position à chaque patch	197x768
Class Token	Un token spécial ajouté pour la classification	1x768
Transformer Layers (12 blocs)	Chaque bloc contient : - Multi-head Self-Attention (12 heads) - Norm Layer - MLP Layer (2 linéaires, GELU)	197x768 par bloc
Layer Norm (LN)	Normalisation des couches	197x768
Classification Head	Projection linéaire pour obtenir les classes finales	1000 (sortie finale)

ViT full training

Hyperparamètres	valeurs
Batch size	32
Epoch	10
Learning rate	1e-3
Optimizer	Adam
Loss function	Cross Entropy Loss



Les résultats ne sont pas concluant

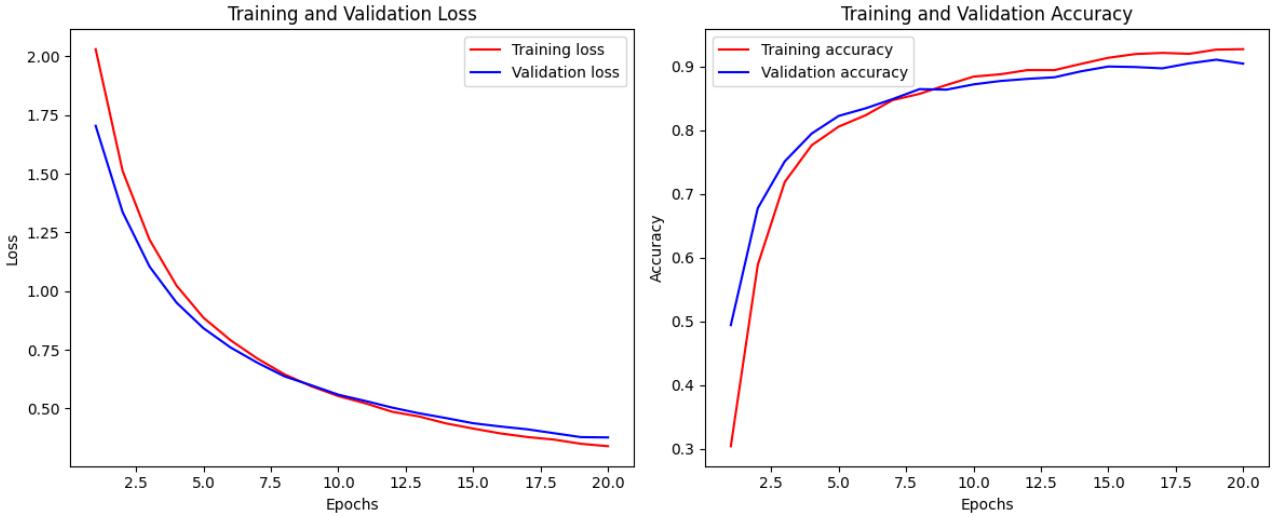
Les ViT nécessites de très grosses bases de données d’apprentissage

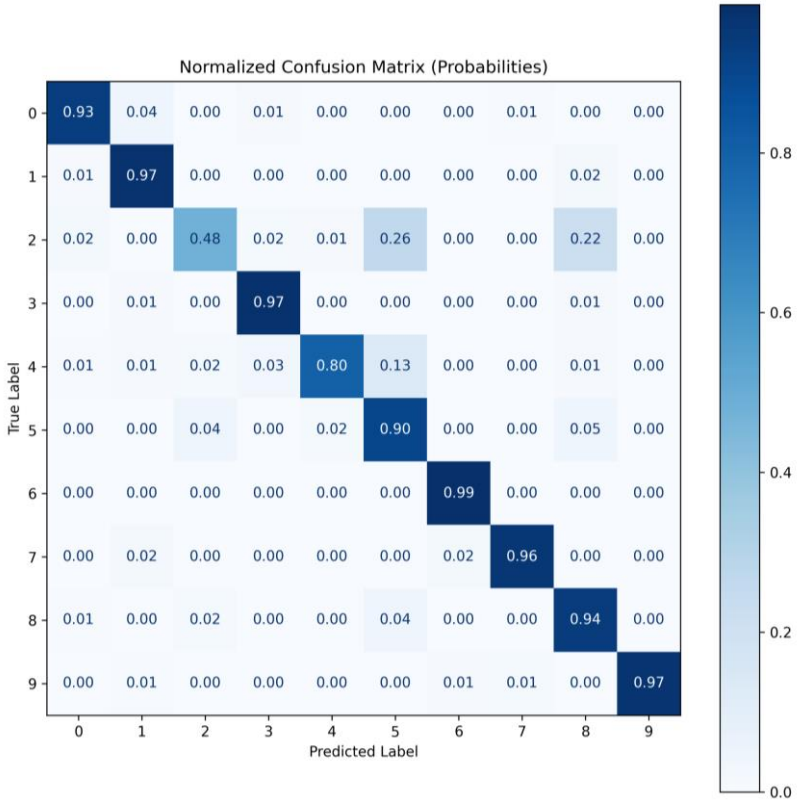
ViT fined Tuning of the classification Head

Hyperparamètres	valeurs
Batch size	32
Epoch	20
Learning rate	1e-4
Optimizer	Adam
Loss function	Cross Entropy Loss

Validation accuracy : 90%

Major issues with class 2



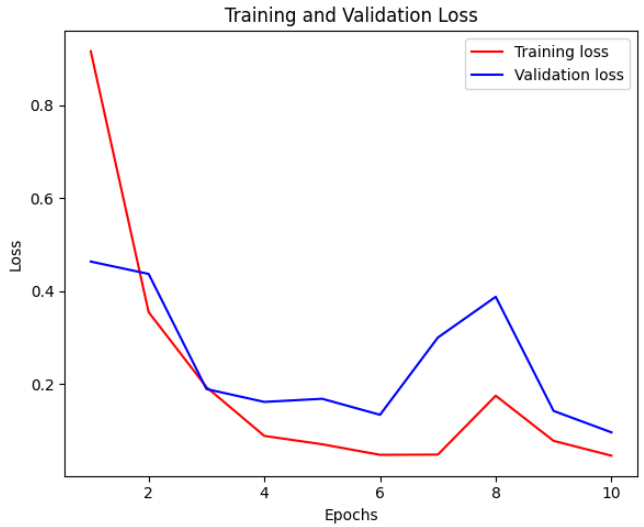


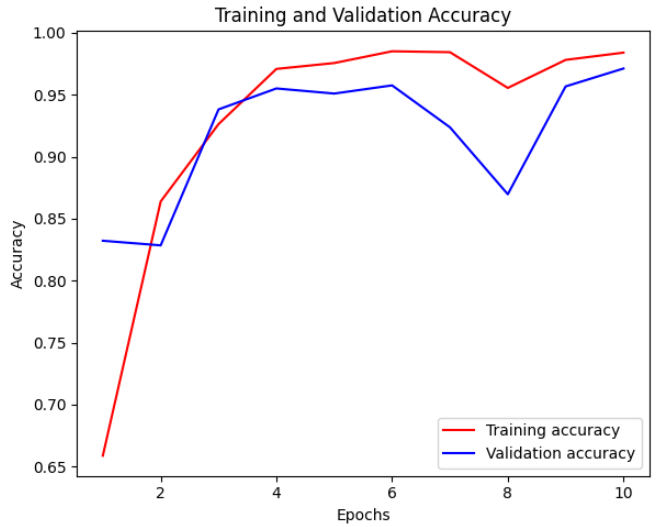
ViT Fined Tuned

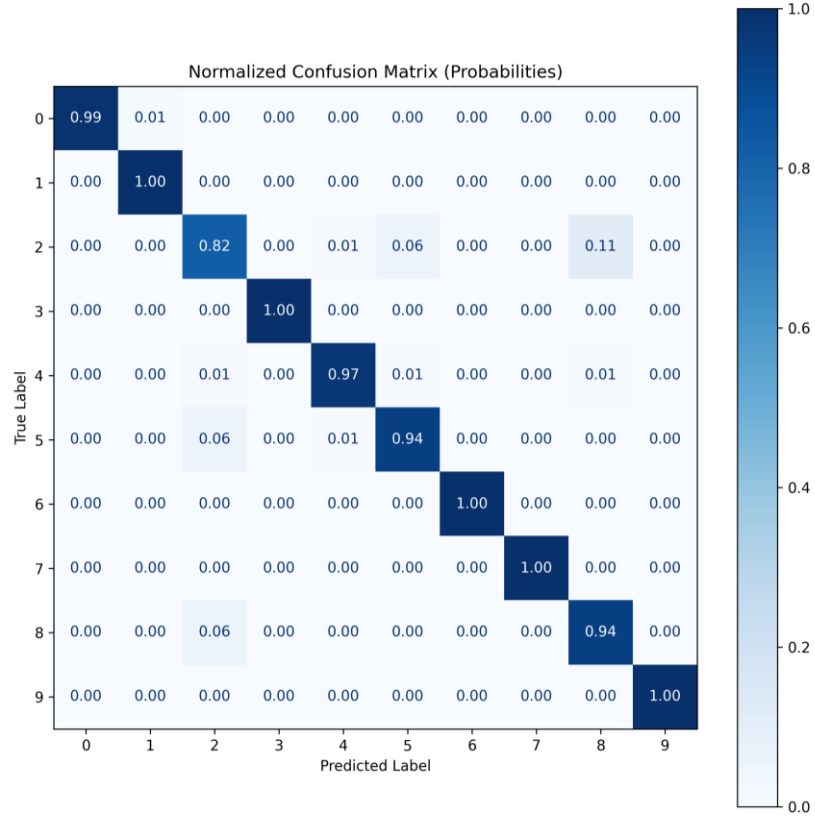
Hyperparamètres	valeurs
Batch size	32
Epoch	10
Learning rate	1e-4
Optimizer	Adam
Loss function	Cross Entropy Loss

Validation accuracy : 96%

Issues with class partially corrected







Attention Modules

- Channel attention
- Spatial attention

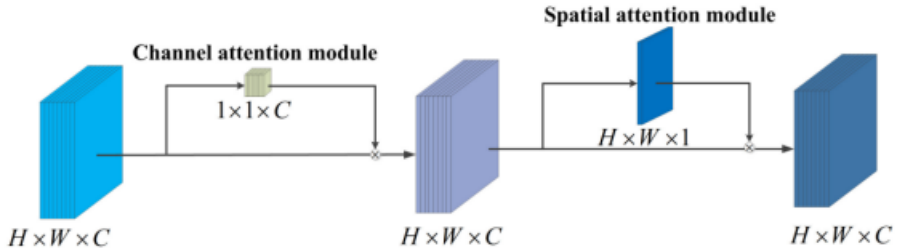


FIGURE 4.2 – FCAB proposé.

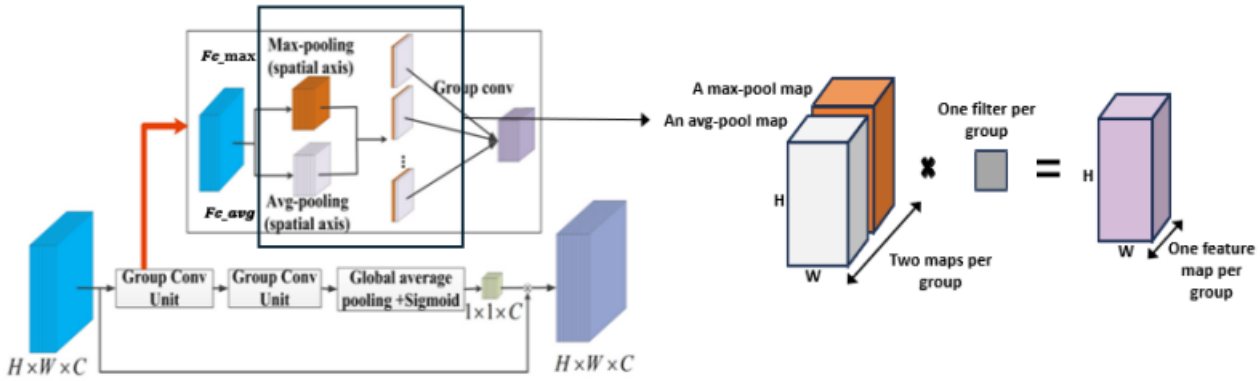


FIGURE 4.3 – Module d'attention par canal.

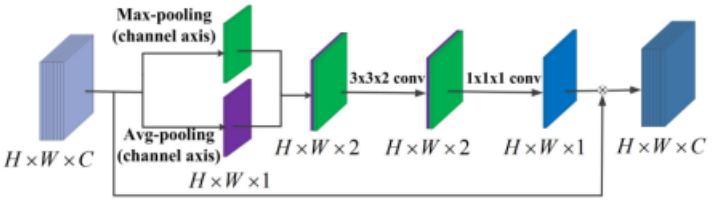
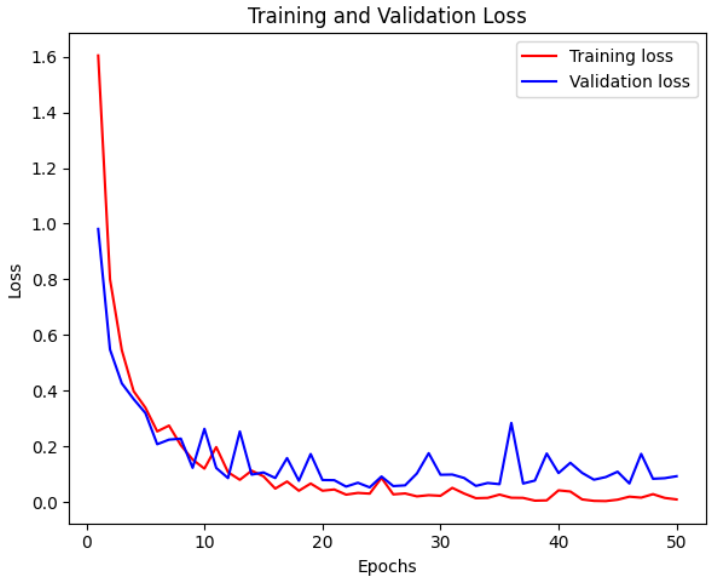
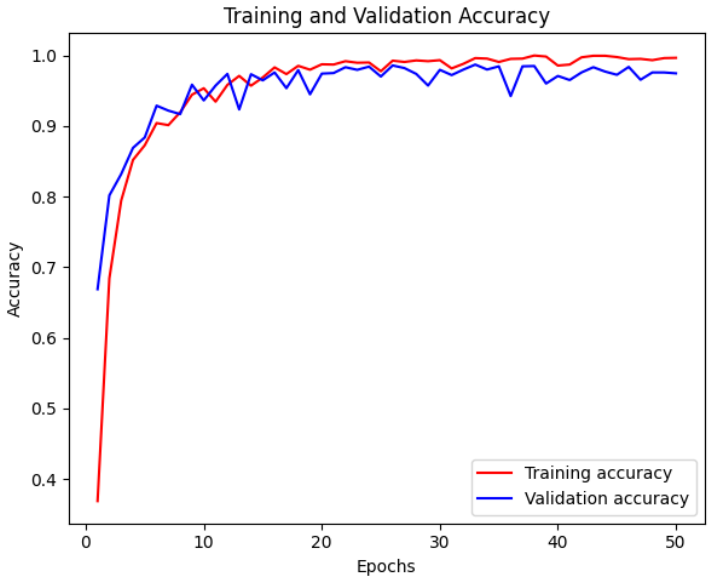
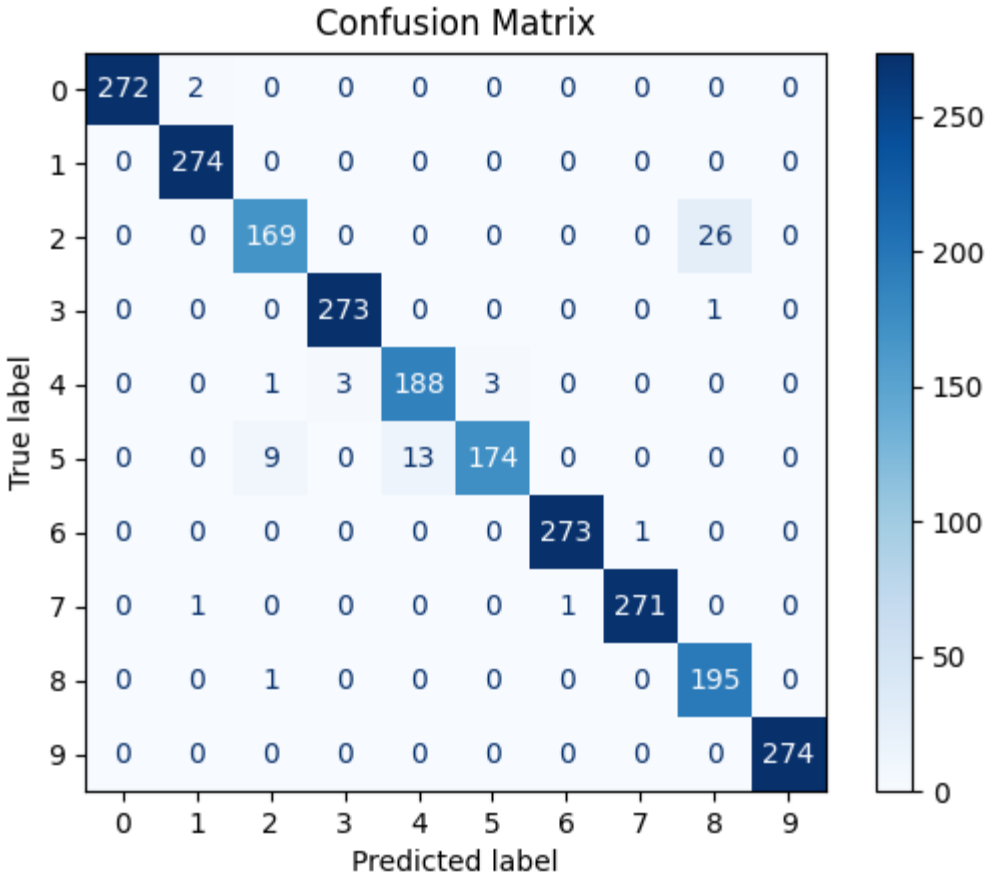


FIGURE 4.4 – Module d'attention spatiale.

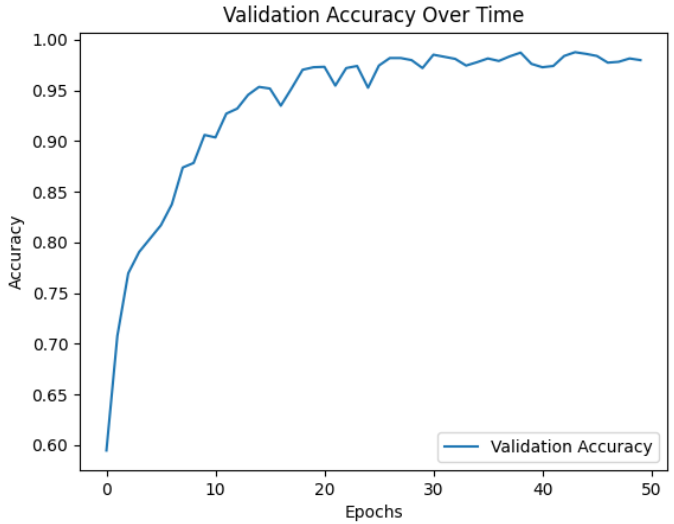
CNN1

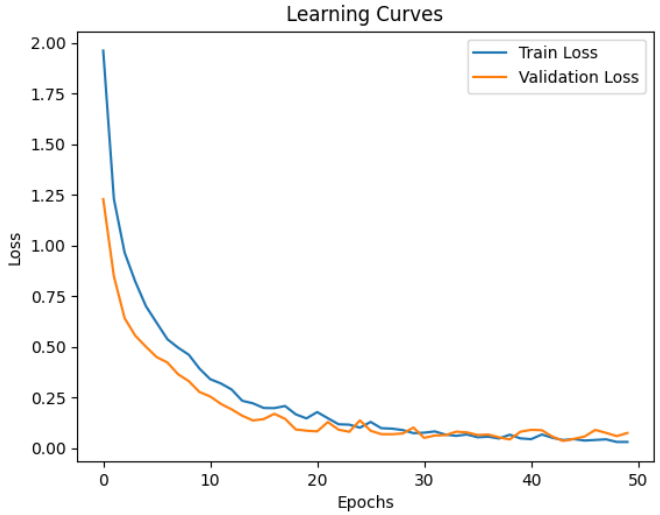

Training and Validation Loss
Training loss
Validation loss
Loss
Epochs

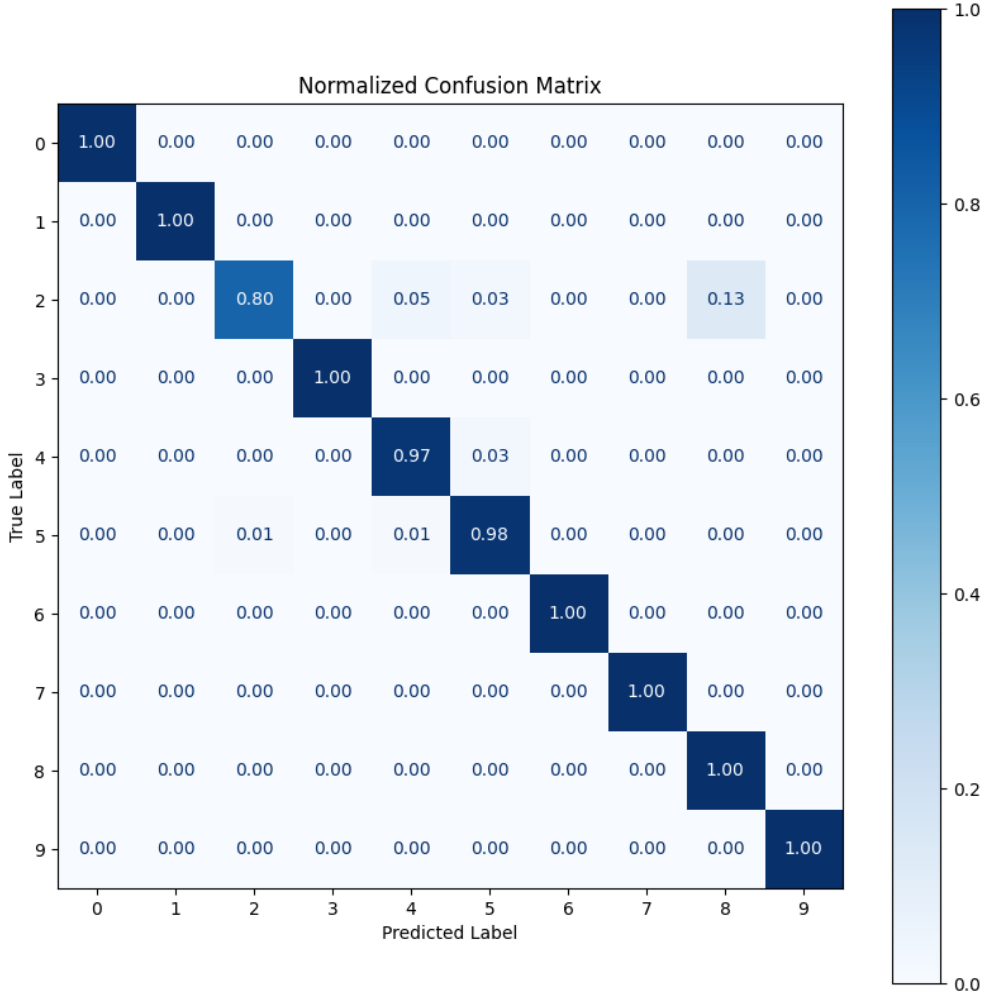

Training and Validation Accuracy
Training accuracy
Validation accuracy
Accuracy
Epochs


Confusion Matrix
True label
Predicted label

CNN2



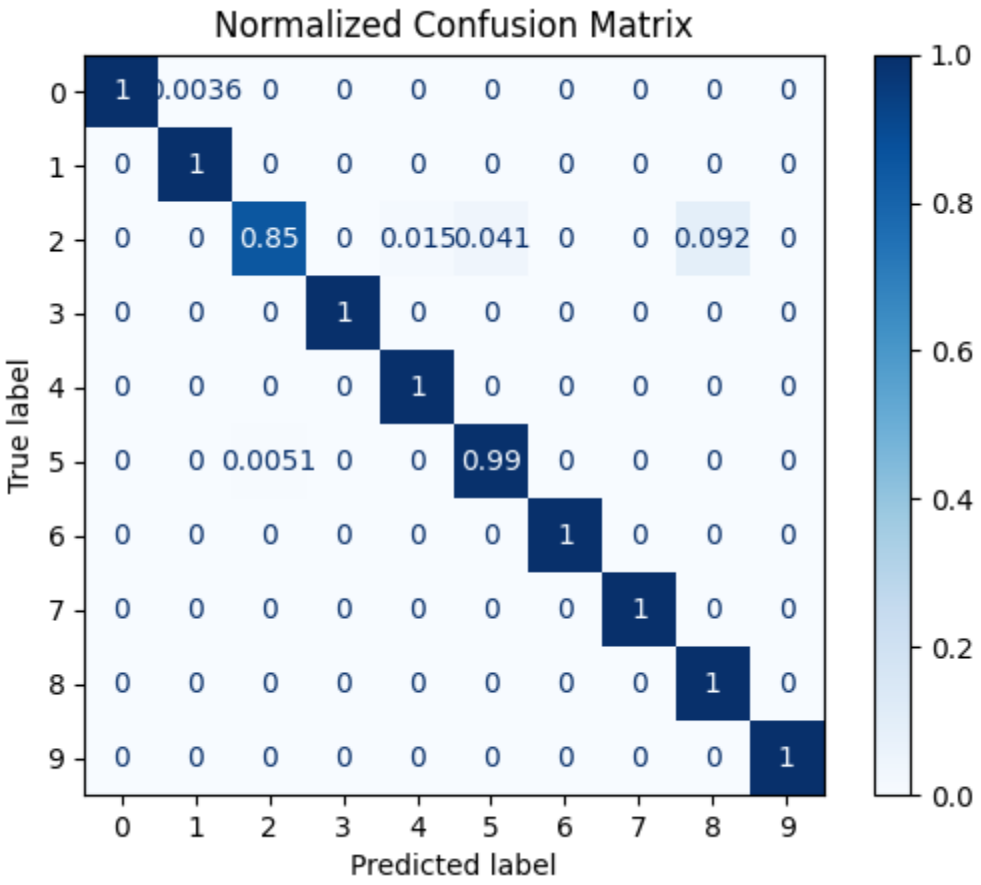


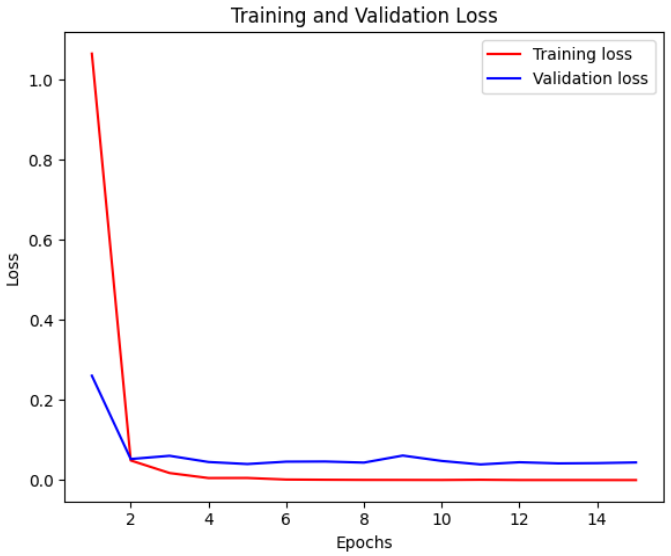


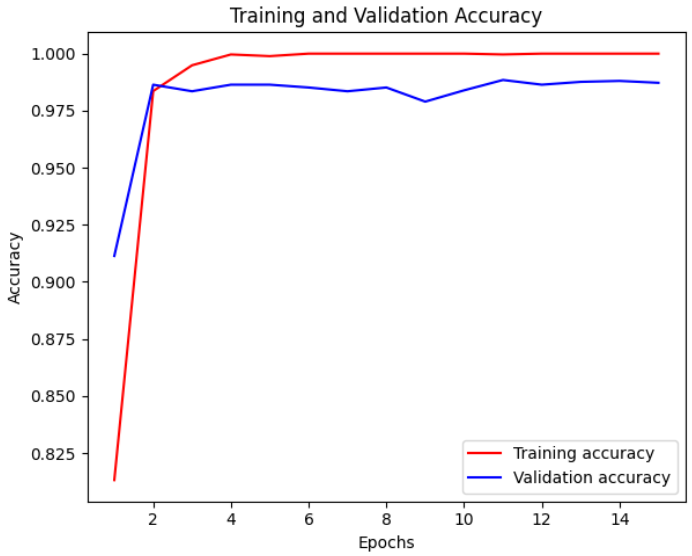
- Validation accuracy : 98.6%

- Learnable parameters :

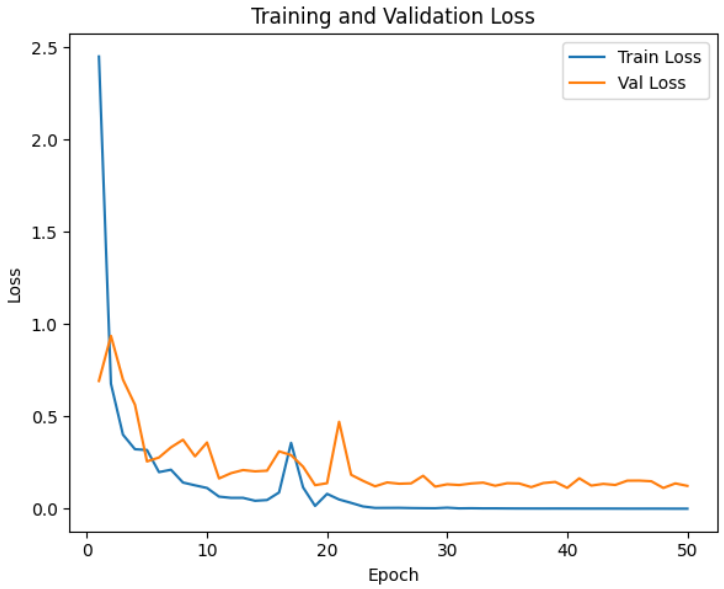
Spatial attention model

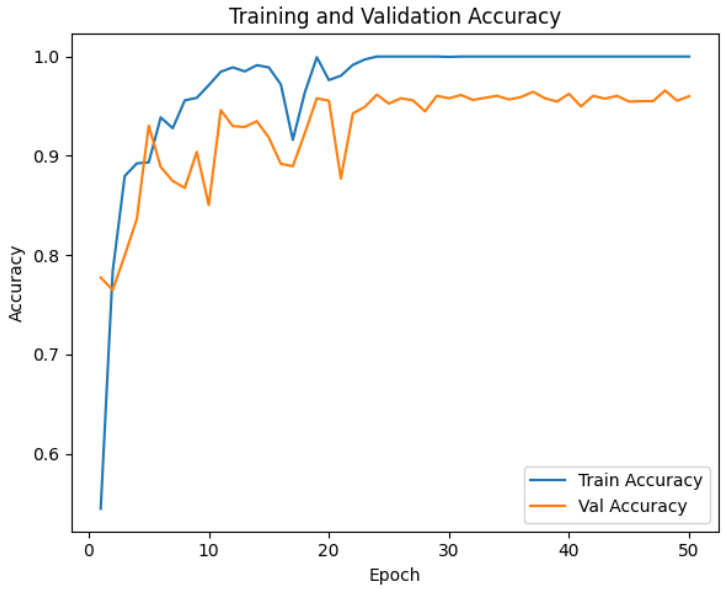


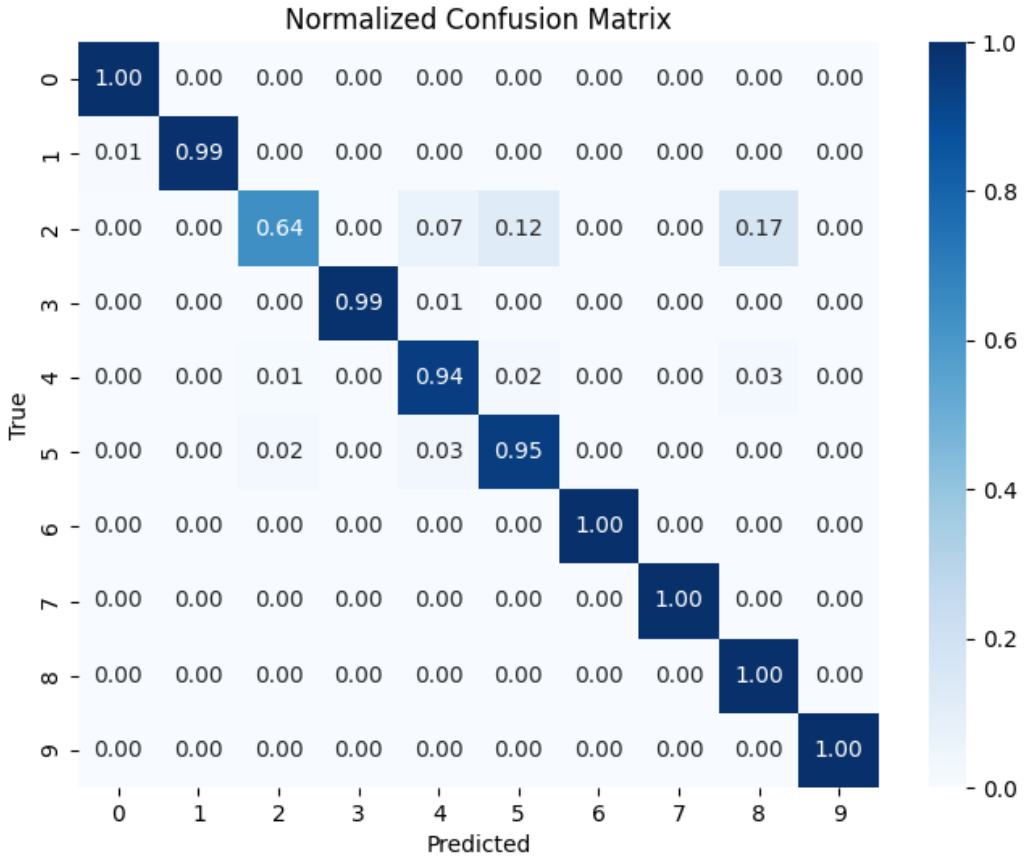




CBAM model



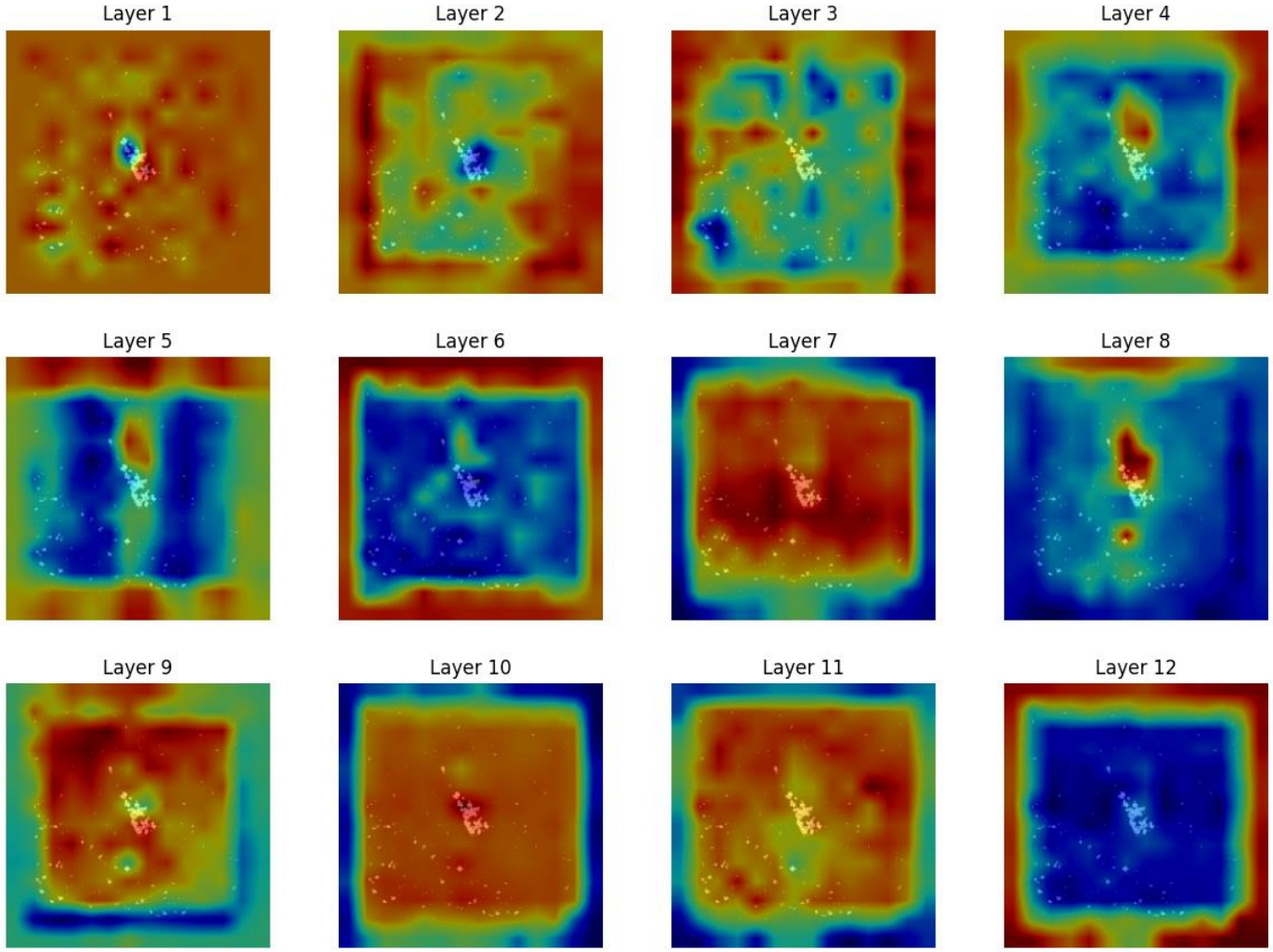




- Validation accuracy : 96.0%

- Learnable parameters

Self-attention



Attention spatiale