Implémentation d'un réseau de neurones sur FPGA

Hugues de Valon
Paul Luperini
Lucas Mahieu

23 janvier 2017

Table des matières

1 Introduction

De nos jours, les réseaux de neurones artificiels reprennent de plus en plus d'importance car la puissance de calculs disponible permet d'obtenir des résultats satisfaisant en temps raisonnable. Le traitement d'images, la reconnaissance vocale ou le traitements lexicaux sont des applications qui pourraient être intégrées dans des systèmes embarqués. Pour ce type d'application, il est possible d'implémenter sur des CPU ou GPU des algorithmes neuronales, mais la consommation et la vitesse de traitement deviendraient vite limitant.

Créer un composant électronique (ASIC ou une IP FPGA dans un premier temps) implémentant un réseau de neurones réduirait drastiquement sa consommation et améliorerait la vitesse du réseau par rapport à un CPU. De plus, étant donnée que l'IP serait spécialisé à cette application permettrait de rendre paramétrable dynamiquement le composant lui permettrait de s'adapter à de multiples applications.

Le projet "Réseau de neurones sur FPGA" s'inscrit dans ce cadre. Sous le tutorat de Frédéric Pétrot et Adrien Prost-Boucle, nous devons créer un tel composant, et tester ses performances en vue de le comparer à des systèmes existants tels que la puce Spinnaker ou TrueNorth d'IBM ou encore de systèmes en développement tel que les réseaux de neurones ternaires du laboratoire TIMA. Une fois implémenté et validé, nous utiliserons une carte FPGA Zedboard pour tester notre composant sur une application classique de reconnaissance de chiffres manuscrits, en utilisant la base de données MNIST.

2 Cahier des charges détaillé

2.1 Présentation générale du problème

2.1.1 Projet

Le but du sujet est de réaliser un réseau de neurones sur FPGA. Un réseau de neurones est un algorithme schématiquement inspiré du fonctionnement des neurones biologiques. On représente un réseau de neurones par un certain nombre de niveau de plusieurs neurones. Un neurone est représenté par un noeud qui reçoit des données de la part des neurones du niveau précédent et diffuse sa valeur de sortie aux neurones du niveau suivant. Les opérations effectuées sur les données par chaque neurone sont assez simples, se sont des multiplications à accumulations : $\sum_{i=0}^{n} w_i * d_i$, avec n données entrantes dans le réseau, w_i le poids pour l'entrée i et d_i la données venant du neurones i du niveau précédant. Le nombre de niveau et le nombre de neurones sont des paramètres qui peuvent être dimensionnés en fonction de l'application voulu. Sur la figure 2 page 5 vous trouverez un schéma représentant un exemple de réseau de neurones à 3 niveaux.

2.1.2 Finalités

Ce réseau de neurones sur FPGA a pour but d'établir une référence pour pouvoir comparer un réseau de neurones ternaire en cours de développement au laboratoire TIMA à un réseau de neurones plus classique, tous deux réalisés sur le même matériel.

Un objectif secondaire serait de comparer le composant réalisé à d'autres produits similaires tels que la puce neuromorphique Spinnaker de l'université de Manchester, ou bien la puce TrueNorth d'IBM.

2.1.3 Contexte

Le projet sera réalisé au CIME Nanotech, à Grenoble.

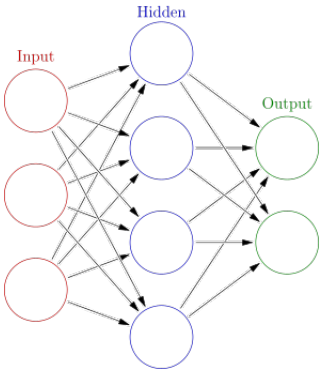

FIGURE 1 – Schéma d'un réseau de neurones à 1 niveau "caché" de 4 neurones avec 3 entrées et 2 sorties

Quatre heures par semaine sont allouées, dans notre emploi du temps, à la réalisation de ce projet. Cependant, il est nécessaire de travailler en dehors des horaires prévus pour terminer le projet.

Les tests sur carte FPGA seront réalisés dans un premier temps sur une carte Zybo, puis une fois la fonctionnalité du composant validée, les tests se poursuivront sur carte ZedBoard.

2.1.4 Énoncé du besoin

Les finalités du produit pour le futur utilisateur tel que prévu par le demandeur sont :

- Il faut concevoir un programme purement logiciel de référence. Celui ci doit respecter les caractéristiques du réseau de neurones demandé, et produire un résultat qui servira de référence à l'IP conçue.
- Le composant matériel doit pouvoir être générique, c'est-à-dire qu'il doit être possible de changer le nombre de neurones, les coefficients de chaque neurones ou encore la taille des données d'entrées, sans la moindre adaptation de code (simplement en changeant les variables voulu dans un fichier).
- Le composant matériel doit produire le résultat attendu, c'est-à-dire qu'il doit produire un résultat identique au programme de référence.
- Le produit étant fortement technique et devant être utilisé par des personnes n'ayant pas conçu cet IP, une documentation détaillée doit être fournie.

2.2 Expression fonctionnelle du besoin

2.2.1 Fonctions de service et de contrainte

Fonctions de service principales

Le produit doit calculer le résultat d'une donnée d'entrée soumise à un réseau de neurones, paramétré selon les coefficients précédemment donnés au composant. Dans le cadre de ce projet, les données d'entrée seront les images issues de la base de données MNIST composée d'image de caractères manuscrits. Ainsi le but premier du réseau sera de trouver quel caractère lui a été donné en entrée.

Fonctions de service complémentaires

Pour une question d'évolution et de réutilisation du projet, le produit rendu doit être reconfigurable très rapidement. En modifiant seulement quelques variables, le produit doit être capable de s'adapter à d'autres applications que celle prévu pour ce projet.

Contraintes

L'architecture globale du composant et ses interfaces (bus, interconnexions, ...) sont déterminées par le laboratoire TIMA.

Le composant devra utiliser les cellules DSP du FPGA pour réaliser la multiplication à accumulation d'un neurone. De plus, il devra utiliser des BRAM pour stocker les coefficients nécessaire aux calculs.

Décomposition en modules, sous-ensembles

L'IP conçue s'intègre dans un environnement complexe. Dans un même chip, se trouvent un FPGA, deux coeurs ARM et de nombreuses autres IP tel qu'un DMA, de la mémoire et un bus permettant de faire communiquer toutes ces IP entre elles. Le réseau de neurones est donc implanté dans le FPGA, et est composé dans l'ordre de propagation des données de :

- Une FIFO permettant de récupérer les données envoyées par le micro-processeur via le bus.
- Un premier niveau de plus de cents neurones.
- Chaque neurone produit un résultat stocké dans une deuxième FIFO.
- Les données produites par le premier niveau de neurones passent par une fonction non linéaire.
- Ces données ainsi traitées sont envoyées par une FIFO au deuxième et dernier étage de neurones comptant 10 neurones (1 par digit à prédire).
- Les résultats finaux sont ainsi renvoyés vers le micro-processeur via le bus AXI du chip.

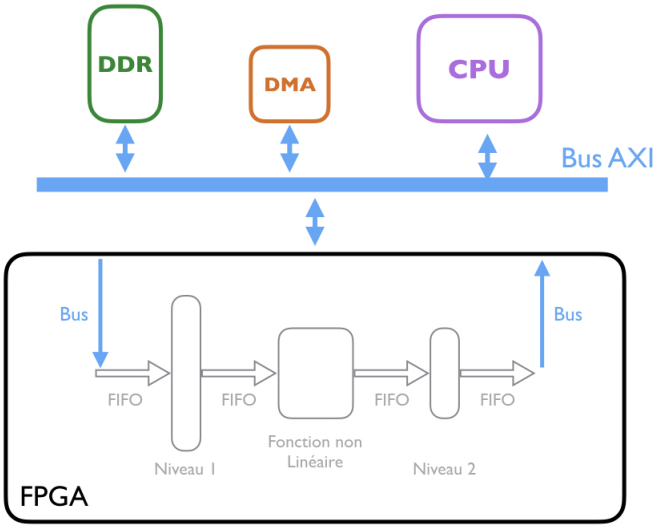

FIGURE 2 – Schéma haut niveau de l'architecture du réseau de neurones et de son environnement dans le FPGA Xilinx

2.2.2 Critères d'appréciation

Les critères permettant de mesurer la qualité du composant produit sont :

- Correction du composant : le résultat doit être celui attendu.
- Taux d'utilisation des cellules du FPGA

— Performances du composant (fréquence, nombre de cycles pour calculer le résultat ...)
— Généricité du composant

2.2.3 Niveaux des critères d'appréciation

Niveaux dont l'obtention est imposée

Il est nécessaire que le composant satisfasse les niveaux de critères suivants :
— Correction : le résultat doit être correct
— Taux d'utilisation des cellules du FPGA : un neurone doit utiliser une cellule DSP.

Niveaux souhaités mais révisables

Il est souhaitable que le composant satisfasse les niveaux de critères suivants :
— Performances : Le résultat d'un calcul du composant doit être plus rapide que son équivalent réalisé sur un processeur classique.
— Généricité : Les fichiers HDL du composant doivent pouvoir être modifié de façon mineure pour changer les paramètres du réseau de neurones.

3 Etapes de conception

A remplir

4 Validation

A remplir

5 Manuel utilisateur

A remplir

6 Planning et répartition des tâches

A remplir

7 Présentation de la démonstration

A remplir

8 Conclusion

A remplir

Appendices

A Potentielle annexe

Références

List of Symbols

DSP Digital Signal Processing : Composant électronique disponible dans le FPGA permettant de faire des multiplication et addition de façon optimisée.

FIFO First In First Out : structure de données permettant de stocker des données et de les restitué dans l'ordre d'arrivée.

FPGA Field-Programmable Gate Array : circuits intégrés re-programmables

IP Intellectual Property : Bloc hardware réutilisable ayant une fonctionnalité spécifique

MNIST Mixed National Institute of Standards and Technology : Base de données de caractère manuscrit servant de tests à de nombreux algorithmes de reconnaissance d'écriture