Deep networks: TensorFlow

Wojciech Działowski

TensorFlow to biblioteka programistyczna typu open source do obliczeń numerycznych z wykorzystaniem wykresów przepływu danych o wysokiej wydajności. Wykorzystywana jest w uczeniu maszynowym i głębokich sieciach neuronowych.

Biblioteka używa wykresy przepływu danych. Węzły grafów reprezentują operacje matematyczne, a krawędzie wykresów reprezentują wielowymiarowe macierze danych (tensory), które przepływają między nimi.

Biblioteka może do działania wykorzystywać zarówno karty graficzne, procesory , jak i wyspecjalizowane mikroprocesory nazywane akceleratorami AI (tensor processing unit).

Elastyczna architektura pozwala na łatwe wdrażanie obliczeń na różnych platformach, a także na komputerach stacjonarnych, klastrach serwerów i urządzeniach przenośnych.

Biblioteka napisana została przez zespół Google Brain do wewnętrznych zastosowań, a następnie została wydana na licencji open source Apache 2.0 9 listopada 2015 r.

TensorFlow zapewnia stabilne API Pythona i C, bez gwarancji kompatybilności wstecznej API takich jak C ++, Go, Java, JavaScript i Swift.

Biblioteka składa się z kilku modułów:

1. W jej najniższej warstwie znajduje się rozproszony silnik wykonawczy, który w celu podniesienia wydajności został zaimplementowany w C++.
2. Nad nią znajdują się interfejsy użytkownika napisane w kilku językach programowania m.in. w Pythonie oraz C++.
3. Powyżej umieszczona została warstwa API, która zapewnia prostszy interfejs dla powszechnie używanych warstw w modelach głębokiego uczenia.
4. Na następną warstwę składają się wysoko poziomowe API, m.in. Keras oraz Estimator API, które ułatwiają tworzenie modeli i ich ocenę.
5. Ponad tym znajdują się przygotowane przez twórców biblioteki gotowe do użycia modele.

Tensor processing unit

Tensor processing unit

W maju 2016 r. Firma Google wydała moduł przetwarzania tensorowego - TPU, układ scalony specyficzny dla aplikacji zbudowany specjalnie do uczenia maszynowego i dostosowany do biblioteki TensorFlow.

TPU to programowalny akcelerator sztucznej inteligencji zaprojektowany w celu zapewnienia wysokiej przepustowości arytmetyki mało precyzyjnej (na przykład 8-bitowej) i ukierunkowany na używanie lub uruchamianie modeli zamiast ich szkolenia.

Tensor processing unit

Jednostki TPU są zastrzeżone i nie są dostępne na rynku, ale Google umożliwia dostęp do tych chipów za pośrednictwem usługi chmury obliczeniowej.

Google użyło czipów TPU między innymi do projektów AlphaGo, AlphaZero, RankBrain i Google Street View

Tensor processing unit

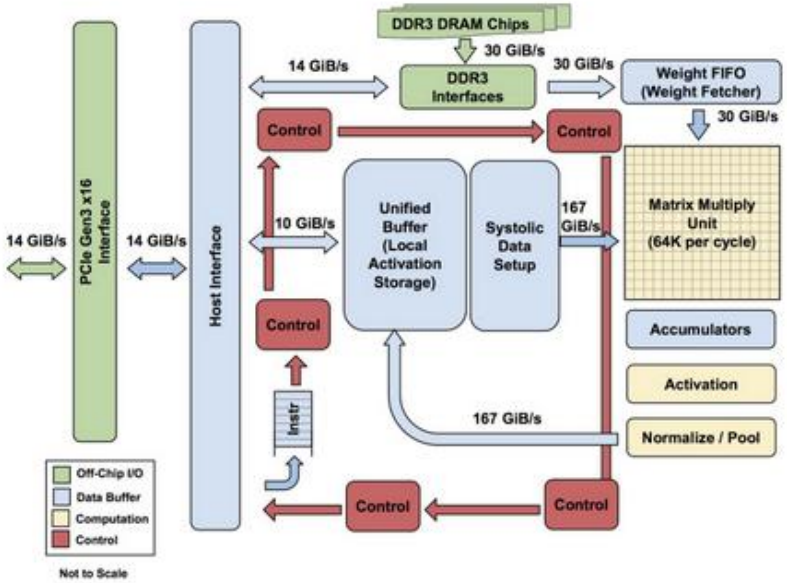

Figure 1. TPU Block Diagram. The main computation part is the yellow Matrix Multiply unit in the upper right hand corner. Its inputs are the blue Weight FIFO and the blue Unified Buffer (UB) and its output is the blue Accumulators (Acc). The yellow Activation Unit performs the nonlinear functions on the Acc, which go to the UB.

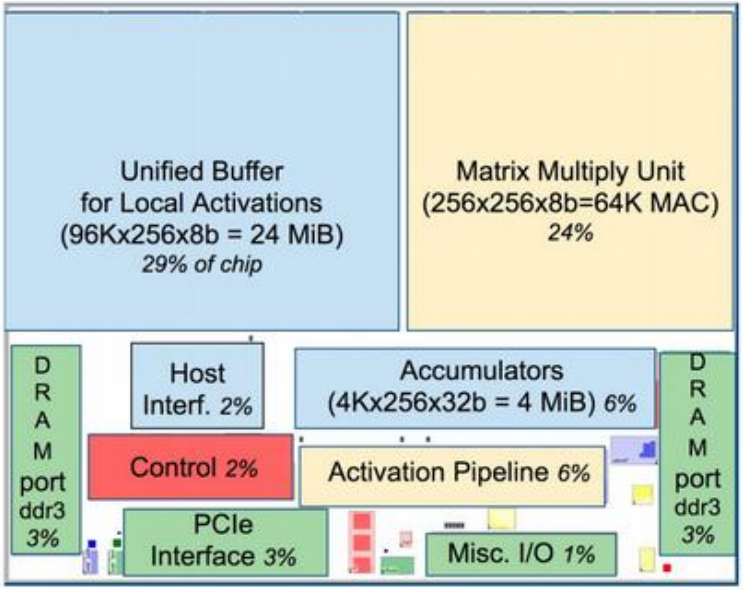

Figure 2. Floor Plan of TPU die. The shading follows Figure 1. The light (blue) data buffers are 37% of the die, the light (yellow) compute is 30%, the medium (green) I/O is 10%, and the dark (red) control is just 2%. Control is much larger (and much more difficult to design) in a CPU or GPU