ĐẠI HỌC BÁCH KHOA HÀ NỘI
KHOA TOÁN - TIN

Báo cáo môn học
Kiến trúc máy tính

Chủ đề
GPU và vai trò của GPU đối với học sâu

Giảng viên hướng dẫn:	ThS. Phạm Huyền Linh	
Mã lớp:	155350	
Nhóm 12:	Vũ Việt Anh	20210031
	Nguyễn Bá Anh	20203309
	Bùi Khánh Duy	20227104
	Phạm Thị Khánh	20227127
	Lê Tiến Thành	20227070
	Phan Thu Trang	20227156

Hà nội, ngày 19 tháng 11 năm 2024

Mục lục

Mục lục 2

Danh sách hình vẽ 4

Giới thiệu 5

1 Giới thiệu GPU 6
- 1.1 GPU là gì? . . . 6
- 1.2 Lịch sử phát triển GPU . . . 6
- 1.3 Kiến trúc GPU . . . 8
- 1.4 Chức năng nổi bật của GPU . . . 10
- 1.5 Phân loại GPU phổ biến hiện nay . . . 11
- 1.6 Ứng dụng GPU trong đời sống . . . 12
- 1.7 Một số thương hiệu GPU được sử dụng phổ biến . . . 14

2 Học sâu 17
- 2.1 Học sâu là gì? . . . 17
- 2.2 GPU cho học sâu . . . 17
  - 2.2.1 Xử lý song song quy mô lớn . . . 17
  - 2.2.2 Tối ưu hóa tính toán cho mạng nơ-ron . . . 17
  - 2.2.3 Tốc độ huấn luyện nhanh hơn . . . 18
  - 2.2.4 Khả năng mở rộng . . . 18
  - 2.2.5 Hỗ trợ từ các thư viện và framework học sâu . . . 18
  - 2.2.6 Hiệu suất với các mô hình lớn . . . 19
  - 2.2.7 Các loại GPU phổ biến cho học sâu . . . 19

3 Vai trò của GPU với học sâu 20
- 3.1 Tăng tốc độ tính toán cho các mô hình học sâu . . . 20
- 3.2 Hỗ trợ các thuật toán xử lý hình ảnh và thị giác máy tính . . . 20
- 3.3 Đáp ứng nhu cầu của các ứng dụng học sâu doanh nghiệp . . . 20
- 3.4 Tiết kiệm năng lượng và tối ưu hóa chi phí . . . 21
- 3.5 Tích hợp trong hệ thống đám mây học sâu . . . 21

4 Thách thức và tương lai **22**
4.1 Thách thức 22
4.2 Tương lai 23

5 Thực nghiệm và kết quả **26**
5.1 Tổng quan 26
5.2 Thực nghiệm và kết quả 27
5.2.1 Thông số thiết bị 27
5.2.2 Thực nghiệm và kết quả với bộ dữ liệu nhỏ 27
5.2.3 Thực nghiệm và kết quả với bộ dữ liệu trung bình 31

Tài liệu **35**

Lời cảm ơn **37**

Danh sách hình vẽ

1 GPU . . . 6
2 Câu tao GPU . . . 10
3 GPU tích hợp . . . 11
4 GPU rời . . . 12
5 Ứng dụng trong game . . . 12
6 Ứng dụng trong đồ họa hình ảnh và video . . . 13
7 Ứng dụng trong y khoa . . . 13
8 GPU NVIDIA . . . 14
9 GPU AMD . . . 14
10 GPU Intel . . . 15
11 GPU Adreno Series . . . 15
12 GPU Mali . . . 16
13 NVDIA . . . 19
14 AMD . . . 19
15 Thông số GPU khi huấn luyện . . . 28
16 Biểu đồ thể hiện độ chính xác khi học (GPU) . . . 28
17 Mức sử dụng bộ nhớ (CPU) . . . 29
18 Mức độ tiêu hao năng lượng (CPU) . . . 29
19 Nhiệt độ (CPU) . . . 30
20 Độ chính xác mô hình (CPU) . . . 30
21 Kết quả huấn luyện (GPU) . . . 32
22 Độ chính xác của mô hình (GPU) . . . 32
23 Mức tiêu hao bộ nhớ (CPU) . . . 33
24 Nhiệt độ (CPU) . . . 33
25 Độ chính xác của mô hình (CPU) . . . 34

Giới thiệu

Mục đích, mục tiêu và phạm vi đề tài

Trong thời tại công nghệ dữ liệu lớn, trí tuệ nhân tạo phát triển nhanh chóng, học sâu (Deep Learning) đã ra đời như một bước ngoặt lớn, nó cho phép xây dựng nhiều mô hình xử lý, nhận dạng hình ảnh, ngôn ngữ tự nhiên, data... có độ chính xác cao. Từ đó góp vai trò quan trọng trong lĩnh vực khai thác, xử lý và phân tích dữ liệu.

Học sâu là một nhánh nhỏ của trí tuệ nhân tạo, được phát triển dựa trên các mô hình mạng nơ-ron nhân tạo (Neural Networks) để xử lý, phân tích dữ liệu và mô phỏng bộ não con người. Các ứng dụng của Deep Learning đã chứng tỏ được hiệu quả vượt trội trong nhiều lĩnh vực như nhận dạng hình ảnh, xử lý ngôn ngữ tự nhiên, tự động lái xe, phân tích cảm xúc, trợ lý ảo - virtual assistant, phân tích dữ liệu lớn, các mô hình trong lĩnh vực y tế như mô hình dự đoán bệnh, chẩn đoán ung thư, phân tích kết quả chụp MRI, X-quang và nhiều ứng dụng khác.

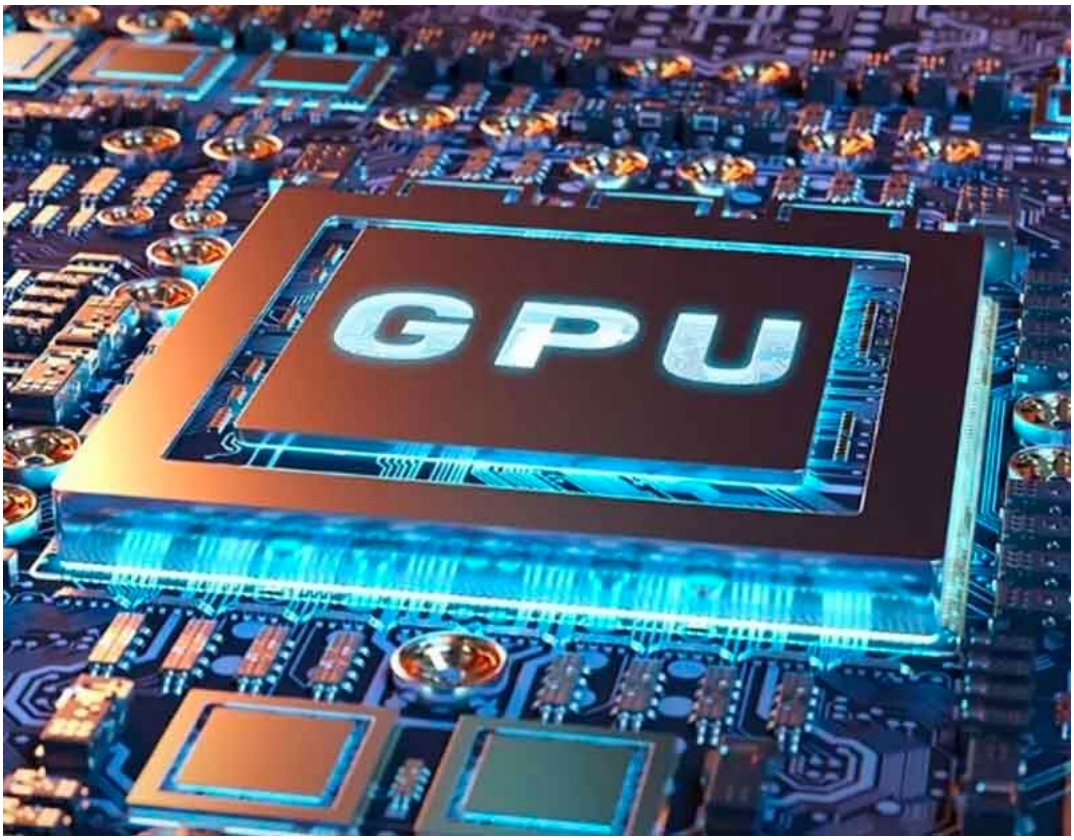
Tuy nhiên, một trong những thách thức lớn nhất trong việc triển khai các mô hình học sâu là yêu cầu tính toán cực kỳ cao đối với các phép toán ma trận, vectơ và hàm số phức tạp. Đây là một khó khăn lớn với các bộ xử lý truyền thống - CPU, vốn chỉ có thể xử một lượng giới hạn các tác vụ song song. Trong khi đó GPU (Graphics Processing Unit), vốn được thiết kế để xử lý đồ họa trong các trò chơi và ứng dụng đa phương tiện đã giúp tăng tốc đáng kể quá trình huấn luyện mô hình học sâu thông qua khả năng tính toán vượt trội các bài toán song song. Điều mà CPU truyền thống không thể đạt được.

Thông qua bài báo cáo này, mục tiêu của nhóm em là mang đến cái nhìn sâu sắc về vai trò của GPU đối với học sâu, đồng thời khám phá và đánh giá sự khác biệt giữa việc sử dụng CPU và GPU trong việc huấn luyện các mô hình học sâu. Nhóm em hy vọng có thể chứng minh lợi thế rõ rệt của GPU trong học sâu qua đó hiểu rõ hơn về cách tối ưu hóa hiệu quả huấn luyện mô hình học sâu bằng phần cứng phù hợp.

Phần 1. Giới thiệu GPU

1.1. GPU là gì?

GPU (Graphics Processing Unit) là bộ xử lý đồ họa, được thiết kế đặc biệt để xử lý các tác vụ liên quan đến đồ họa và hình ảnh. Ban đầu, GPU chủ yếu được sử dụng để tăng cường khả năng xử lý đồ họa trong các ứng dụng như trò chơi điện tử, dựng hình ảnh 3D, và phát lại video. GPU ngày càng được sử dụng rộng rãi trong các ứng dụng tính toán khác, đặc biệt là trí tuệ nhân tạo (AI) và học sâu (Deep Learning), do khả năng xử lý song song mạnh mẽ. GPU có thể thực hiện nhiều tác vụ cùng lúc, điều này làm cho chúng cực kỳ hiệu quả trong các lĩnh vực yêu cầu tính toán khối lượng lớn và tốc độ cao.



Hình 1: GPU

1.2. Lịch sử phát triển GPU

Lịch sử phát triển của GPU là một quá trình dài với nhiều giai đoạn khác nhau, từ sự ra đời của các bộ xử lý đồ họa cơ bản cho đến những siêu máy tính hiện đại chuyên dụng

cho AI và học máy.

- Thập niên 1970 - 1980: Khởi đầu
  - **1970s - 1980s**: GPU ban đầu không thực sự tồn tại như một phần cứng riêng biệt. Các máy tính và hệ thống trò chơi điện tử sử dụng các bộ vi xử lý đơn giản để điều khiển đầu ra đồ họa. Các hệ thống như Atari 2600 (1977) đã có các chip xử lý đơn giản để hiển thị hình ảnh trên màn hình.
  - **IBM PC và CGA (1981)**: Một trong những hệ thống máy tính đầu tiên có khả năng hiển thị đồ họa cơ bản là IBM PC với Card đồ họa CGA (Color Graphics Adapter), hỗ trợ hiển thị 4 bit - 16 màu.
- Thập niên 1990: Sự ra đời của GPU chuyên dụng
  - **1990s**: Sự bùng nổ của các trò chơi điện tử và đồ họa máy tính trong thập niên 1990 đã dẫn đến nhu cầu phát triển các bộ xử lý chuyên dụng cho đồ họa. Các hãng sản xuất như 3dfx, NVIDIA, và ATI (sau này trở thành một phần của AMD) bắt đầu phát triển các card đồ họa có GPU riêng biệt.
  - **3dfx Voodoo (1996)**: Đây là một trong những card đồ họa đầu tiên thực sự tạo ra một cuộc cách mạng trong ngành công nghiệp, mang lại khả năng hiển thị đồ họa 3D mượt mà và phong phú cho các trò chơi máy tính.
  - **NVIDIA RIVA 128 (1997)**: Một trong những sản phẩm đột phá của NVIDIA, được xem là cột mốc quan trọng trong sự phát triển của GPU, với khả năng xử lý 2D và 3D, đánh dấu sự khởi đầu của kỷ nguyên GPU hiện đại.
- Thập niên 2000: GPU hiện đại và sự phát triển của NVIDIA
  - **NVIDIA GeForce 256 (1999)**: Được gọi là GPU đầu tiên trên thế giới, GeForce 256 tích hợp khả năng xử lý đồ họa 3D, bao gồm cả ánh sáng, bóng đổ và xử lý hình học. Đây là sản phẩm giúp NVIDIA trở thành một trong những nhà sản xuất GPU hàng đầu thế giới.
  - **2006 - NVIDIA CUDA**: NVIDIA giới thiệu kiến trúc CUDA, cho phép các nhà phát triển sử dụng GPU không chỉ để xử lý đồ họa mà còn để thực hiện các tác vụ tính toán khác (tính toán song song). Điều này đã mở ra cơ hội cho GPU tham gia vào các lĩnh vực mới như khoa học dữ liệu, học máy và AI.
- Thập niên 2010: GPU và Trí tuệ nhân tạo
  - **2010s**: GPU dần trở thành trung tâm của các ứng dụng không chỉ trong đồ họa mà còn trong trí tuệ nhân tạo (AI) và học sâu (Deep Learning). Do GPU có khả

năng xử lý song song cực kỳ mạnh mẽ, chúng được sử dụng để tăng tốc quá trình huấn luyện các mô hình học máy, đặc biệt là các mạng nơ-ron sâu.

  - **NVIDIA Tesla (2007 - hiện tại)**: NVIDIA đã phát triển dòng GPU Tesla, được thiết kế đặc biệt cho siêu máy tính và trí tuệ nhân tạo. Tesla GPU trở thành tiêu chuẩn trong các trung tâm dữ liệu AI.

- Hiện tại và tương lai:

  - **2020s**: GPU hiện đại như NVIDIA Ampere, NVIDIA RTX và AMD Radeon không chỉ tập trung vào hiệu năng đồ họa cao cấp cho trò chơi điện tử và thiết kế đồ họa mà còn đóng vai trò quan trọng trong việc phát triển AI, học máy và các ứng dụng tính toán khoa học.
  - **AI & Học Máy**: GPU hiện là thành phần chủ chốt trong việc xây dựng các hệ thống trí tuệ nhân tạo hiện đại. Các lĩnh vực như thị giác máy tính, nhận dạng giọng nói, và xử lý ngôn ngữ tự nhiên đều dựa vào sức mạnh của GPU.
  - **GPU Quantum (tương lai)**: Với sự phát triển của máy tính lượng tử, tương lai có thể chứng kiến sự kết hợp giữa GPU truyền thống và các công nghệ mới như GPU lượng tử để mở rộng khả năng tính toán đến những giới hạn chưa từng có.

1.3. Kiến trúc GPU

Kiến trúc của GPU được thiết kế đặc biệt để xử lý các tác vụ đồ họa và tính toán song song một cách hiệu quả. GPU có hàng ngàn lõi xử lý nhỏ giúp xử lý nhiều tác vụ đồng thời, rất hữu ích trong các ứng dụng đòi hỏi tính toán song song như đồ họa 3D, AI, và học sâu. Dưới đây là các thành phần chính trong kiến trúc GPU:

- Lõi xử lý song song (CUDA Cores/Stream Processors)

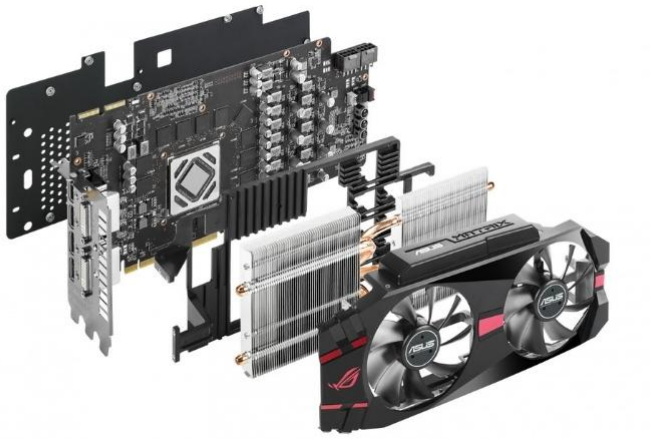
  - Lõi CUDA (Compute Unified Device Architecture) của NVIDIA hay Stream Processors của AMD là các đơn vị tính toán cơ bản của GPU. Một GPU có thể có hàng ngàn lõi xử lý như vậy, cho phép thực hiện hàng nghìn phép tính đồng thời.
  - Các lõi này được thiết kế để xử lý các tác vụ đơn giản nhưng khối lượng lớn, ví dụ như tính toán cho từng điểm ảnh (pixel) trong một hình ảnh hoặc từng phần tử trong một tập dữ liệu lớn.

- Đơn vị xử lý hình học (Geometry Processing Units): GPU được trang bị các đơn vị xử lý hình học chuyên biệt để xử lý các tác vụ liên quan đến hình học 3D như dựng hình

tam giác, xử lý đỉnh (vertices), và tạo lưới đa giác. Điều này rất quan trọng trong đồ họa máy tính 3D khi phải hiển thị các mô hình phức tạp.

- Bộ đổ bóng (Shader Units)
  - Pixel Shader: Xử lý các tác vụ liên quan đến từng pixel trong hình ảnh, bao gồm cả tô màu, độ sáng, và các hiệu ứng đổ bóng.
  - Vertex Shader: Tính toán vị trí, màu sắc và ánh sáng của các điểm (đỉnh) trong mô hình 3D.
  - Geometry Shader: Tạo và biến đổi hình học từ các đỉnh, ví dụ như tạo thêm các tam giác hoặc đường cong.
- Bộ nhớ GPU (GPU Memory): VRAM (Video RAM): Đây là loại bộ nhớ đặc biệt được sử dụng để lưu trữ dữ liệu đồ họa, bao gồm kết cấu (textures), khung hình (frame buffer), và các thông tin khác cần thiết cho quá trình dựng hình. VRAM có băng thông rất cao, giúp tăng tốc quá trình truyền dữ liệu giữa bộ nhớ và các lõi xử lý.
- Đơn vị xử lý văn bản (Texture Mapping Units - TMUs): TMUs chịu trách nhiệm xử lý các kết cấu (textures) và áp dụng chúng lên các bề mặt của mô hình 3D. Chúng xử lý việc ánh xạ, lọc, và các hiệu ứng liên quan đến kết cấu để tạo ra hình ảnh mượt mà, chân thực.
- Rasterizers (Bộ chuyển đổi điểm thành pixel): Rasterizer chuyển đổi các hình dạng và đa giác từ không gian 3D sang không gian 2D (pixel) để hiển thị trên màn hình. Đây là quá trình quan trọng trong việc hiển thị hình ảnh 3D trên các thiết bị hiển thị 2D như màn hình.
- Bộ điều phối tác vụ (Task Scheduler): GPU có một bộ điều phối tác vụ (scheduler) để quản lý việc phân phối các tác vụ tính toán đến các lõi xử lý song song. Điều này đảm bảo rằng GPU có thể tận dụng tối đa các tài nguyên xử lý của mình.
- Bộ xử lý video (Video Decode/Encode Engines): Nhiều GPU hiện đại được trang bị các bộ xử lý video chuyên dụng để mã hóa và giải mã các định dạng video, chẳng hạn như H.264, H.265, VP9. Điều này giúp giảm tải cho CPU khi thực hiện các tác vụ liên quan đến video.
- Bộ điều khiển hiển thị (Display Controller): Bộ điều khiển này chịu trách nhiệm xuất hình ảnh đã xử lý ra các thiết bị hiển thị như màn hình máy tính, TV, hoặc máy chiếu. Nó hỗ trợ nhiều chuẩn kết nối khác nhau như HDMI, DisplayPort, và VGA.

- Kiến trúc đa luồng (SIMD - Single Instruction Multiple Data): GPU sử dụng kiến trúc SIMD hoặc SIMT (Single Instruction, Multiple Threads) để thực thi cùng một lệnh trên nhiều dữ liệu song song. Điều này cho phép GPU đạt được hiệu suất cao trong các tác vụ tính toán song song.



Hình 2: Câu tao GPU

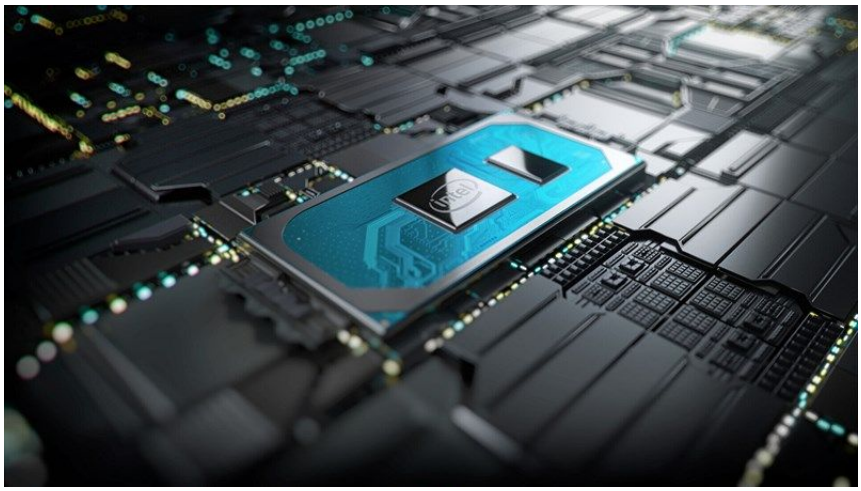
1.4. Chức năng nổi bật của GPU

- GPU có hàng ngàn lõi xử lý có thể thực hiện hàng triệu phép tính toán đồng thời. Điều này làm cho GPU rất hiệu quả trong việc xử lý các tác vụ song song như render video, mô phỏng khoa học, và học máy.
- GPU giúp giảm tải lượng công việc lớn cho CPU. Từ đó giúp tiết kiệm thời gian tối ưu để tạo ra những sản phẩm chất lượng tốt nhất.
- GPU giúp đem lại hình ảnh đồ họa vô cùng sắc nét, khả năng nâng cao quá trình xử lý video và hình ảnh cực đỉnh. Trong đó có các phần mềm được hỗ trợ từ GPU như: After Effects, Adobe Premiere, Camtasia,...
- GPU là công cụ hữu ích trong việc vận hành các tựa game 3D cực mượt mà. Đồng thời nó còn giúp các phần mềm kiến trúc hoạt động ổn định hơn.
- GPU có khả năng chia ra những lõi con khác nhau để xử lý hình ảnh trong vùng tam giác. Kể cả những mặt phẳng phức tạp của vật thể, GPU cũng dễ dàng xử lý một cách nhanh gọn. Đây là ưu điểm vượt trội mà CPU không thể thực hiện được.

1.5. Phân loại GPU phổ biến hiện nay

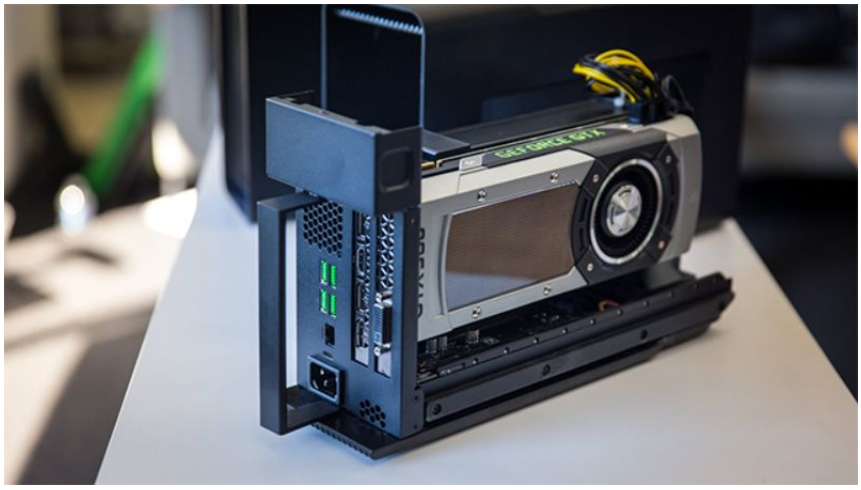
Hiện nay, GPU được chia làm 2 loại chính là: GPU rời và GPU tích hợp. Mỗi một loại GPU sẽ có những nhiệm vụ và ưu điểm khác nhau. Cụ thể:

- GPU tích hợp:
  - Hiện nay, GPU tích hợp chiếm phần lớn và được ưa chuộng bởi nhiều người dùng. GPU tích hợp được nhúng vào CPU để làm nhiệm vụ thay vì đi kèm với card màn hình riêng. Việc kết hợp với CPU cho khả năng giảm thiểu tiêu hao năng lượng và chi phí tối ưu. Đồng thời nhờ đó mà hệ thống cũng vận hành một cách đơn giản, trơn tru hơn rất nhiều.
  - GPU tích hợp còn kết hợp với RAM cho sức mạnh xử lý càng nhân đôi mạnh mẽ hơn nữa. Từ đó có thể xử lý các vấn đề về hình ảnh, tốc độ phân giải một cách nhanh chóng. Kể cả những ứng dụng liên quan đến lĩnh vực thiết kế đồ họa cũng trở nên dễ dàng hơn.



Hình 3: GPU tích hợp

- GPU rời:
  - GPU rời là con chip hoạt động riêng biệt được gắn trong khe cắm PCI Express. GPU rời phù hợp để xử lý các công việc với những phần mềm có nguồn tài nguyên lớn.
  - GPU rời có thể tăng tính năng xử lý với lượng tiêu hao năng lượng và tạo nhiệt. Thông thường, GPU rời sẽ yêu cầu làm mát để tăng hiệu quả làm việc cao hơn.



Hình 4: GPU rời

1.6. Ứng dụng GPU trong đời sống

- Ứng dụng GPU trong game: Trong các tựa game, GPU được ứng dụng vô cùng rộng rãi. Nhờ có GPU mà đồ họa trong game trở nên mượt mà, hình ảnh sống động hơn. Trên thực tế, những tựa game có hiệu năng lớn đều cần có sự hỗ trợ của GPU. Đặc biệt là các dòng game như: LoL, PUBG hay Call of Duty.

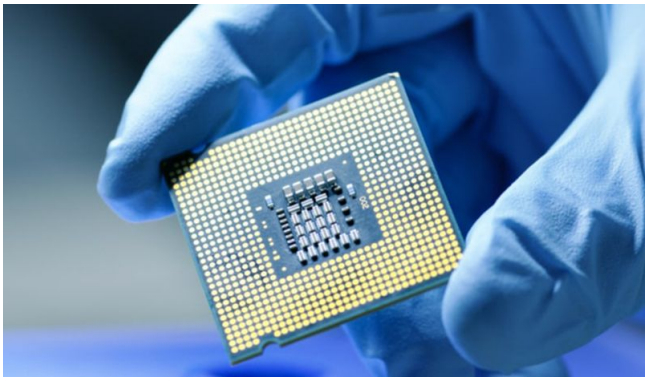
Hình 5: Ứng dụng trong game

- Ứng dụng GPU trong đồ họa hình ảnh và video: GPU có tính ứng dụng cao hơn cả đối với việc xây dựng các hình ảnh và video. Đây là một công việc vô cùng quan trọng của những ai làm kỹ sư thiết kế đồ họa. Đối với việc xây dựng video chất lượng cao, GPU có nhiệm vụ tiếp nhận và xử lý thông tin. Đặc biệt trong quá trình làm video

chất lượng 2K hay 4K thì GPU còn quan trọng hơn cả. Quá trình xử lý và nâng cao hiệu ứng của video sẽ được nhanh nhạy mà không bị giật lag.

Hình 6: Ứng dụng trong đồ họa hình ảnh và video

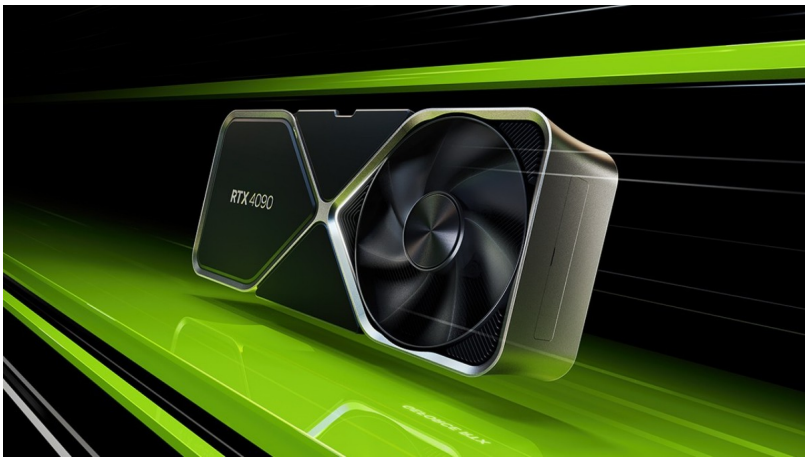
- Ứng dụng của GPU trong khoa học, y khoa: GPU được xem là công cụ hữu ích trong nhiều lĩnh vực khác nhau của đời sống. Cụ thể hơn GPU có thể ứng dụng trong lĩnh vực điện tử, nghiên cứu khoa học, y khoa, thăm dò dầu khí, mô hình tài chính... Có thể nói, GPU chính là một sản phẩm tuyệt vời đáng mong đợi nhất. Việc tận dụng GPU để tạo ra nhiều kỹ thuật tiên tiến mới, làm việc thay thế cho con người.



Hình 7: Ứng dụng trong y khoa

1.7. Một số thương hiệu GPU được sử dụng phổ biến

- **NVIDIA**: Nổi tiếng với dòng card đồ họa GeForce, được đánh giá cao về hiệu năng mạnh mẽ, khả năng xử lý đồ họa ấn tượng và hỗ trợ tốt cho các ứng dụng sáng tạo, chơi game. Một số dòng sản phẩm tiêu biểu bao gồm GeForce RTX 30 Series, GeForce GTX 16 Series, v.v.



Hình 8: GPU NVIDIA

- **AMD**: Đối thủ cạnh tranh trực tiếp của NVIDIA với dòng card đồ họa Radeon RX. AMD tập trung vào việc cung cấp các sản phẩm có hiệu năng cao, giá thành hợp lý, phù hợp cho nhiều đối tượng người dùng. Các dòng sản phẩm nổi bật bao gồm Radeon RX 6000 Series, Radeon RX 7000 Series, v.v.

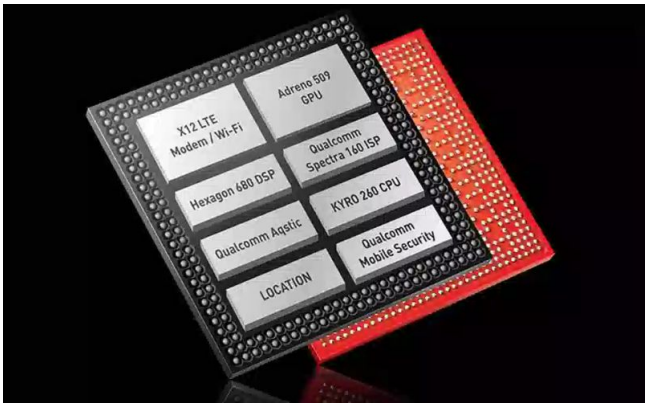
Hình 9: GPU AMD

- **Intel**: Intel, gã khổng lồ trong ngành chip máy tính, đang tạo nên tiếng vang trong thị trường GPU với những dòng sản phẩm đa dạng. Đơn cử như dòng Intel Iris Xe

Graphics mang đến hiệu năng đồ họa mạnh mẽ cho các tác vụ văn phòng, giải trí đa phương tiện và chơi game nhẹ nhàng. Ngoài ra, Intel còn cung cấp các dòng sản phẩm GPU rời cao cấp hơn như Intel Arc Alchemist và sắp tới là Intel Arc Battlemage, nhắm đến đối tượng game thủ và người sáng tạo nội dung chuyên nghiệp vốn đòi hỏi hiệu năng đồ họa mạnh mẽ.

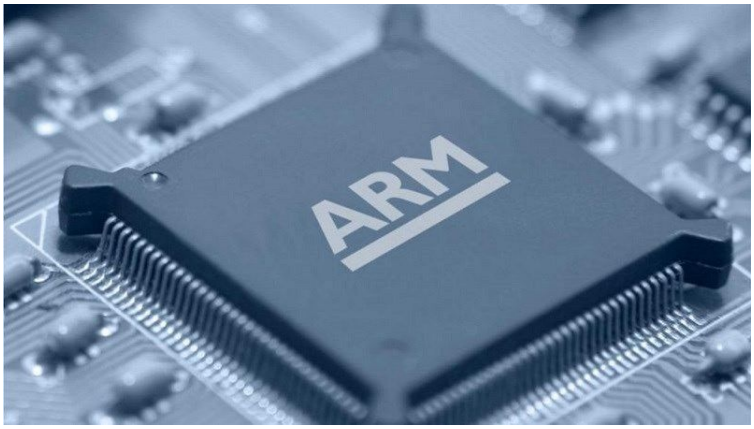
Hình 10: GPU Intel

- **GPU Adreno Series**: GPU Adreno thuộc sở hữu bởi công ty bán dẫn Qualcomm đến từ Mỹ. GPU được sử dụng chủ yếu trong các thiết bị thuộc đơn vị xử lý Snapdragon. Trước đây, Adreno còn có tên gọi cũ là Imageon. Đây là cái tên được sử dụng khi nó được thiết kế bởi thương hiệu ATI Technologies. Sau này đã được thương hiệu AMD mua lại rồi bán cho Qualcomm.



Hình 11: GPU Adreno Series

- **GPU Mali**: GPU Mali thuộc sở hữu bởi thương hiệu thiết kế vi xử lý có tên Advanced RISC đến từ Anh. Tập đoàn này được ra đời vào năm 1990 và nổi tiếng trong giới công nghệ. Các sản phẩm nổi tiếng được tích hợp GPU Mali như: MediaTek, Exynos...



Hình 12: GPU Mali

Ưu điểm nổi bật của chúng đó là tiêu hao điện năng ít giúp giảm thiểu chi phí.

Phần 2. Học sâu

2.1. Học sâu là gì?

Học sâu (deep learning) là một tập hợp con của học máy (machine learning), tập trung vào việc xây dựng và huấn luyện mạng nơ-ron nhiều lớp, được gọi là mạng nơ-ron sâu (DNN – Deep neural networks) để chúng có thể tự động học, hiểu dữ liệu, mô phỏng khả năng ra quyết định phức tạp của bộ não con người.

Mô hình học sâu có thể nhận diện nhiều hình mẫu phức tạp trong hình ảnh, văn bản, âm thanh và các dữ liệu khác để tạo ra thông tin chuyên sâu và dự đoán chính xác. Bạn có thể sử dụng các phương pháp học sâu để tự động hóa các tác vụ thường đòi hỏi trí tuệ con người, chẳng hạn như phân loại hình ảnh hoặc chép lời một tập tin âm thanh.

2.2. GPU cho học sâu

GPU cho học sâu là một bộ xử lý đồ họa được sử dụng để tăng tốc quá trình huấn luyện và triển khai các mô hình trí tuệ nhân tạo (AI) và học sâu (deep learning). Ban đầu, GPU được thiết kế để xử lý đồ họa và hình ảnh trong các ứng dụng như trò chơi điện tử và xử lý video, nhưng nhờ khả năng tính toán song song cao, nó đã trở thành một công cụ quan trọng trong học sâu.

2.2.1. Xử lý song song quy mô lớn

- GPU có hàng nghìn lõi xử lý hoạt động song song, giúp thực hiện các phép toán ma trận và tính toán tensor (rất phổ biến trong học sâu) nhanh chóng hơn so với CPU (Central Processing Unit).
- Điều này đặc biệt quan trọng khi làm việc với các mô hình lớn và khối lượng dữ liệu khổng lồ, vì các phép toán được thực hiện đồng thời trên các phần khác nhau của dữ liệu.

2.2.2. Tối ưu hóa tính toán cho mạng nơ-ron

Trong học sâu, các mô hình như mạng nơ-ron tích chập (Convolutional Neural Networks - CNNs), mạng nơ-ron hồi quy (Recurrent Neural Networks - RNNs), và Transformer yêu

cầu tính toán nhiều phép toán trên ma trận. GPU tối ưu hóa cho các loại phép tính này, giúp giảm thời gian huấn luyện mô hình từ hàng tuần xuống còn vài giờ hoặc thậm chí vài phút.

2.2.3. Tốc độ huấn luyện nhanh hơn

- **Vì sao GPU nhanh hơn CPU trong học sâu?**
  - Kiến trúc tính toán song song: GPU có hàng nghìn lõi xử lý nhỏ, tối ưu cho tính toán đồng thời, trong khi CPU chỉ có vài chục lõi, phù hợp với xử lý tuần tự.
  - Băng thông bộ nhớ cao: GPU có băng thông bộ nhớ lớn hơn nhiều (lên tới 3TB/s), giúp truyền tải dữ liệu nhanh hơn so với CPU.
  - Tối ưu cho công việc học sâu: GPU được thiết kế để xử lý ma trận và tensor hiệu quả hơn, phù hợp với nhu cầu của mạng nơ-ron sâu.
  - Hỗ trợ phần mềm chuyên biệt: Thư viện như CUDA, cuDNN giúp tối ưu hiệu suất GPU trong học sâu, vượt xa CPU.
  - Xử lý dữ liệu hàng loạt: GPU xử lý hàng nghìn dữ liệu cùng lúc, trong khi CPU bị giới hạn bởi khả năng xử lý từng bước.
- Với sự hỗ trợ của GPU, các nhà nghiên cứu và kỹ sư có thể thử nghiệm nhiều cấu trúc mô hình khác nhau trong thời gian ngắn, giúp cải thiện hiệu suất và độ chính xác của mô hình.

2.2.4. Khả năng mở rộng

Nhiều GPU có thể được sử dụng cùng lúc để xử lý các tác vụ lớn hơn, cho phép huấn luyện các mô hình cực lớn với dữ liệu khổng lồ trên nhiều máy chủ (distributed training). Điều này giúp giải quyết các bài toán phức tạp mà CPU không thể xử lý kịp thời.

2.2.5. Hỗ trợ từ các thư viện và framework học sâu

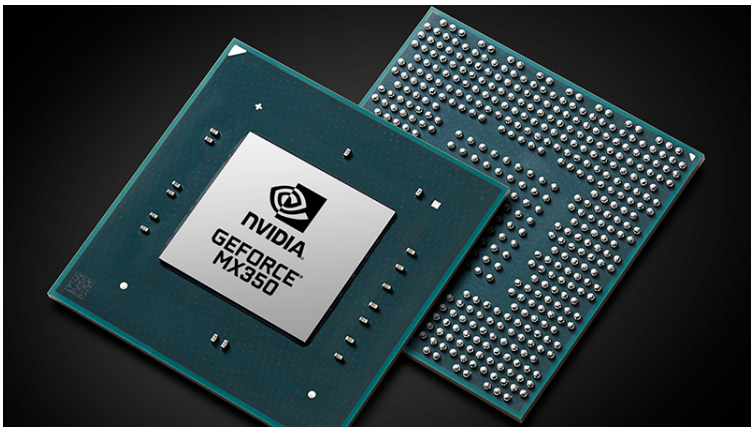
- Các framework học sâu phổ biến như TensorFlow, PyTorch, Keras đều hỗ trợ GPU, cung cấp các công cụ và hàm tích hợp sẵn để tận dụng sức mạnh tính toán của GPU mà không cần can thiệp nhiều vào mã nguồn.
- Người dùng chỉ cần cấu hình để sử dụng GPU thay vì CPU, các phép toán sẽ được tự động xử lý trên GPU một cách tối ưu.

2.2.6. Hiệu suất với các mô hình lớn

GPU rất hiệu quả trong việc huấn luyện các mô hình lớn như GPT, BERT, hoặc các mạng GAN (Generative Adversarial Networks). Các mô hình này có hàng tỷ tham số, và GPU giúp giảm đáng kể thời gian tính toán cần thiết để huấn luyện chúng.

2.2.7. Các loại GPU phổ biến cho học sâu

- **NVIDIA**: Đây là hãng sản xuất GPU hàng đầu trong lĩnh vực học sâu, với dòng GPU chuyên dụng như NVIDIA Tesla, Quadro, A100 và RTX được tối ưu hóa cho các mô hình AI.



Hình 13: NVDIA

- **AMD**: GPU AMD cũng được sử dụng, nhưng ít phổ biến hơn so với NVIDIA trong lĩnh vực học sâu do sự hỗ trợ phần mềm ít phong phú hơn.

Hình 14: AMD

Phần 3. Vai trò của GPU với học sâu

3.1. Tăng tốc độ tính toán cho các mô hình học sâu

- Xử lý tính toán song song:
  GPU với hàng nghìn lõi xử lý song song, thực hiện đồng thời nhiều phép tính một cách nhanh chóng. Điều này giúp GPU vượt trội hơn CPU trong các tác vụ học sâu, đặc biệt khi xử lý lượng dữ liệu lớn.

- Giảm thời gian huấn luyện:
  Với mạng nơ-ron nhân tạo (ANN), đặc biệt là các mô hình phức tạp như CNN (nơ-ron tích chập) và RNN (nơ-ron hồi quy), GPU giúp giảm đáng kể thời gian huấn luyện bằng cách xử lý song song các lớp và tầng trong mạng.

- Tối ưu hóa phép tính:
  Các kiến trúc GPU hiện đại đang chuyển từ phép tính 32-bit truyền thống sang 16-bit Các cải tiến như chuyển từ phép tính 32 bit sang 16 bit, mạng nơ-ron nhị phân, và bỏ qua các trọng số bằng không (ZeroSkip) giúp GPU tiết kiệm bộ nhớ và tăng tốc độ xử lý.

3.2. Hỗ trợ các thuật toán xử lý hình ảnh và thị giác máy tính

- Ứng dụng trong thị giác máy tính:
  Các mô hình học sâu như CNN thường được sử dụng trong nhận diện và phân loại hình ảnh, phát hiện vật thể, và nhận diện khuôn mặt. Các tác vụ này đòi hỏi GPU do khối lượng tính toán đồ họa và xử lý hình ảnh khổng lồ.

- Xử lý thời gian thực:
  Trong các ứng dụng yêu cầu xử lý hình ảnh và video liên tục (xe tự lái, robot,...) GPU giúp đảm bảo các mô hình học sâu có thể xử lý thông tin đầu vào trong thời gian thực, đảm bảo độ chính xác và an toàn trong các tình huống cấp bách.

3.3. Đáp ứng nhu cầu của các ứng dụng học sâu doanh nghiệp

- Hiệu suất cao cho các tác vụ quan trọng:
  Các công ty lớn sử dụng GPU để triển khai học sâu trong đám mây, cho phép các hệ

thống thực hiện các tác vụ quan trọng như: phân tích dữ liệu lớn và tối ưu hóa sản phẩm.

- Dễ dàng mở rộng quy mô:
  Khi lượng dữ liệu và độ phức tạp của mô hình tăng lên, doanh nghiệp chỉ cần tăng số lượng GPU hoặc sử dụng GPU mạnh hơn mà không cần thay đổi phần cứng.
  Ví dụ: OpenAI sử dụng NVIDIA H100 để huấn luyện các mô hình ngôn ngữ lớn như GPT-4, xử lý hàng tỷ tham số hiệu quả và giảm đáng kể thời gian huấn luyện so với thế hệ GPU trước.

3.4. Tiết kiệm năng lượng và tối ưu hóa chi phí

- Tiết kiệm năng lượng:
  GPU hiện đại sử dụng điện năng hiệu quả hơn trong các tác vụ tính toán nặng. Điều này đặc biệt hữu ích cho các hệ thống yêu cầu hoạt động liên tục hoặc các thiết bị nhúng (Ví dụ: điện thoại thông minh, thiết bị IoT)

- Tối ưu chi phí:
  Chi phí ban đầu của GPU có thể cao nhưng với khả năng xử lý nhanh chóng và hiệu quả → Giảm chi phí tổng thể do tiết kiệm thời gian và năng lượng.
  Ngoài ra, GPU có thể được tái sử dụng cho nhiều dự án khác nhau trong học sâu → Tối ưu chi phí đầu tư.

3.5. Tích hợp trong hệ thống đám mây học sâu

- Tích hợp trong các dịch vụ đám mây:
  Các nhà cung cấp dịch vụ đám mây như Amazon Web Services, Google Cloud Platform và Microsoft Azure cung cấp dịch vụ GPU chuyên dụng để triển khai học sâu trên diện rộng.
  Sử dụng GPU trong đám mây giúp doanh nghiệp dễ dàng triển khai các mô hình học sâu mà không cần đầu tư cơ sở hạ tầng phần cứng.

- Đảm bảo tính linh hoạt và tùy chỉnh: Cho phép người dùng lựa chọn số lượng GPU, dung lượng bộ nhớ, cấu hình máy ảo để tối ưu hiệu suất cho từng nhu cầu cụ thể.
  → Đảm bảo các mô hình học sâu có thể hoạt động tối ưu theo yêu cầu cụ thể của từng dự án.

Phần 4. Thách thức và tương lai

4.1. Thách thức

1. **Hạn chế về hiệu năng và tiêu thụ năng lượng:**
   GPU hiệu suất cao tiêu thụ một lượng lớn điện năng, dẫn đến chi phí vận hành cao, gây khó khăn cho việc triển khai trong các hệ thống lớn và trung tâm dữ liệu quy mô lớn. Hơn nữa, sự tiêu thụ năng lượng này còn đặt ra vấn đề về bền vững và tác động đến môi trường.

2. **Giới hạn bộ nhớ:**
   Bộ nhớ trên GPU (VRAM) thường có dung lượng giới hạn (ví dụ: 16GB hoặc 32GB), khiến việc xử lý các mô hình phức tạp như GPT-4 hay mô hình đa nhiệm trở nên khó khăn.
   Để giải quyết, các kỹ thuật như chia nhỏ mô hình (model partitioning) hoặc sử dụng bộ nhớ ngoài đã được áp dụng, nhưng chúng làm tăng độ phức tạp và thời gian tính toán.

3. **Lỗi phần cứng và chi phí:**
   GPU hoạt động ở cường độ cao có nguy cơ bị lỗi phần cứng như hỏng DRAM, lỗi bit-flip do tác động của môi trường hoặc quá nhiệt.
   Bên cạnh đó, chi phí để xây dựng một cụm GPU hiệu năng cao có thể rất lớn, từ vài triệu USD cho đến hàng chục triệu USD cho các trung tâm dữ liệu lớn.

4. **Tối ưu hóa kiến trúc và mô hình:**
   Để đạt hiệu suất tối đa, GPU cần có cấu hình phù hợp với khối lượng công việc cụ thể. Tuy nhiên, việc tối ưu hóa các yếu tố như kích thước batch, độ chính xác số học, và sử dụng tài nguyên bộ nhớ là một thách thức lớn đối với các đội ngũ kỹ thuật.
   Ví dụ: Một GPU có thể không tận dụng hết khả năng của nó nếu không được cấu hình đúng, dẫn đến tình trạng lãng phí tài nguyên và hiệu suất thấp hơn mong đợi.

5. **Thách thức trong công nghệ sản xuất và giới hạn kích thước**
   Sản xuất GPU đòi hỏi công nghệ tiên tiến, nhưng đang gặp thách thức lớn:

   - Giới hạn kích thước bóng bán dẫn: Các GPU hiện nay sử dụng công nghệ bán dẫn 3nm hoặc 5nm, nhưng việc thu nhỏ kích thước bóng bán dẫn hơn nữa (dưới 3nm) gặp trở ngại về vật liệu và độ ổn định của thiết bị.

- Chi phí sản xuất tăng: Khi giảm kích thước bóng bán dẫn, chi phí sản xuất tăng theo cấp số nhân, khiến GPU tiên tiến trở nên đắt đỏ và khó tiếp cận đối với nhiều tổ chức.
  Ví dụ: NVIDIA và AMD đang đối mặt với các vấn đề về năng suất sản xuất khi chuyển sang công nghệ 2nm, làm tăng thời gian và chi phí đưa sản phẩm ra thị trường.

4.2. Tương lai

1. **Tối ưu hóa hiệu suất và năng lượng**

   - Sử dụng toán học độ chính xác thấp: Các mạng học sâu không cần độ chính xác cao trong mọi tầng. Sử dụng số học 16-bit (hoặc thậm chí 8-bit) cho các trọng số và phép tính giúp giảm kích thước mô hình và tiết kiệm năng lượng.
   - Khai thác ma trận thưa (Sparse Matrix): Khai thác ma trận thưa bằng kỹ thuật ZeroSkip để giảm phép tính không cần thiết.
   - Tối ưu hóa GPU: Các kiến trúc như AccUDNN giảm yêu cầu bộ nhớ từ 24GB xuống còn 8GB bằng cách quản lý thông minh bộ nhớ và siêu tham số

2. **Hệ thống phân tán và môi trường kết hợp**

   - Điện toán đám mây và tại biên (Cloud and Edge Computing):
     - Điện toán đám mây sử dụng GPU (như AWS, Azure) để xử lý lượng dữ liệu lớn, phục vụ AI-as-a-Service.
     - Điện toán tại biên cho phép triển khai GPU trên các thiết bị nhỏ như camera thông minh, giảm độ trễ trong xử lý dữ liệu nhạy cảm theo thời gian thực, đặc biệt trong các ứng dụng IoT
   - Hệ thống GPU phân tán:
     - Các kiến trúc như DeepSpotCloud cho phép phân phối công việc trên nhiều GPU với khả năng giảm chi phí và đảm bảo tính liên tục khi hệ thống gặp lỗi.
     - HeteroSpark tích hợp GPU với nền tảng xử lý Big Data Spark, mang lại hiệu suất gấp 18 lần so với Spark truyền thống

3. **Phát triển ứng dụng thực tiễn**

   - Y tế và dự đoán:

  - GPU hỗ trợ xử lý nhanh các tín hiệu y tế (như ECG) và hình ảnh y tế (MRI, CT).
    Ví dụ, các mô hình học sâu có thể được huấn luyện trên GPU để phát hiện ung thư qua hình ảnh nhanh hơn và chính xác
  - Các ứng dụng như dự đoán thiên tai (lũ lụt, cháy rừng) cũng hưởng lợi từ sức mạnh của GPU
- Công nghiệp và an ninh:
  - GPU được sử dụng để tăng tốc phát hiện xâm nhập, giúp cải thiện khả năng bảo vệ hệ thống mạng trong thời gian thực.
  - Trong sản xuất, GPU hỗ trợ robot học chuyển động phức tạp và tính toán quỹ đạo tránh va chạm cho các nhóm robot.

4. **Tăng tính tùy biến**

- GPU và ReRAM (Resistive Random-Access Memory):
  Kiến trúc như GRAMARCH kết hợp GPU với ReRAM giúp tăng cường hiệu suất tính toán và tiết kiệm năng lượng, đặc biệt trong các tác vụ như phân đoạn hình ảnh
- Mạng neural nhị phân (BNNs):
  BNNs sử dụng trọng số và giá trị kích hoạt nhị phân (0 hoặc 1), giảm mạnh chi phí lưu trữ và tính toán

5. **Hướng tới tính bền vững**

- Dynamic Voltage and Frequency Scaling (DVFS):
  DVFS điều chỉnh điện áp và tần số hoạt động của GPU tùy theo nhu cầu, giúp tiết kiệm năng lượng mà không làm giảm hiệu suất.
- Điện toán xanh:
  Các công nghệ GPU mới đang cố gắng giảm lượng khí thải carbon trong quá trình sử dụng, hướng tới mục tiêu "Net-Zero Emission".

6. **Nghiên cứu và phát triển mới**

- Mô hình hóa GPU:
  Các mô hình như DeLTA phân tích hiệu suất GPU theo băng thông bộ nhớ và thông lượng tính toán, giúp tối ưu hóa thiết kế.
  Ví dụ: Mô hình DeLTA phân tích hiệu suất GPU trong dự án của Google Cloud, tối ưu thiết kế phần cứng để tăng hiệu quả xử lý các tác vụ lớn như ChatGPT.

- Học máy tối ưu tài nguyên GPU:
  Các phương pháp học máy được áp dụng để dự đoán và tối ưu hóa yêu cầu bộ nhớ của các ứng dụng CUDA, giúp giảm thiểu lãng phí tài nguyên.
  Ví dụ: Netflix áp dụng học máy để tối ưu hóa bộ nhớ GPU khi xử lý các thuật toán nén video, giúp giảm tới 40%

Phần 5. Thực nghiệm và kết quả

5.1. Tổng quan

Có hàng nghìn lõi với khả năng xử lý và tính toán song song, GPU cho thấy một vai trò quan trọng trong học sâu (Deep Learning). GPU giúp các mô hình học sâu như CNN(Convolutional Neural Network) hay YOLO(You Only Look Once) tăng tốc tính toán đối, tiết kiệm thời gian huấn luyện mô hình hay cải thiện độ chính xác của các mô hình huấn luyện đối với những bộ dữ liệu lớn.

Trong báo cáo này, chúng em sử dụng mô hình CNN(Convolutional Neural Network) để thực hiện việc xử lý và nhận dạng ảnh trên GPU và CPU, qua đó đánh giá hiệu năng cũng như cho thấy được sự hiệu quả của GPU trong lĩnh vực này.

Có 5 thông số chính sẽ được đo khi huấn luyện mô hình để phân tích hiệu quả của GPU khi so sánh GPU với CPU, là:

- Mức sử dụng bộ nhớ
- Mức tiêu hao năng lượng
- Nhiệt độ
- Thời gian huấn luyện
- Độ chính xác

Các thông số sẽ được đo trên trình quản lý tác vụ Task Manager, thông qua Command Promt với các câu lệnh hay sử dụng mã nguồn trong lập trình mô hình. Chi tiết về bài toán, mô hình, bộ dữ liệu, sẽ được trình bày chi tiết trong dưới.[1]

5.2. Thực nghiệm và kết quả

5.2.1. Thông số thiết bị

Việc thực nghiệm độ hiệu quả của GPU so với CPU trong huấn luyện các mô hình học sâu (Deep Learing) được thực hành trên Laptop Asus ROG Strix G513IH với các thông số:

- Hệ điều hành: Windows 11 Home Single Language 64-bit
- Hãng sản xuất: ASUSTek COMPUTER INC.
- CPU: AMD Ryzen 7 4800H (16 CPUs), $\sim$ 2.9GHz
- GPU: NVIDIA GeForce GTX 1650
- RAM: 16 GB
- Disk: 512 GB

5.2.2. Thực nghiệm và kết quả với bộ dữ liệu nhỏ

Thực nghiệm đầu tiên sẽ là bài toán phân loại ảnh chó/mèo. Bộ dữ liệu để huấn luyện mô hình được tải xuống miễn phí từ Kaggle. Ở thực nghiệm này sử dụng một mô hình học sâu nổi bật trong nhận diện và phân loại ảnh là mô hình Inception-v3 được phát triển bởi các nhà khoa học tại Google Brain. Tuy nhiên, mô hình sẽ được thay đổi các lớp cuối của mạng nơ ron để phù hợp với bộ dữ liệu. Tổng quát lại, ta có kiến trúc mô hình như sau:

- Mô hình: Inception-V3 đã chỉnh sửa
- Bộ huấn luyện: 557 ảnh, 52.6MB
- Bộ kiểm tra: 140 ảnh, 14.7MB
- Số lớp: 2 lớp
- Epochs: 10
- Trình tối ưu: Adam
- Hàm mất mát: Mất mát nhị phân
- Kích thước batch: 32
- Tốc độ học: 0.001

Dưới đây là thông số đã đề cập được đo từ thực nghiệm đối với GPU và CPU

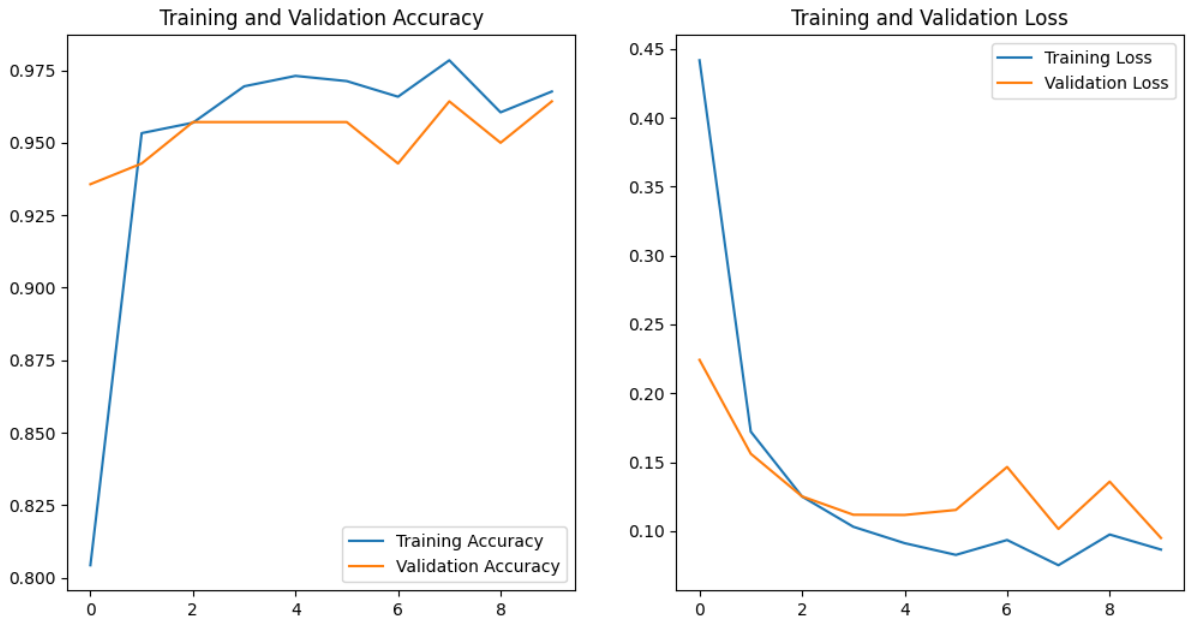
1. GPU

- Mức sử dụng bộ nhớ: 2785 MB/ 4096 MB
- Mức tiêu hao năng lượng: 26%
- Nhiệt độ: 62°C.

```
C:\Users\BUI KHANH DUY>nvidia-smi
Sun Nov 17 17:42:13 2024
+-----------------------------------------------------------------------------------------+
| NVIDIA-SMI 560.94                 Driver Version: 560.94         CUDA Version: 12.6     |
|-----------------------------------------+------------------------+----------------------+
| GPU  Name                  Driver-Model | Bus-Id          Disp.A | Volatile Uncorr. ECC |
| Fan  Temp   Perf          Pwr:Usage/Cap |           Memory-Usage | GPU-Util  Compute M. |
|                                         |                        |                  MIG M. |
|=========================================+========================+======================|
|   0  NVIDIA GeForce GTX 1650      WDDM  |   00000000:01:00.0 Off |                  N/A |
| N/A   62C    P0             20W /   65W |    2785MiB /   4096MiB |     26%      Default |
|                                         |                        |                  N/A |
+-----------------------------------------+------------------------+----------------------+
```

Hình 15: Thông số GPU khi huấn luyện

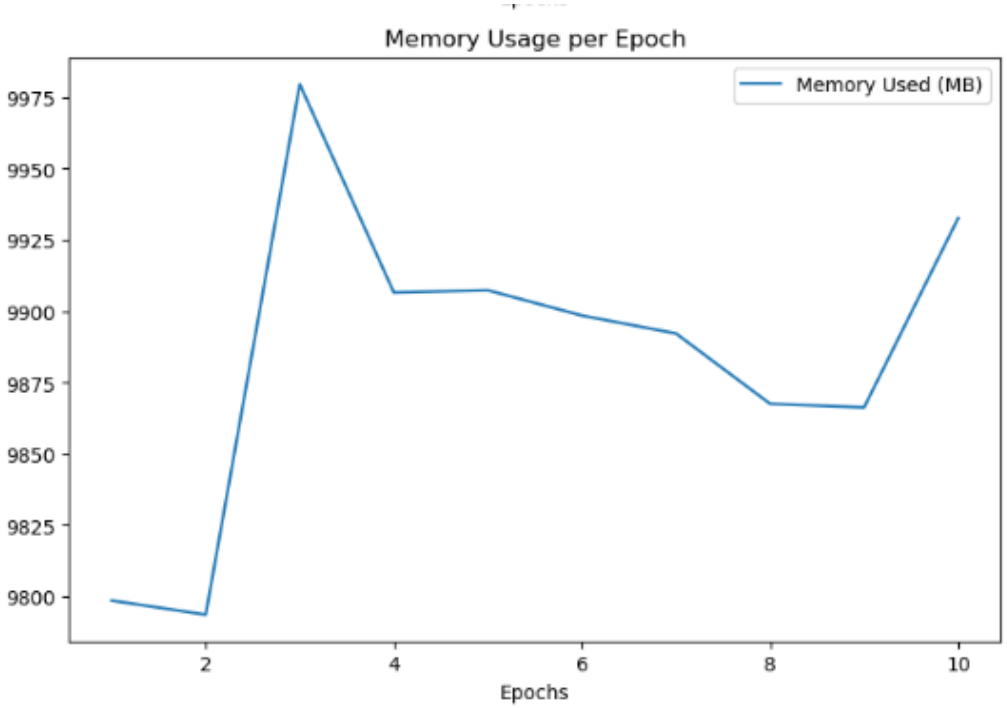
- Thời gian huấn luyện: 142.85s $\sim$ 2 phút 22s
- Độ chính xác: Đạt độ chính xác cao, dao động quanh 97.5%



Hình 16: Biểu đồ thể hiện độ chính xác khi học (GPU)

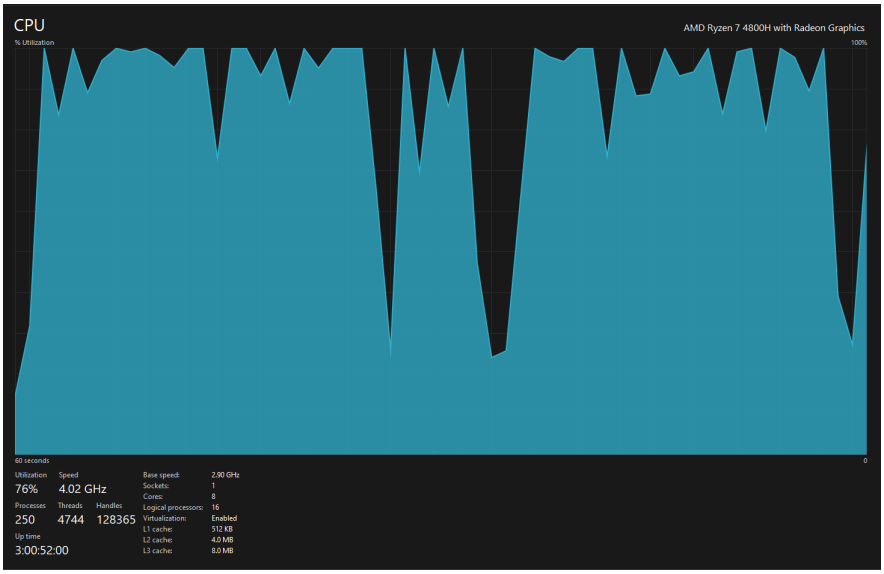
2. CPU

- Mức sử dụng bộ nhớ: Dao đông từ 9800-9900 MB



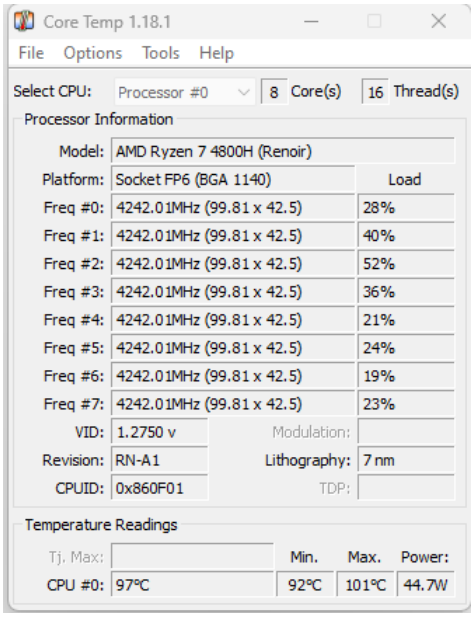
Hình 17: Mức sử dụng bộ nhớ (CPU)

- Mức tiêu hao năng lượng: 76% với tốc độ xung nhịp 4.02 GHz



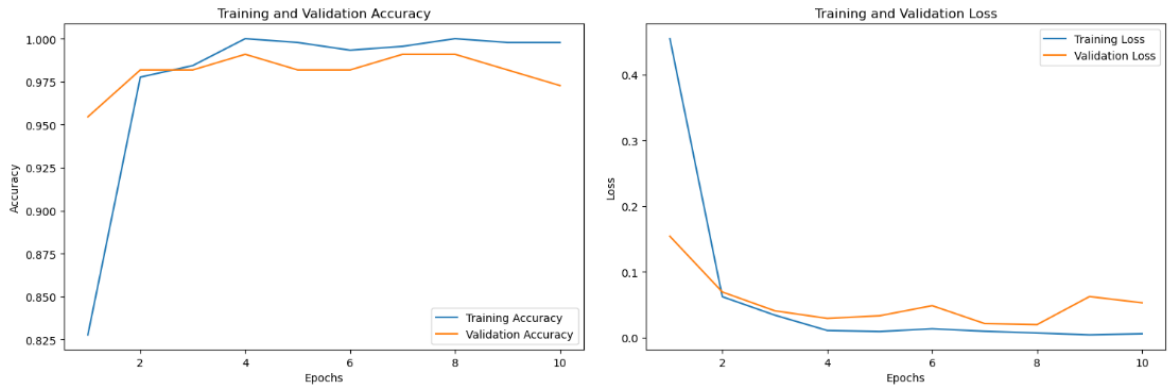
Hình 18: Mức độ tiêu hao năng lượng (CPU)

- Nhiệt độ: Trung bình 97°C, dao động từ 92-100°C



Hình 19: Nhiệt độ (CPU)

- Thời gian huấn luyện: 328.29s ∼ 5 phút 28s
- Độ chính xác: Đạt độ chính xác cao 97.5% - 99%



Hình 20: Độ chính xác mô hình (CPU)

3. So sánh

Cùng với một bộ dữ liệu nhỏ, một bộ mô hình và cùng một cấu hình thiết bị, ta thấy GPU cần sử dụng bộ nhớ ít hơn, năng lượng tiêu hao ít hơn, nhiệt độ thấp hơn và cần ít thời gian để huấn luyện hơn so với CPU. Về độ chính xác, cả GPU và CPU đều cho thấy mức độ chính xác cao, từ 97-99%. Tuy nhiên, đây chỉ là thử nghiệm trên bộ dữ liệu nhỏ nên các thông số không thể hiện rõ sự vượt trội của GPU đối với CPU.

Tiêu chí	GPU	CPU
Mức sử dụng bộ nhớ	2785 MB / 4096 MB	9800-9900 MB
Mức tiêu hao năng lượng	26%	76%
Nhiệt độ	62°C	92°C - 100°C
Thời gian huấn luyện	142.85s ∼ 2 phút 22s	328.29s ∼ 5 phút 28s
Độ chính xác	97.5%	97.5% - 99%

Bảng 1: So sánh giữa GPU và CPU

5.2.3. Thực nghiệm và kết quả với bộ dữ liệu trung bình

Thực nghiệm thứ hai sẽ là bài toán phân loại côn trùng với bộ dữ liệu để huấn luyện mô hình Insect Dataset được tải xuống miễn phí từ Kaggle. Ở thực nghiệm này sử dụng một mô hình học sâu nổi bật là mô hình EfficientNetB0. Tuy nhiên, mô hình sẽ được thay đổi các lớp cuối của mạng nơ ron để phù hợp với bộ dữ liệu. Tổng quát lại, ta có kiến trúc mô hình như sau:

- Mô hình: EfficientNetB0 đã chỉnh sửa
- Bộ huấn luyện: 1591 ảnh, 455MB
- Bộ kiểm tra: 157 ảnh, 46MB
- Số lớp: 15 lớp
- Epochs: 10
- Trình tối ưu: Adam
- Hàm mất mát: Categorical Crossentropy
- Kích thước batch: 32
- Tốc độ học: 0.001

Dưới đây là thông số đã đề cập được đo từ thực nghiệm đối với GPU và CPU

1. GPU

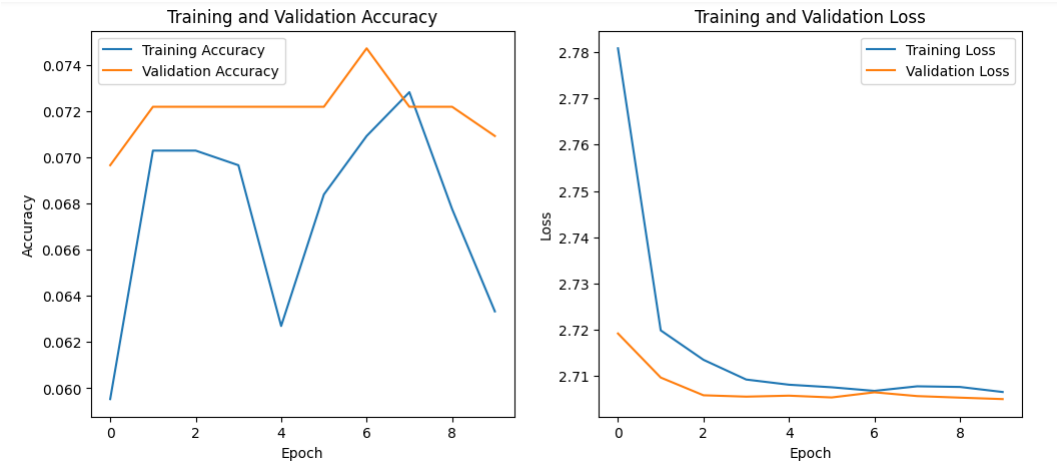
- Mức sử dụng bộ nhớ: 2785MB/4096MB
- Mức tiêu hao năng lượng: 17%
- Nhiệt độ: 54-59°C
- Thời gian huấn luyện: 2564.44s ∼ 42 phút

```
C:\Users\BUI KHANH DUY>nvidia-smi
Mon Nov 18 21:43:52 2024
+-----------------------------------------------------------------------------------------+
| NVIDIA-SMI 560.94                 Driver Version: 560.94         CUDA Version: 12.6     |
|-----------------------------------------+------------------------+----------------------+
| GPU  Name                  Driver-Model | Bus-Id          Disp.A | Volatile Uncorr. ECC |
| Fan  Temp   Perf          Pwr:Usage/Cap |           Memory-Usage | GPU-Util  Compute M. |
|                                         |                        |               MIG M. |
|=========================================+========================+======================|
|   0  NVIDIA GeForce GTX 1650      WDDM  |   00000000:01:00.0 Off |                  N/A |
| N/A   54C    P0             19W /   65W |    2785MiB /   4096MiB |     17%      Default |
|                                         |                        |                  N/A |
+-----------------------------------------+------------------------+----------------------+

+-----------------------------------------------------------------------------------------+
| Processes:                                                                              |
|  GPU   GI   CI        PID   Type   Process name                              GPU Memory |
|        ID   ID                                                               Usage      |
|=========================================================================================|
|    0   N/A  N/A      3756      C   C:\anaconda3\envs\train\python.exe            N/A      |
+-----------------------------------------------------------------------------------------+
```

Hình 21: Kết quả huấn luyện (GPU)

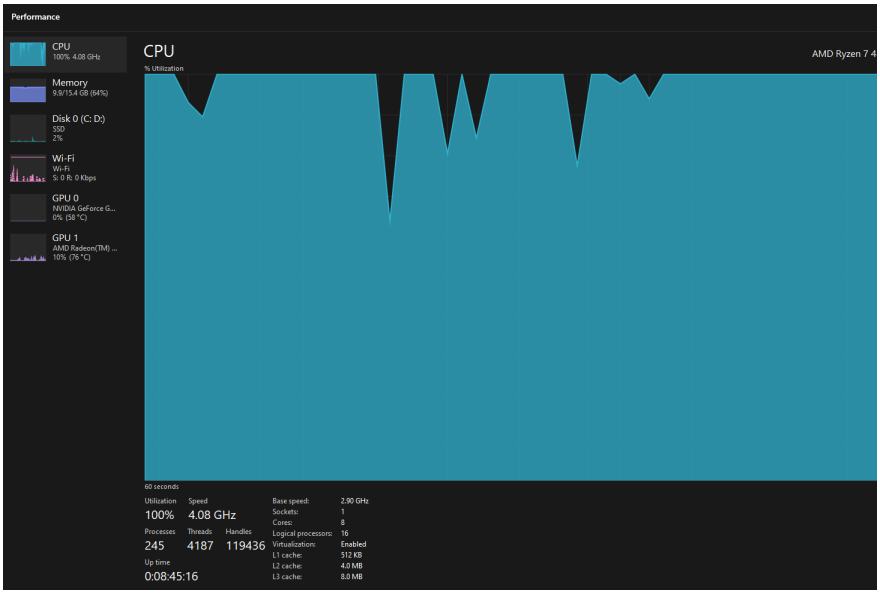
- Độ chính xác: Độ chính xác tương đối cao, 72-74%



Hình 22: Độ chính xác của mô hình (GPU)

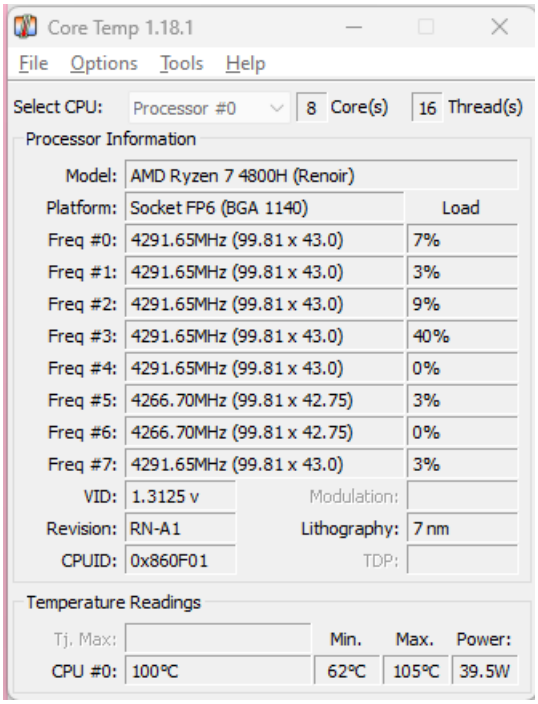
2. CPU

- Mức sử dụng bộ nhớ: 11000-12000MB
- Mức tiêu hao năng lượng: 96-100%



Hình 23: Mức tiêu hao bộ nhớ (CPU)

- Nhiệt độ: 100-105°C



Hình 24: Nhiệt độ (CPU)

- Thời gian huấn luyện: 8479.2s $\sim$ 2 giờ 20 phút
- Độ chính xác: Độ chính xác tương đối thấp, 14-16%

Hình 25: Độ chính xác của mô hình (CPU)

3. So sánh Có thể thấy, với bộ dữ liệu lớn hơn, thì đã có sự khác biệt tươn đối rõ rệt.Với bộ dữ liệu kích thước vài trăm MB thì CPU đã phải gần như sử dụng hết bộ nhớ để huấn luyện mô hình, tuy nhiên cho ra kết quả huấn luyện tệ hơn rất nhiều so với GPU và mất rất nhiều thời gian. Nhiệt độ của CPU cũng là gấp đôi đối với GPU. Qua thử nghiệm này thì ta thấy được sự ưu việt của GPU đối với học sâu khi phải huấn luyện và kiểm thử các mô hình. Đối với các bộ dữ liệu có kích thước thật sự lớn ($\geq$ 5GB) thì CPU sẽ không thể huấn luyến được mô hình và sẽ bị crash.

Tiêu chí	**GPU**	**CPU**
Mức sử dụng bộ nhớ	2785MB / 4096MB	11000-12000MB
Mức tiêu hao năng lượng	17%	96-100%
Nhiệt độ	54°C - 59°C	100°C - 105°C
Thời gian huấn luyện	2564.44s $\sim$ 42 phút	8479.2s $\sim$ 2 giờ 20 phút
Độ chính xác	72-74%	14-16%

Bảng 2: So sánh giữa GPU và CPU

Tài liệu

[1] Dipesh Gyawal, *Comparative Analysis of CPU and GPU Profiling for Deep Learning Models*

[2] Samuel Cortinhas, *Cats and Dogs Image Classification*, 2023, https://www.kaggle.com/datasets/samuelcortinhas/cats-and-dogs-image-classification. Accessed: 2024-11-19.

Đánh giá thành viên

Trong bài báo cáo này, chúng em đã phân chia các nhiệm vụ cho mỗi thành viên trong nhóm. Tuy nhiên ở mỗi giai đoạn đều có sự đóng góp và xây dựng từ tất cả các thành viên để bài báo cáo được hoàn thiện tốt nhất. Dưới đây là Phân công công việc và đánh giá đóng góp cho mỗi thành viên như sau:

- Nguyễn Bá Anh: Đề xuất chủ đề báo cáo, kiểm chứng kết quả thực nghiệm, đề xuất và góp ý cho nội dung các phần, tổng kết báo cáo.
- Vũ Việt Anh: Đề xuất chủ đề báo cáo, xác định hướng đi chính cho báo cáo, xây dựng mô hình học sâu và bộ dữ liệu, kiểm chứng kết quả thực nghiệm.
- Bùi Khánh Duy: Xây dựng mô hình học sâu và bộ dữ liệu, thực nghiệm và so sánh, tổng hợp nội dung phần 5.
- Phạm Thị Khánh: Xây dựng mẫu báo cáo, tổng hợp nội dung phần 3, rà soát, đề xuất và góp ý cho nội dung các phần.
- Lê Tiến Thành: Tổng hợp nội dung phần 1 và phần 2, tham gia đánh giá và góp ý các phần còn lại.
- Phan Thu Trang: Lựa chọn chủ đề, phân công công việc thành viên, tổng hợp nội dung phần 4, đánh giá thành viên trong nhóm, góp ý chỉnh sửa và tổng kết báo cáo.

STT	Họ và tên	MSSV	Đánh giá đóng góp
1	Vũ Việt Anh	20210031	1
2	Nguyễn Bá Anh	20203309	1
3	Bùi Khánh Duy	20227104	1
4	Phạm Thị Khánh	20227127	1
5	Lê Tiến Thành	20227070	1
6	Phan Thu Trang **(Leader)**	20227156	1

Lời cảm ơn

Báo cáo này là kết quả của quá trình học tập, tìm hiểu về GPU và vai trò của GPU đối với học sâu cũng như môn học Kiến trúc máy tính của nhóm em. Qua đây chúng em xin gửi lời cảm ơn đến cô Phạm Huyền Linh đã giảng dạy môn học và mang đến cho chúng em nhiều kiến thức bổ ích. Trong quá trình làm bài báo cáo này, nhóm chúng em không thể tránh khỏi thiếu sót, mong cô nhận xét và góp ý để nhóm em có thể hoàn thiện nội dung tốt hơn.

Chúng em xin chân thành cảm ơn cô.