**Co-Design of DNN Models and Hardware: Sparsity**

1. **Mục tiêu chính:**

Giảm số lượng phép toán và chi phí lưu trữ/tính toán bằng cách tận dụng độ thưa thớt (nhiều giá trị lặp lại, đặc biệt là số 0).

Khai thác độ thưa tự nhiên trong dữ liệu và tạo độ thưa thông qua kỹ thuật pruning (cắt tỉa).

Lợi ích kiến trúc:

* Giảm chi phí di chuyển và lưu trữ dữ liệu.
* Giảm số lượng phép toán không cần thiết.

1. **Nguồn gốc của độ thưa:**

* Độ thưa của activation: Do hàm ReLU (chuyển giá trị âm thành 0), tương quan trong dữ liệu đầu vào, và cấu trúc biểu diễn (ví dụ: đồ thị).
* Độ thưa của trọng số (weight): Thông qua sắp xếp lại trọng số, tái sử dụng và pruning mạng.

1. **Khai thác tính thưa thớt:**

* Dữ liệu thưa thớt có thể được nén để tiết kiệm không gian và năng lượng bằng cách tránh lưu trữ và di chuyển các giá trị bằng không....
* Có thể tiết kiệm thời gian và năng lượng bằng cách tránh tìm nạp các toán hạng không cần thiết và tránh các tính toán không hiệu quả

A diagram of a graph

AI-generated content may be incorrect.

1. **Các kỹ thuật cụ thể:**
2. **Nén dữ liệu thưa**

* Giảm chi phí di chuyển dữ liệu (băng thông bộ nhớ) và chi phí lưu trữ.
* Các yêu cầu bao gồm khả năng giải mã duy nhất, thuật toán nhẹ và thường không mất dữ liệu để không ảnh hưởng đến độ chính xác.
* **Yêu cầu:** Sử dụng mã hóa đơn nhất (uniquely decodable) và thuật toán nhẹ, không ảnh hưởng đến độ chính xác.

1. **Cnvlutin**

* Xử lý các lớp tích chập.
* Được xây dựng trên DaDianNao (chi phí diện tích 4,49%).
* Bỏ qua các activation bằng 0 trong tầng convolution, tăng tốc 1.37x-1.52x.

1. **Pruning activation**

* Loại bỏ các giá trị kích hoạt nhỏ (ảnh hưởng đến độ chính xác).
* Tăng tốc độ 11% (ImageNet).
* Giảm 2 lần (MNIST)

A graph of different types of data

AI-generated content may be incorrect.

1. **Khai thác ReLU:**

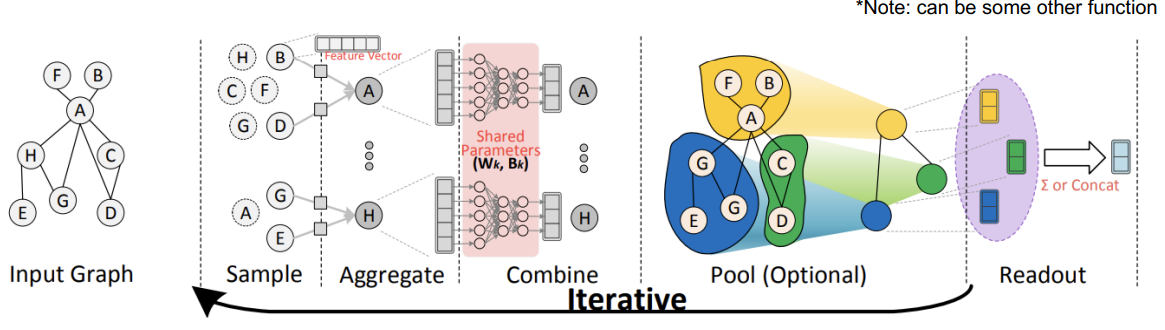
* Giảm số lượng hoạt động khi kích hoạt kết quả sẽ âm vì ReLU sẽ xuất ra số không.
* Yêu cầu thêm phần cứng để kiểm tra quyết định thời điểm kết thúc.

1. **Tương quan không gian/thời gian**

* Tận dụng sự tương quan giữa các activation hoặc khung hình liên tiếp để giảm tính toán.
* Giảm số lượng phép toán cần thiết bằng cách tận dụng sự tương đồng giữa các đầu vào lân cận(không gian).
* Việc áp dụng phương pháp này phụ thuộc vào ứng dụng cụ thể, chẳng hạn như xử lý video, nơi mà một số phép toán cần được thực hiện cho mỗi khung hình.

1. **Graph Neural Networks (GNN):**

* Cấu trúc liên kết của đồ thị có thể được biểu diễn bằng Ma trận liền kề, thường là thưa thớt.
* Mỗi nút có thể được biểu diễn bằng một vectơ đặc trưng và tập hợp các nút được biểu diễn bằng một ma trận đặc trưng.
* Các bước chính: Aggregate (tổng hợp đặc trưng từ hàng xóm để tính toán được tính năng đặc trưng trung bình) và Combine (kết hợp với trọng số để cho tính năng nút trung gian để có lớp tiếp theo tính năng nút).



1. **Khai thác trọng số dư thừa:**

* Tiền xử lý để sắp xếp lại trọng số (ok vì đã biết trọng số).
* Thực hiện phép cộng trước phép nhân để giảm số lượng phép nhân và đọc trọng số.

A black text with red text

AI-generated content may be incorrect.

1. **Pruning – Make Weights Sparse**

* Loại bỏ trọng số dựa trên độ lớn của trọng số.
* **Các bước bao gồm:**

1. Chọn kiến trúc mạng hợp lý

* Xác định số lượng lớp, số lượng nơ-ron trong mỗi lớp, v.v.

1. Huấn luyện mạng nơ-ron đến khi đạt kết quả hợp lý

* Đây là bước học ban đầu để có một mô hình hoạt động tốt.

1. Tính đạo hàm bậc hai (second derivative) của từng trọng số

* Đạo hàm bậc hai giúp xác định mức độ ảnh hưởng của trọng số đến hàm lỗi.

1. Tính "saliency" (tầm quan trọng) của từng trọng số

* Saliency thể hiện tác động của trọng số đối với lỗi huấn luyện. Trọng số có giá trị saliency thấp là ít quan trọng.

1. Sắp xếp trọng số theo saliency và xóa đi những trọng số ít quan trọng

* Những trọng số có ít ảnh hưởng sẽ bị loại bỏ để giảm số lượng tham số.

1. Quay lại bước 2 và lặp lại quá trình

* Mô hình sẽ tiếp tục được huấn luyện để điều chỉnh lại các trọng số còn lại.

|  |  |
| --- | --- |
| A graph showing the fine tuning  AI-generated content may be incorrect. | A diagram of a training scheme  AI-generated content may be incorrect. |

A diagram of a weight removal program

AI-generated content may be incorrect.

1. **Loại bỏ trọng số: Chấm điểm**

* Gán điểm cho mỗi trọng số hoặc một nhóm trọng số dựa trên tác động đến một số tiêu chí (thường là độ chính xác).
* Cắt tỉa dựa trên độ lớn (phổ biến nhất).
* Cắt tỉa dựa trên đặc trưng.
* Đồng thời xem xét tác động của mỗi trọng số đối với hiệu quả năng lượng và thông lượng.

1. **Loại bỏ trọng số: Gom nhóm**:

* Tăng mức độ thô (cấu trúc dễ hơn cho phần cứng), ít báo hiệu hơn cho vị trí của các số không (nén tốt hơn).

A diagram of a diagram

AI-generated content may be incorrect.

1. **Coarse-grained pruning:**

* Tạo cấu trúc thưa dễ khai thác trên phần cứng (ví dụ: Scalpel, tốc độ tăng 1.92x).

1. Pattern-based pruning:

* Cắt tỉa theo mẫu thay vì hàng.

1. Ranking và Fine-tuning:

* Xếp hạng trọng số/group để loại bỏ, sau đó tinh chỉnh để khôi phục độ chính xác.

A diagram of a diagram of a diagram

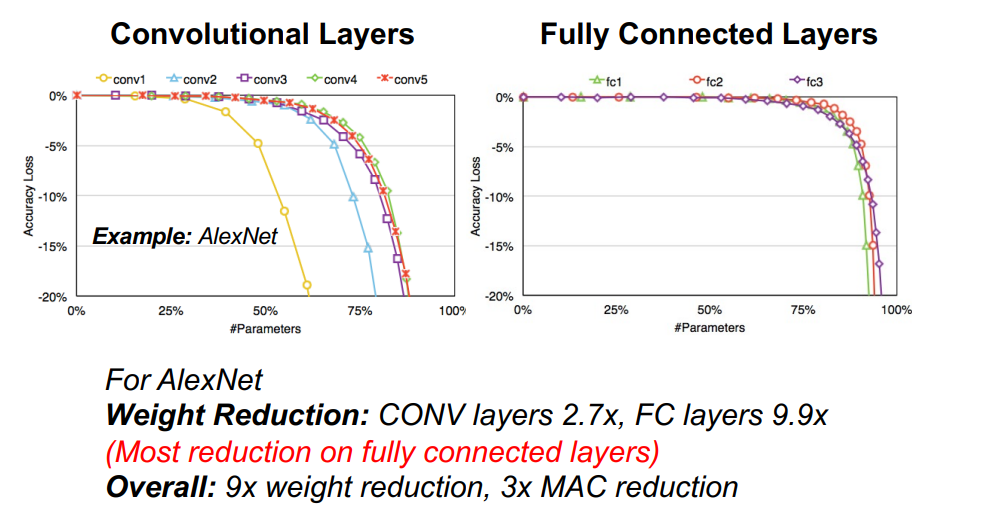
AI-generated content may be incorrect.

1. **Tinh chỉnh và Lập lịch trình**:

* Cập nhật các giá trị của các trọng số còn lại để khôi phục độ chính xác, xác định số lượng trọng số cần tỉa bớt trong mỗi lần lặp.

1. **Hiệu quả và tradeoff:**

* Pruning giảm trọng số (9x) và MAC (3x) trên AlexNet, nhưng hiệu quả khác nhau giữa các tầng (CONV vs FC).
* Các mô hình DNN hiện đại hiệu quả hơn có thể bị giảm độ chính xác nhanh hơn khi pruning.
* Sử dụng **direct metrics** (năng lượng, độ trễ) thay vì proxy (MACs) mang lại tradeoff tốt hơn.



1. **Phần cứng hỗ trợ:**

* Gating: Tắt các phép toán/lấy dữ liệu khi gặp 0 để tiết kiệm năng lượng.
* Skipping: Bỏ qua chu kỳ để tiết kiệm thời gian và năng lượng.
* Format: Chọn biểu diễn tensor tối ưu hóa lưu trữ.

1. **Kết luận:**

* Độ thưa đạt 30-70%, cần phần cứng chuyên dụng để khai thác (khác với thư viện phần mềm yêu cầu >99%).
* Pruning thô và fine-tuning tăng cường hiệu quả, nhưng cần cân nhắc tradeoff giữa tốc độ, kích thước và độ chính xác.