

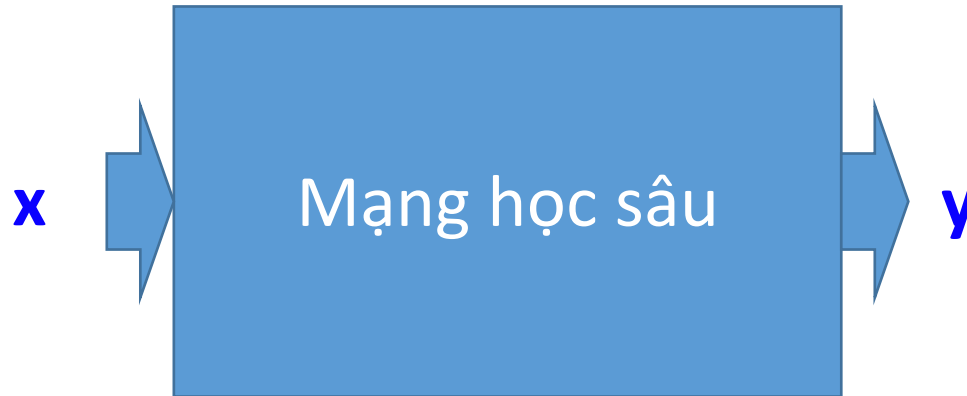
Chuẩn bị dữ liệu huấn luyện

Lê Thành Sách

NỘI DUNG

- ❖ Mô hình mạng học sâu
- ❖ Các tập dữ liệu
- ❖ Cách dùng các tập dữ liệu
- ❖ Dữ liệu có nhãn cho các bài toán
- ❖ Các yêu cầu về dữ liệu
- ❖ Xử lý thiếu dữ liệu

Mô hình mạng học sâu



Mạng học sâu = Một hàm:

$$\mathbf{y} = f_{\mathbf{w}}(\mathbf{x})$$

w: là tham số của mạng, được xác định lúc huấn luyện

Các tập dữ liệu

❖ Dữ liệu được chia thành 3 tập

➤ Tập huấn luyện (training set)

- Gồm các cặp $\langle \mathbf{x}^{(i)}, \mathbf{y}^{(i)} \rangle$
- Dùng để huấn luyện mạng, tìm ra các \mathbf{w} .

➤ Tập Kiểm thử (validation set)

- Gồm các cặp $\langle \mathbf{x}^{(i)}, \mathbf{y}^{(i)} \rangle$
- Dùng để đánh giá thử mạng làm việc như thế nào với dữ liệu không nằm trong tập huấn luyện.

Huấn luyện mạng không phải để tìm ra bộ \mathbf{w} làm cho sai số trên tập huấn luyện nhỏ, mà để làm cho sai số trên tập kiểm thử nhỏ và hy vọng rằng sai số lúc kiểm tra cũng nhỏ.

Các tập dữ liệu

❖ Dữ liệu được chia thành 3 tập

➤ Tập huấn luyện (training set)

- Gồm các cặp $\langle \mathbf{x}^{(i)}, \mathbf{y}^{(i)} \rangle$
- Dùng để huấn luyện mạng, tìm ra các \mathbf{w} .

➤ Tập Kiểm thử (validation set)

- Gồm các cặp $\langle \mathbf{x}^{(i)}, \mathbf{y}^{(i)} \rangle$
- Dùng để đánh giá thử mạng làm việc như thế nào với dữ liệu không nằm trong tập huấn luyện.

➔ Nên có kế hoạch huấn luyện mạng với các siêu tham số khác nhau nhằm chọn bộ siêu tham số cho ra sai số kiểm thử nhỏ nhất! (gọi là: hyper-parameter tuning)

Các tập dữ liệu

❖ Dữ liệu được chia thành 3 tập

- Tập huấn luyện (training set)
- Tập Kiểm thử (validation set)
- Tập Kiểm tra (Test set)
 - Gồm các cặp $\langle \mathbf{x}^{(i)}, \mathbf{y}^{(i)} \rangle$
 - Dùng để đánh giá mạng và công bố các chỉ tiêu về hiệu quả của mạng

Các tập dữ liệu

❖ Cách chia dữ liệu thành các tập

➤ Với tập dữ liệu không nhiều (~ 1000 s cặp $\langle \mathbf{x}^{(i)}, \mathbf{y}^{(i)} \rangle$)

Tập huấn luyện 60% - 70%	Tập Kiểm thử 15%-20%	Tập Kiểm tra 15%-20%

Các tập dữ liệu

❖ Cách chia dữ liệu thành các tập

➤ Với tập dữ liệu lớn ($\sim 10^6$ s cặp $\langle \mathbf{x}^{(i)}, \mathbf{y}^{(i)} \rangle$)

Tập huấn luyện 90%-95%	Tập Kiểm thử ~5%	Tập Kiểm tra ~5%

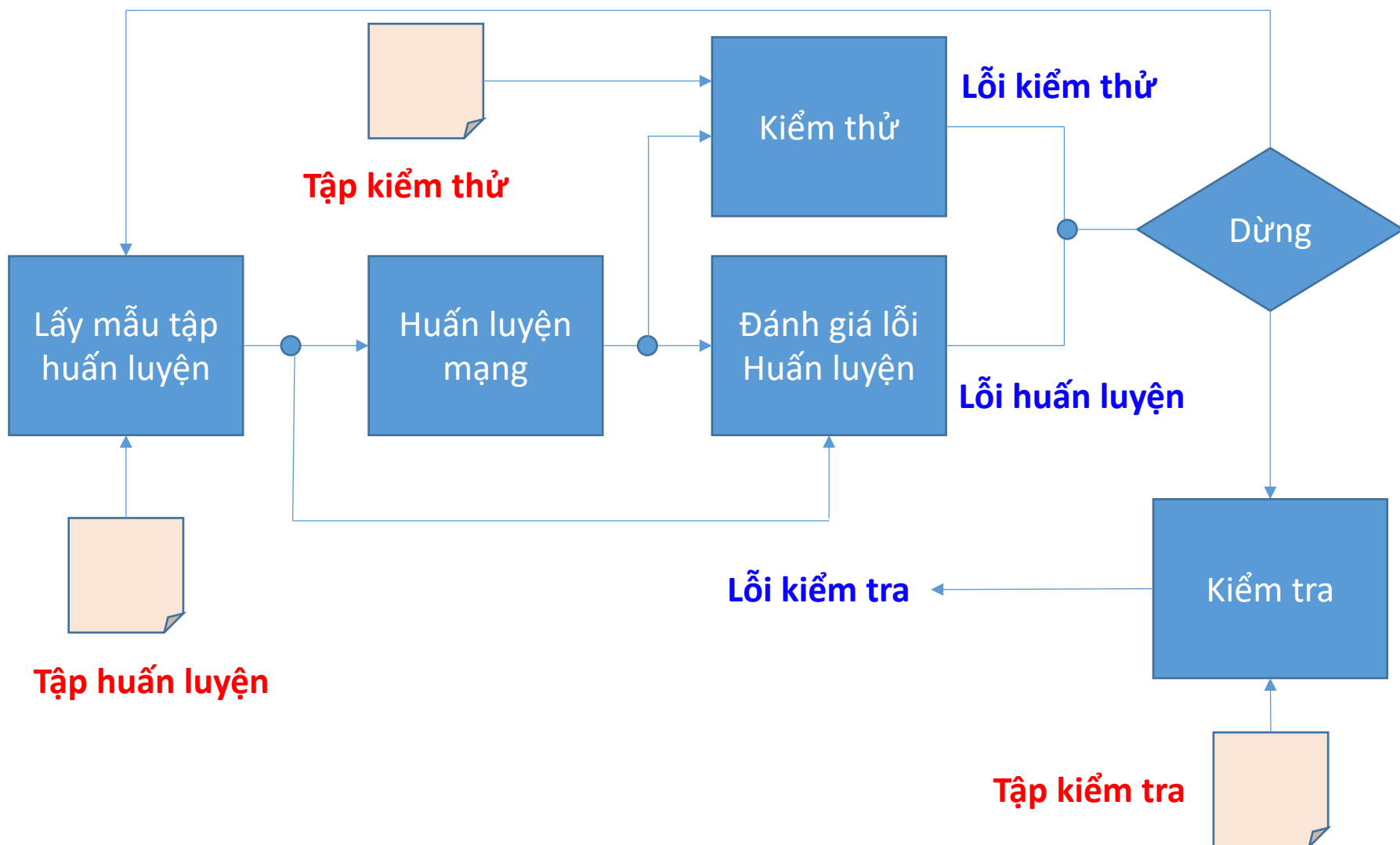
Các tập dữ liệu

❖ Cách chia dữ liệu thành các tập



➤ Với tập dữ liệu lớn ($\sim 10^6$ s cặp $\langle \mathbf{x}^{(i)}, \mathbf{y}^{(i)} \rangle$)

Tập huấn luyện 90%-95%	Tập Kiểm thử ~5%	Tập Kiểm tra ~5%

Cách dùng các tập dữ liệu






Dữ liệu có nhãn cho các bài toán

Bài toán	x	y
Nhận dạng / Phân loại đối tượng trên ảnh	<p>Ảnh đối tượng</p>  	<p>Tên nhãn (hoặc chỉ số)</p> <p>Cat (0)</p> <p>Dog (1)</p>


...

Dữ liệu có nhãn cho các bài toán

Bài toán	x	y
Phát hiện đối tượng trên ảnh	Ảnh đầu  (đã được vẽ chồng các box)	Vùng bao + Nhãn cho vùng bao  Tọa độ: x,y, w, h + dog  Tọa độ: x,y, w, h + human

...

Dữ liệu có nhãn cho các bài toán

Bài toán	x	y
Nhận dạng tiếng nói	Âm thanh 	Nội dung của đoạn âm “Tôi đi học”

...

Dữ liệu có nhãn cho các bài toán

Bài toán	x	y
Chatbot	Câu hỏi	Câu trả lời
	Bạn bao nhiêu tuổi?	Tôi 10 tuổi

...

Các yêu cầu về dữ liệu

❖ Các tập dữ liệu phải có cùng phân phối, ví dụ

➤ Bài toán phát hiện người

- Nếu muốn kiểm tra trên tập dữ liệu mà hình dáng người bất kỳ thì
 - Tập huấn luyện và kiểm thử không thể chỉ bao gồm hình người đứng hay nằm, mà phải bao gồm các hình dạng khác nhau của hình người
- Nếu muốn kiểm tra trên tập dữ liệu mà hình người xuất hiện với bối cảnh bất kỳ, thì
 - Tập huấn luyện và kiểm thử không thể chỉ có bối cảnh trong nhà

❖ Do đó, DL thực sự có thách thức về việc thu thập dữ liệu có nhãn

- Cần công cụ gán nhãn phù hợp!
- (giới thiệu về công cụ AIT của GVLab)

Xử lý thiếu dữ liệu

- ❖ Thu thập và làm nhãn thêm!
- ❖ Làm giàu nguồn dữ liệu có nhãn, ví dụ,
 - Biến đổi hình học trên ảnh: lật, xoay, dịch chuyển
 - Làm nhiều ảnh: mờ ảnh, gieo nhiễu, v.v
 - ...
- ❖ Lưu ý:
 - Phải có yếu tố ngẫu nhiên khi biến dữ liệu nguồn
 - Phải biến đổi nhãn đi kèm, ví dụ,
 - Random crop => thay đổi vị trí khung bao