Chuẩn hóa lớp

Jimmy Lei Ba Đại
học Toronto
jimmy@psi.toronto.edu

Jamie Ryan Kiros Đại
học Toronto
rkiros@cs.toronto.edu

Geoffrey E. Hinton Đại
học Toronto và Google Inc.
gợi ý@cs.toronto.edu

trừu tượng

Việc đào tạo các mạng lưới thần kinh sâu, hiện đại rất tốn kém về mặt tính toán. Một cách để giảm thời gian huấn luyện là bình thường hóa hoạt động của các tế bào thần kinh. Một kỹ thuật được giới thiệu gần đây được gọi là chuẩn hóa hàng loạt sử dụng phân phối đầu vào tổng hợp cho nơ-ron qua một loạt trường hợp huấn luyện nhỏ để tính giá trị trung bình và phương sai, sau đó được sử dụng để chuẩn hóa đầu vào tổng hợp cho nơ-ron đó trên mỗi trường hợp huấn luyện. Điều này làm giảm đáng kể thời gian đào tạo trong mạng lưới thần kinh chuyển tiếp nguồn cấp dữ liệu. Tuy nhiên, hiệu quả của việc chuẩn hóa hàng loạt phụ thuộc vào kích thước lô nhỏ và không rõ ràng về cách áp dụng nó cho các mạng thần kinh tái phát. Trong bài báo này, chúng tôi chuyển việc chuẩn hóa hàng loạt thành chuẩn hóa lớp bằng cách tính toán giá trị trung bình và phương sai được sử dụng để chuẩn hóa từ tất cả các đầu vào tổng hợp đến các nơ-ron trong một lớp trên một trường hợp huấn luyện duy nhất. Giống như chuẩn hóa hàng loạt, chúng tôi cũng cung cấp cho mỗi nơ-ron độ lệch và mức tăng thích ứng riêng được áp dụng sau khi chuẩn hóa nhưng trước phi tuyến tính. Không giống như chuẩn hóa theo lô, chuẩn hóa lớp thực hiện chính xác tính toán tương tự tại thời điểm huấn luyện và kiểm tra.

Việc áp dụng cho các mạng thần kinh hồi quy cũng đơn giản bằng cách tính toán số liệu thống kê chuẩn hóa riêng biệt ở mỗi bước thời gian. Chuẩn hóa lớp rất hiệu quả trong việc ổn định động lực trạng thái ẩn trong các mạng hồi quy. Theo kinh nghiệm, chúng tôi cho thấy rằng chuẩn hóa lớp có thể giảm đáng kể thời gian đào tạo so với các kỹ thuật đã được công bố trước đó.

1. Giới thiệu

Mạng lưới thần kinh sâu được huấn luyện bằng một số phiên bản của Stochastic gradient Descent đã được chứng minh là hoạt động tốt hơn đáng kể so với các phương pháp trước đây đối với các nhiệm vụ học tập có giám sát khác nhau trong thị giác máy tính [Krizhevsky và cộng sự, 2012] và xử lý giọng nói [Hinton và cộng sự, 2012]. Nhưng mạng lưới thần kinh sâu hiện đại thường đòi hỏi nhiều ngày đào tạo. Có thể tăng tốc độ học bằng cách tính toán gradient cho các tập hợp con khác nhau của trường hợp huấn luyện trên các máy khác nhau hoặc phân chia mạng lưới thần kinh trên nhiều máy [Dean và cộng sự, 2012], nhưng điều này có thể đòi hỏi nhiều giao tiếp và phần mềm phức tạp. Nó cũng có xu hướng dẫn đến lợi nhuận giảm dần khi mức độ song song hóa tăng lên. Một cách tiếp cận trực giao là sửa đổi các tính toán được thực hiện trong quá trình truyền tiếp của mạng nơron để giúp việc học dễ dàng hơn. Gần đây, chuẩn hóa hàng loạt [Ioffe và Szegedy, 2015] đã được đề xuất để giảm thời gian đào tạo bằng cách bao gồm các giai đoạn chuẩn hóa bổ sung trong mạng lưới thần kinh sâu. Quá trình chuẩn hóa sẽ chuẩn hóa từng đầu vào tổng hợp bằng cách sử dụng giá trị trung bình và độ lệch chuẩn của nó trên dữ liệu huấn luyện. Mạng nơ-ron Feedforward được đào tạo bằng cách sử dụng chuẩn hóa hàng loạt sẽ hội tụ nhanh hơn ngay cả với SGD đơn giản. Ngoài việc cải thiện thời gian đào tạo, tính ngẫu nhiên từ số liệu thống kê theo lô đóng vai trò điều chỉnh trong quá trình đào tạo.

Mặc dù đơn giản nhưng việc chuẩn hóa hàng loạt yêu cầu tính trung bình của các thống kê đầu vào được tính tổng. Trong các mạng chuyển tiếp nguồn cấp dữ liệu có độ sâu cố định, việc lưu trữ số liệu thống kê riêng biệt cho từng lớp ẩn là điều đơn giản. Tuy nhiên, tổng đầu vào của các nơ-ron hồi quy trong mạng nơ-ron hồi quy (RNN) thường thay đổi theo độ dài của chuỗi nên việc áp dụng chuẩn hóa hàng loạt cho RNN dường như yêu cầu các số liệu thống kê khác nhau cho các bước thời gian khác nhau. Hơn nữa, chuẩn hóa hàng loạt

không thể áp dụng điều này cho các nhiệm vụ học tập trực tuyến hoặc cho các mô hình phân tán cực lớn trong đó các đợt nhỏ phải nhỏ.

Bài báo này giới thiệu chuẩn hóa lớp, một phương pháp chuẩn hóa đơn giản để cải thiện tốc độ huấn luyện cho các mô hình mạng nơ-ron khác nhau. Không giống như chuẩn hóa hàng loạt, phương pháp được đề xuất ước tính trực tiếp số liệu thống kê chuẩn hóa từ các đầu vào được tổng hợp đến các nơ-ron trong một lớp ẩn để việc chuẩn hóa không đưa ra bất kỳ sự phụ thuộc mới nào giữa các trường hợp huấn luyện. Chúng tôi cho thấy rằng chuẩn hóa lớp hoạt động tốt cho RNN và cải thiện cả thời gian đào tạo cũng như hiệu suất tổng quát hóa của một số mô hình RNN hiện có.

2 Bối cảnh

Mạng nơ-ron chuyển tiếp nguồn cấp dữ liệu là ánh xạ phi tuyến tính từ mẫu đầu vào x đến vectơ đầu ra y. Hãy coi l là biểu diễn vectơ của các đầu vào tổng hợp cho l^{th} lớp ẩn trong mạng thần kinh chuyển tiếp sâu (bỏ qua chỉ số trường hợp). Các đầu vào tổng hợp này được tính toán thông qua phép chiếu tuyến tính với ma trận trọng số W^l và các đầu vào h từ dưới lên được cho như sau:

$$a_i^{l+1} = w_i^l h_i^l \qquad h_i^{l+1} = f(a_i^{l+1} + b_i^{l+1}) \tag{1}$$

trong đó f(·) là hàm phi tuyến tính theo phần tử và w_i^l là trọng số đến của i th đơn vị ẩn và b_i^l là tham số độ lệch vô hướng. Các tham số W^l trong mạng nơ-ron được học bằng thuật toán tối ưu hóa dựa trên độ dốc với độ dốc được tính toán bằng cách truyền ngược.

Một trong những thách thức của học sâu là độ dốc liên quan đến trọng số trong một lớp phụ thuộc nhiều vào đầu ra của các nơ-ron ở lớp trước, đặc biệt nếu các đầu ra này thay đổi theo cách có mối tương quan cao. Chuẩn hóa hàng loạt [Ioffe và Szegedy, 2015] đã được đề xuất để giảm "sự dịch chuyển đồng biến" không mong muốn như vậy. Phương pháp này chuẩn hóa các đầu vào tổng hợp cho từng đơn vị ẩn trong các trường hợp huấn luyện. Cụ thể, đối với lớp i, phương pháp chuẩn hóa hàng loạt sẽ điều chỉnh lại các đầu vào tổng hợp theo phương sai của chúng theo phân phối tổng hợp trong l^{th} lớp và đầu vào tổng hợp trong l th lớp sẽ thu được phân phối chuẩn với giá trị trung bình bằng 0 và phương sai đơn vị:

$$\bar{a}_i^l = \frac{g_i^l}{\sigma_i^l}(a_i^l - \mu_i^l) \qquad \mu_i^l = E_x[a_i^l] \qquad \sigma_i^l = \sqrt{E_x[(a_i^l - \mu_i^l)^2]} \tag{2}$$

tôi ở đâu $\bar{a}_i^l$ được chuẩn hóa tổng các đầu vào cho i th đơn vị ẩn thứ trong l lớp và gi là tham số khuếch đại chia tỷ lệ kích hoạt chuẩn hóa trước hàm kích hoạt phi tuyến tính. Lưu ý rằng kỳ vọng nằm trong toàn bộ quá trình phân phối dữ liệu đào tạo. Thông thường, việc tính toán các kỳ vọng trong biểu thức là không thực tế. (2) chính xác, vì nó sẽ yêu cầu chuyển tiếp qua toàn bộ tập dữ liệu huấn luyện với tập trọng số hiện tại. Thay vào đó, μ và σ được ước tính bằng cách sử dụng các mẫu thực nghiệm từ lô nhỏ hiện tại. Điều này đặt ra những hạn chế về kích thước của một lô nhỏ và khó áp dụng cho các mạng thần kinh tái phát.

Chuẩn hóa 3 lớp

Bây giờ chúng ta xem xét phương pháp chuẩn hóa lớp được thiết kế để khắc phục những hạn chế của chuẩn hóa hàng loạt.

Lưu ý rằng những thay đổi về đầu ra của một lớp sẽ có xu hướng gây ra những thay đổi tương quan cao về tổng đầu vào của lớp tiếp theo, đặc biệt là với các đơn vị ReLU có đầu ra có thể thay đổi rất nhiều. Điều này cho thấy vấn đề "dịch chuyển hiệp phương sai" có thể được giảm bớt bằng cách sửa giá trị trung bình và phương sai của tổng các đầu vào trong mỗi lớp. Do đó, chúng tôi tính toán số liệu thống kê chuẩn hóa lớp trên tất cả các đơn vị ẩn trong cùng một lớp như sau:

$$\mu^l = \frac{1}{H}\sum_{i=1}^{H} a_i^l \qquad \sigma^l = \sqrt{\frac{1}{H}\sum_{i=1}^{H}(a_i^l - \mu^l)^2} \tag{3}$$

trong đó H biểu thị số lượng đơn vị ẩn trong một lớp. Sự khác biệt giữa phương trình. (2) và phương trình. (3) là trong quá trình chuẩn hóa lớp, tất cả các đơn vị ẩn trong một lớp có chung các thuật ngữ chuẩn hóa μ và σ, nhưng các trường hợp huấn luyện khác nhau có các thuật ngữ chuẩn hóa khác nhau. Không giống như chuẩn hóa hàng loạt, chuẩn hóa lớp không áp đặt bất kỳ ràng buộc nào đối với kích thước của lô nhỏ và nó có thể được sử dụng trong chế độ trực tuyến thuần túy với kích thước lô 1.

3.1 Mạng nơ-ron tái phát được chuẩn hóa theo lớp

Các mô hình trình tự gần đây [Sutskever và cộng sự, 2014] sử dụng mạng thần kinh hồi quy nhỏ gọn để giải quyết các vấn đề dự đoán tuần tự trong xử lý ngôn ngữ tự nhiên. Thông thường, các nhiệm vụ NLP có độ dài câu khác nhau cho các trường hợp đào tạo khác nhau. Điều này rất dễ giải quyết trong RNN vì các trọng số giống nhau được sử dụng ở mọi bước. Nhưng khi áp dụng chuẩn hóa hàng loạt cho RNN một cách rõ ràng, chúng ta cần tính toán và lưu trữ số liệu thống kê riêng biệt cho từng bước thời gian trong một chuỗi. Đây là vấn đề nếu chuỗi kiểm tra dài hơn bất kỳ chuỗi huấn luyện nào. Chuẩn hóa lớp không gặp phải vấn đề như vậy vì các thuật ngữ chuẩn hóa của nó chỉ phụ thuộc vào tổng đầu vào của một lớp ở bước thời gian hiện tại. Nó cũng chỉ có một bộ tham số độ lợi và độ lệch được chia sẻ trên tất cả các bước thời gian.

Trong RNN tiêu chuẩn, các đầu vào tổng hợp trong lớp hồi quy được tính từ đầu vào hiện tại. x^t và vectơ trước đó của trạng thái ẩn h^{t-1} được tính toán như một $a^t = Ở^{t-1} + Wxhx^t$. lớp tái phát được chuẩn hóa lớp tái căn giữa và chia tỷ lệ lại các kích hoạt của nó bằng cách sử dụng các thuật ngữ chuẩn hóa bổ sung tương tự như biểu thức. (3):

$$h^t = f\left(\frac{g}{\sigma^t}(a^t - \mu^t) + b\right) \qquad \mu^t = \frac{1}{H}\sum_{i=1}^{H} a_i^t \qquad \sigma^t = \sqrt{\frac{1}{H}\sum_{i=1}^{H}(a_i^t - \mu^t)^2} \tag{4}$$

trong đó Whh là trọng số ẩn truy hồi đối với các trọng số ẩn và Wxh là đầu vào từ dưới lên của các trọng số ẩn. là phép nhân phần tử giữa hai vectơ . b và g được định nghĩa là các tham số độ lệch và độ lợi có cùng kích thước với h^t.

Trong RNN tiêu chuẩn, có xu hướng độ lớn trung bình của các đầu vào tổng hợp đối với các đơn vị cho thuê định kỳ tăng hoặc giảm ở mỗi bước thời gian, dẫn đến độ dốc bùng nổ hoặc biến mất. Trong RNN được chuẩn hóa của lớp, các thuật ngữ chuẩn hóa làm cho việc chia tỷ lệ lại tất cả các đầu vào được tổng hợp thành một lớp là bất biến, điều này dẫn đến động lực ẩn-ẩn ổn định hơn nhiều.

4 Công việc liên quan

Chuẩn hóa hàng loạt trước đây đã được mở rộng cho các mạng thần kinh tái phát [Laurent và cộng sự, 2015, Amodei và cộng sự, 2015, Cooijmans và cộng sự, 2016]. Công trình trước đây [Cooijmans và cộng sự, 2016] cho thấy hiệu suất tốt nhất của việc chuẩn hóa hàng loạt định kỳ đạt được bằng cách giữ số liệu thống kê chuẩn hóa độc lập cho từng bước thời gian. Các tác giả cho thấy rằng việc khởi tạo tham số khuếch đại trong lớp chuẩn hóa lô lặp lại thành 0,1 sẽ tạo ra sự khác biệt đáng kể về hiệu suất cuối cùng của mô hình. Công việc của chúng tôi cũng liên quan đến việc bình thường hóa cân nặng [Salimans và Kingma, 2016]. Trong chuẩn hóa trọng số, thay vì phương sai, định mức L2 của các trọng số đến được sử dụng để chuẩn hóa tổng các đầu vào của một nơ-ron. Việc áp dụng chuẩn hóa trọng lượng hoặc chuẩn hóa hàng loạt bằng cách sử dụng số liệu thống kê dự kiến tương đương với việc có một tham số hóa khác của mạng thần kinh chuyển tiếp nguồn cấp dữ liệu ban đầu. Việc tái tham số hóa trong mạng ReLU đã được nghiên cứu trong SGD chuẩn hóa đường dẫn [Neyshabur và cộng sự, 2015]. Tuy nhiên, phương pháp chuẩn hóa lớp được đề xuất của chúng tôi không phải là tái tham số hóa mạng nơ-ron ban đầu. Do đó, mô hình chuẩn hóa lớp có các thuộc tính bất biến khác với các phương pháp khác mà chúng ta sẽ nghiên cứu trong phần sau.

5 Phân tích

Trong phần này, chúng tôi điều tra các thuộc tính bất biến của các sơ đồ chuẩn hóa khác nhau.

5.1 Tính bất biến theo trọng số và chuyển đổi dữ liệu

Việc chuẩn hóa lớp được đề xuất có liên quan đến chuẩn hóa lô và chuẩn hóa trọng số. Mặc dù, các đại lượng vô hướng chuẩn hóa của chúng được tính toán khác nhau, nhưng các phương pháp này có thể được tóm tắt là chuẩn hóa tổng các đầu vào ai cho một nơ-ron thông qua hai đại lượng vô hướng μ và σ. Họ cũng học được độ lệch thích ứng b và đạt được g cho mỗi nơ-ron sau khi chuẩn hóa.

$$\text{xin chào} = f\left(\frac{gi}{\sigma i}(ai - \mu i) + bi\right) \tag{5}$$

Lưu ý rằng đối với chuẩn hóa lớp và chuẩn hóa lô, μ và σ được tính theo biểu thức. 2 và 3. Trong chuẩn hóa trọng số, μ là 0 và σ = w2.

	Ma trận trọng số Định tâm lại	Ma trận trọng số Định lại tỷ lệ	vectơ trọng lượng	Chia tỷ lệ lại tập dữ liệu	Định tâm lại tập dữ liệu	Trường hợp đào tạo duy nhất mở rộng quy mô lại
định mức hàng loạt	bất biến	KHÔNG	bất biến	bất biến	bất biến	KHÔNG
Định mức cân nặng	bất biến	KHÔNG	bất biến	KHÔNG	KHÔNG	KHÔNG
Định mức lớp	bất biến	bất biến	KHÔNG	bất biến	KHÔNG	bất biến

Bảng 1: Thuộc tính bất biến theo phương pháp chuẩn hóa.

Bảng 1 nêu bật các kết quả bất biến sau đây cho ba phương pháp chuẩn hóa.

Định lại tỷ lệ và định tâm lại trọng lượng: Trước tiên, hãy quan sát rằng trong quá trình chuẩn hóa và trọng lượng theo lô chuẩn hóa, bất kỳ việc điều chỉnh lại tỷ lệ nào theo các trọng số wi của một nơ-ron đơn lẻ đều không ảnh hưởng đến tổng hợp đầu vào được chuẩn hóa cho một nơ-ron. Nói chính xác, theo chuẩn hóa lô và trọng lượng, nếu vectơ trọng số được chia tỷ lệ theo δ, hai đại lượng vô hướng μ và σ cũng sẽ được chia tỷ lệ theo δ. Việc bình thường hóa tổng đầu vào vẫn giữ nguyên trước và sau khi chia tỷ lệ. Vì vậy, việc chuẩn hóa lô và trọng lượng là bất biến đối với việc thay đổi tỷ lệ của các trọng số. Mặt khác, chuẩn hóa lớp không phải là bất biến đến tỷ lệ riêng của các vectơ trọng số đơn. Thay vào đó, việc chuẩn hóa lớp là bất biến đối với chia tỷ lệ của toàn bộ ma trận trọng số và bất biến đối với sự dịch chuyển sang tất cả các trọng số đến trong ma trận trọng số. Cho có hai bộ tham số mô hình θ, θ có ma trận trọng số W và W khác nhau bởi hệ số tỷ lệ δ và tất cả các trọng số đến trong W cũng được dịch chuyển bởi một hằng số vectơ γ, tức là W = δW + 1γ . Dưới sự chuẩn hóa lớp, hai mô hình tính toán hiệu quả cùng một đầu ra:

$$h = f\left(\frac{g}{\sigma}(Wx - \mu) + b\right) = f\left(\frac{g}{\sigma}(\delta W + 1\gamma)x - \mu + b\right)$$
$$= f\left(\frac{g}{\sigma}(Wx - \mu) + b\right) = h. \quad (6)$$

Lưu ý rằng nếu việc chuẩn hóa chỉ được áp dụng cho đầu vào trước các trọng số thì mô hình sẽ không được bất biến đối với việc thay đổi tỷ lệ và định tâm lại các trọng số.

Chia tỷ lệ lại và định tâm lại dữ liệu: Chúng tôi có thể chỉ ra rằng tất cả các phương pháp chuẩn hóa là bất biến để điều chỉnh lại tỷ lệ tập dữ liệu bằng cách xác minh rằng tổng đầu vào của các nơ-ron không đổi trong những thay đổi. Hơn nữa, chuẩn hóa lớp là bất biến đối với việc mở rộng lại quy mô của các trường hợp đào tạo riêng lẻ, bởi vì các vô hướng chuẩn hóa μ và σ trong biểu thức. (3) chỉ phụ thuộc vào dữ liệu đầu vào hiện tại. Hãy để x là điểm dữ liệu mới thu được bằng cách thay đổi tỷ lệ x theo δ. Sau đó chúng tôi có,

$$h' = f\left(\frac{g_i}{\sigma}(w' x - \mu + b_i) = f\left(\frac{g_i}{\delta\sigma}(\delta w' x - \delta\mu + b_i) = h_i\right.\right). \quad (7)$$

Có thể dễ dàng nhận thấy việc thay đổi tỷ lệ các điểm dữ liệu riêng lẻ không làm thay đổi dự đoán của mô hình dưới lớp bình thường hóa. Tương tự như việc định tâm lại ma trận trọng số trong chuẩn hóa lớp, chúng ta cũng có thể cho thấy rằng việc chuẩn hóa hàng loạt là bất biến đối với việc căn giữa lại tập dữ liệu.

5.2 Hình học không gian tham số trong quá trình học

Chúng tôi đã nghiên cứu tính bất biến của dự đoán của mô hình khi định tâm lại và định tỷ lệ lại của những thông số. Tuy nhiên, việc học có thể hoạt động rất khác nhau dưới các tham số hóa khác nhau, mặc dù các mô hình thể hiện chức năng cơ bản giống nhau. Trong phần này chúng ta phân tích việc học hành vi thông qua hình học và đa tạp của không gian tham số. Chúng tôi chỉ ra rằng vô hướng chuẩn hóa σ có thể ngầm làm giảm tốc độ học và làm cho việc học ổn định hơn.

5.2.1 Hệ mét Riemann

Các tham số có thể học được trong một mô hình thống kê tạo thành một đa tạp trơn tru bao gồm tất cả các tham số có thể có. quan hệ đầu vào-đầu ra của mô hình. Đối với các mô hình có đầu ra là phân bố xác suất, cách để đo sự phân tách của hai điểm trên đa tạp này là sự phân kỳ Kullback-Leibler giữa các phân phối đầu ra mô hình của chúng. Theo thước đo phân kỳ KL, không gian tham số là một đa tạp Riemannian.

Độ cong của đa tạp Riemannian hoàn toàn được nắm bắt bởi mêtric Riemannian của nó, mà dạng bậc hai được ký hiệu là ds^2 . Đó là khoảng cách vô cùng nhỏ trong không gian tiếp tuyến tại một điểm trong không gian tham số. Một cách trực quan, nó đo lường những thay đổi trong đầu ra của mô hình từ tham số không gian theo hướng tiếp tuyến. Số liệu Riemannian theo KL đã được nghiên cứu trước đây [Amari, 1998] và được chứng minh là gần đúng với khai triển Taylor bậc hai bằng cách sử dụng Fisher

ma trận thông tin:

$$ds^2 = D_{KL}\left(P(y \mid x;\theta)\,\|\,P(y \mid x;\theta+\delta)\right) \approx \frac{1}{2}\,\delta^T F(\theta)\,\delta, \tag{số 8}$$

$$F(\theta) = \underset{x\sim P(x),\, y\sim P(y\mid x)}{\mathbb{E}}\left[\frac{\partial \log P(y \mid x;\theta)}{\partial \theta}\,\frac{\partial \log P(y \mid x;\theta)}{\partial \theta}\right], \tag{9}$$

trong đó, δ là một thay đổi nhỏ đối với các tham số. Số liệu Riemannian ở trên trình bày một cái nhìn hình học về không gian tham số. Phân tích sau đây về số liệu Riemannian cung cấp một số cái nhìn sâu sắc về cách các phương pháp chuẩn hóa có thể giúp ích trong việc đào tạo mạng lưới thần kinh.

5.2.2 Hình học của mô hình tuyến tính tổng quát chuẩn hóa

Chúng tôi tập trung phân tích hình học vào mô hình tuyến tính tổng quát. Các kết quả từ phân tích sau đây có thể dễ dàng được áp dụng để hiểu các mạng lưới thần kinh sâu với phép tính gần đúng đường chéo khối với ma trận thông tin Fisher, trong đó mỗi khối tương ứng với các tham số cho một tế bào thần kinh.

Một mô hình tuyến tính tổng quát (GLM) có thể được coi là tham số hóa phân bố đầu ra từ họ hàm mũ bằng cách sử dụng vectơ trọng số w và độ lệch vô hướng b. Để nhất quán với các phần trước, khả năng ghi nhật ký của GLM có thể được viết bằng cách sử dụng tổng đầu vào a như sau:

$$\log P(y \mid x; w, b) = \frac{(a+b)y - \eta(a+b)}{\varphi} + c(y,\varphi), \tag{10}$$

$$\mathbb{E}[y \mid x] = f(a+b) = f(wx+b), \quad \mathrm{Var}[y \mid x] = \varphi f'(a+b), \tag{11}$$

trong đó f(·) là hàm truyền tương tự của phi tuyến tính trong mạng nơ ron, f′(·) là đạo hàm của hàm truyền, η(·) là một hàm có giá trị thực và c(·) là hàm phân vùng nhật ký. φ là hằng số đo lường phương sai đầu ra. Giả sử vectơ đầu ra H chiều y = [y1, y2, · · · , yH] được mô hình hóa bằng cách sử dụng H GLM độc lập và log P(y | x; W, b) = $\sum_{i=1}^{H}$ log P(yi | x; wi , bi) . Cho W là ma trận trọng số có các hàng là vectơ trọng số của các GLM riêng lẻ, b biểu thị vectơ thiên vị có độ dài H và vec(·) biểu thị toán tử vectơ Kronecker. Ma trận thông tin Fisher cho GLM đa chiều đối với các tham số của nó θ = [w 1 , b1, · · · bH] = vec([W, b]) đơn giản là tích Kronecker dự kiến, của các đặc tính dữ liệu và đầu ra ma trận hiệp phương sai:

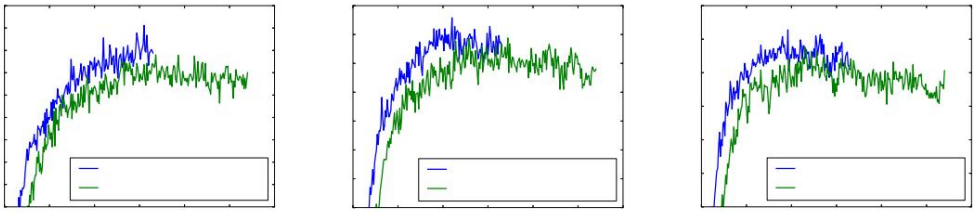
$$F(\theta) = \underset{x\sim P(x)}{\mathbb{E}}\left[\frac{\mathrm{Cov}[y \mid x]}{\varphi^2}\right] \otimes \begin{bmatrix} x x^T & x \\ x^T & 1 \end{bmatrix}. \tag{12}$$

Chúng tôi thu được các GLM chuẩn hóa bằng cách áp dụng các phương pháp chuẩn hóa cho tổng đầu vào a trong mô hình ban đầu thông qua μ và σ. Không mất tính tổng quát, chúng tôi biểu thị $\bar{F}$ là ma trận thông tin Fisher trong GLM đa chiều đã chuẩn hóa với các tham số khuếch đại bổ sung θ = vec([W, b, g]):

$$\bar{F}(\theta) = \begin{bmatrix} \bar{F}_{11} & \cdots & \bar{F}_{1H} \\ \vdots & \ddots & \vdots \\ \bar{F}_{H1} & \cdots & \bar{F}_{HH} \end{bmatrix}, \quad \bar{F}_{ij} = \underset{x\sim P(x)}{\mathbb{E}}\left[\frac{\mathrm{Cov}[y_i, y_j \mid x]}{\varphi^2}\right] \otimes \begin{bmatrix} \frac{g_i g_j}{\sigma_i\sigma_j}\chi_i\chi_j & \chi_i\frac{g_i}{\sigma_i} & \chi_i\frac{g_i(a_j-\mu_j)}{\sigma_i\sigma_j} \\ g_j\chi_j & 1 & \frac{a_j-\mu_j}{\sigma_j} \\ \frac{\sigma_j g_j}{(a_i-\mu_i)\chi_j\sigma_i\sigma_j} & \frac{a_i-\mu_i}{\sigma_i} & \frac{\sigma_j(a_i-\mu_i)(a_j-\mu_j)}{\sigma_i\sigma_j} \end{bmatrix} \tag{13}$$

$$\mu_i\chi_i = \underset{x}{\mathbb{E}}\left[\frac{a_i-\mu_i}{\sigma_i}\right]\frac{\sigma_i}{w_i}. \tag{14}$$

Giảm tốc độ học ngầm thông qua sự tăng trưởng của vectơ trọng số: Lưu ý rằng, so với GLM tiêu chuẩn, khối $\bar{F}_{ij}$ dọc theo hướng của vectơ trọng số wi được chia tỷ lệ bởi các tham số khuếch đại và vô hướng chuẩn hóa σi . Nếu chuẩn của vectơ trọng số wi tăng gấp đôi, mặc dù đầu ra của mô hình vẫn giữ nguyên thì ma trận thông tin Fisher sẽ khác.
Độ cong dọc theo hướng wi sẽ thay đổi theo hệ số vì σi cũng sẽ lớn gấp đôi. Kết quả là, đối với cùng một cập nhật tham số trong mô hình chuẩn hóa, định mức của vectơ trọng số sẽ kiểm soát hiệu quả tốc độ học của vectơ trọng số. Trong quá trình học, việc thay đổi hướng của vectơ trọng số với chuẩn lớn sẽ khó hơn. Do đó, các phương pháp chuẩn hóa



(a) Nhớ lại@1 (b) Thu hồi@5 (c) Thu hồi@10

Hình 1: Các đư ờng cong Recall@K sử dụng cách nhúng thứ tự có và không có chuẩn hóa lớp.

	MSCOCO					
	Truy xuất phụ đề			Thu hồi hình ảnh		
Ngư ời mẫu	R@1	R@5	R@10 Trung bình r	R@1	R@5	R@10 Trung bình r
Sym [Vendrov và cộng sự, 2016] 45,4 OE	88,7	5,8	36,3	85,8	9,0	
[Vendrov và cộng sự, 2016] 46,7 OE (của	88,9	5,7	37,9	85,9	8.1	
chúng tôi) 46,6 79,3 OE + LN 48,5 80,6	89,1	5,2	37,8 73,6 38,9	85,7	7,9	
	89,8	5,1	74,3	86,3	7,6	

Bảng 2: Kết quả trung bình qua 5 lần phân tách thử nghiệm về khả năng truy xuất chú thích và hình ảnh. R@K là Thu hồi@K (cao là tốt). Mean r là thứ hạng trung bình (thấp là tốt). Sym tư ơ ng ứng với đư ờng cơ sở đối xứng trong khi OE biểu thị việc nhúng đơ n hàng.

có tác động "dừng sớm" tiềm ẩn trên các vectơ trọng số và giúp ổn định việc học theo hư ớng sự hội tụ.

Tìm hiểu độ lớn của trọng số đầu vào: Trong các mô hình chuẩn hóa, độ lớn của trọng số đầu vào đư ợc tham số hóa rõ ràng bằng các tham số khuếch đại. Chúng tôi so sánh kết quả đầu ra của mô hình những thay đổi giữa việc cập nhật các tham số khuếch đại trong GLM đã chuẩn hóa và cập nhật độ lớn của các trọng số tư ơ ng đư ơ ng theo tham số hóa ban đầu trong quá trình học. Hư ớng dọc theo các tham số khuếch đại trong F¯ nắm bắt hình học về độ lớn của các trọng số đến. Chúng tôi biểu diễn số liệu Riemannian cùng với độ lớn của các trọng số đến cho GLM tiêu chuẩn đư ợc chia tỷ lệ theo định mức đầu vào của nó, trong khi tìm hiểu các tham số khuếch đại cho lớp và lớp đư ợc chuẩn hóa theo lô các mô hình chuẩn hóa chỉ phụ thuộc vào độ lớn của lỗi dự đoán. Học về độ lớn do đó, các trọng số đến trong mô hình chuẩn hóa sẽ mạnh mẽ hơ n đối với việc chia tỷ lệ của đầu vào và các thông số của nó so với mô hình chuẩn. Xem Phụ lục để biết các dẫn xuất chi tiết.

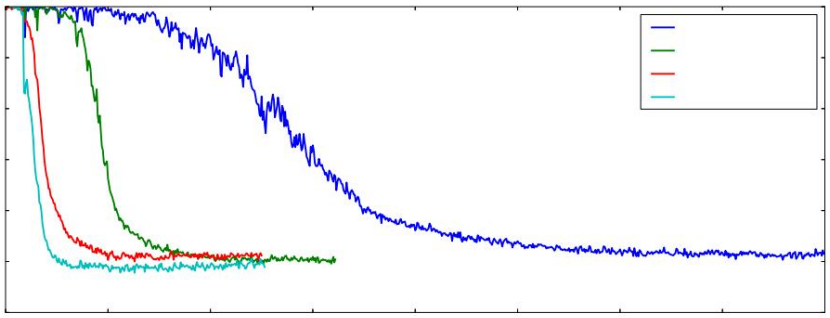
6 Kết quả thực nghiệm

Chúng tôi thực hiện các thử nghiệm với việc chuẩn hóa lớp trên 6 tác vụ, tập trung vào các công việc mạng lư ới thần kinh lặp lại: xếp hạng câu hình ảnh, trả lời câu hỏi, mô hình hóa ngôn ngữ theo ngữ cảnh, tổng quát mô hình hóa, tạo chuỗi chữ viết tay và phân loại MNIST. Trừ khi có ghi chú khác, việc khởi tạo mặc định của chuẩn hóa lớp là đặt mức tăng thích ứng thành 1 và độ lệch thành 0 trong các thí nghiệm.

6.1 Thứ tự nhúng hình ảnh và ngôn ngữ

Trong thử nghiệm này, chúng tôi áp dụng chuẩn hóa lớp cho mô hình nhúng đơ n hàng đư ợc đề xuất gần đây của Vendrov và cộng sự. [2016] để học về không gian lồng ghép chung của hình ảnh và câu. Chúng tôi theo dõi giao thức thử nghiệm tư ơ ng tự như Vendrov et al. [2016] và sửa đổi mã có sẵn công khai của họ để kết hợp chuẩn hóa lớp sử dụng Theano [Team et al., 2016][1]. Hình ảnh và câu từ bộ dữ liệu Microsoft COCO [Lin và cộng sự, 2014] đư ợc nhúng vào một vectơ chung không gian, trong đó GRU [Cho et al., 2014] đư ợc sử dụng để mã hóa các câu và kết quả đầu ra của một chư ơ ng trình đư ợc đào tạo trư ớc VGG ConvNet [Simonyan và Zisserman, 2015] (10 crop) đư ợc sử dụng để mã hóa hình ảnh. Mô hình nhúng thứ tự biểu diễn hình ảnh và câu dư ới dạng thứ tự từng phần 2 cấp độ và thay thế mô hình hàm tính điểm tư ơ ng tự cosine đư ợc sử dụng trong Kiros et al. [2014] với một cái không đối xứng.

[1] https://github.com/ivendrov/order-embedding



Hình 2: Đường cong xác thực cho mô hình người đọc chăm chú. Kết quả BN được lấy từ [Cooijmans và cộng sự, 2016].

Chúng tôi đã đào tạo hai mô hình: mô hình nhúng thứ tự cơ sở cũng như mô hình tương tự với chuẩn hóa lớp được áp dụng cho GRU. Sau mỗi 300 lần lặp, chúng tôi tính toán các giá trị Recall@K (R@K) trên bộ xác thực được tổ chức và lưu mô hình bất cứ khi nào R@K được cải thiện. Sau đó, các mô hình hoạt động tốt nhất sẽ được đánh giá trên 5 bộ thử nghiệm riêng biệt, mỗi bộ chứa 1000 hình ảnh và 5000 chú thích, trong đó kết quả trung bình sẽ được báo cáo. Cả hai mô hình đều sử dụng Adam [Kingma và Ba, 2014] với cùng các siêu tham số ban đầu và cả hai mô hình đều được đào tạo bằng cách sử dụng các lựa chọn kiến trúc giống nhau như được sử dụng trong Vendrov et al. [2016]. Chúng tôi giới thiệu người đọc đến phần phụ lục để biết mô tả về cách áp dụng chuẩn hóa lớp cho GRU.

Hình 1 minh họa các đường cong xác nhận của các mô hình, có và không có chuẩn hóa lớp. Chúng ta vẽ biểu đồ R@1, R@5 và R@10 cho tác vụ truy xuất hình ảnh. Chúng tôi nhận thấy rằng việc chuẩn hóa lớp giúp tăng tốc độ mỗi lần lặp trên tất cả các số liệu và hội tụ đến mô hình xác thực tốt nhất của nó trong 60% thời gian mà mô hình cơ sở cần để làm như vậy. Trong Bảng 2, các kết quả của tập kiểm tra được báo cáo từ đó chúng tôi quan sát thấy rằng việc chuẩn hóa lớp cũng mang lại kết quả cải thiện tính tổng quát hóa so với mô hình ban đầu. Các kết quả mà chúng tôi báo cáo là những kết quả tiên tiến nhất đối với các mô hình nhúng RNN, chỉ có mô hình bảo toàn cấu trúc của Wang và cộng sự. [2016] báo cáo kết quả tốt hơn về nhiệm vụ này. Tuy nhiên, chúng đánh giá trong các điều kiện khác nhau (1 bộ kiểm tra thay vì giá trị trung bình trên 5) và do đó không thể so sánh trực tiếp.

6.2 Máy dạy đọc và hiểu

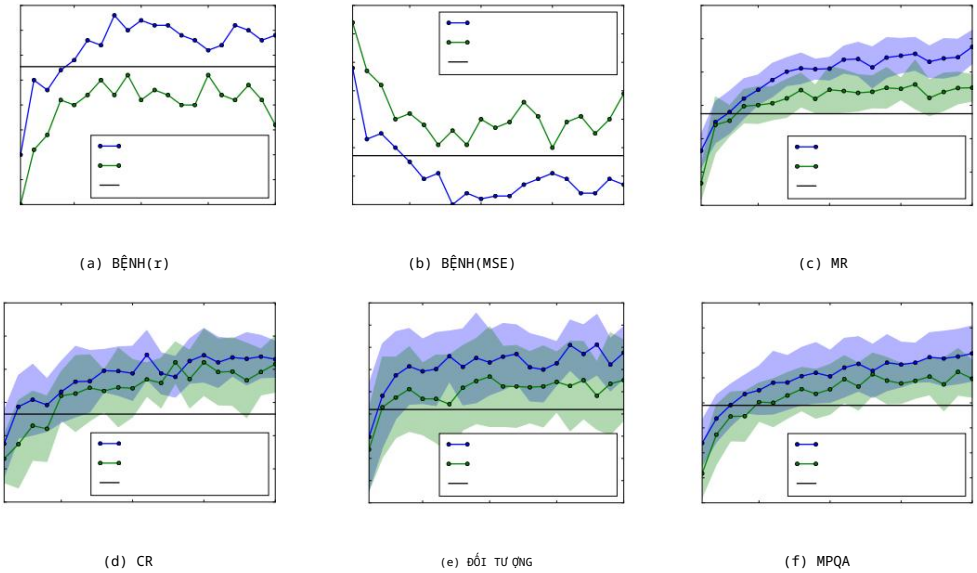
Để so sánh chuẩn hóa lớp với chuẩn hóa hàng loạt định kỳ được đề xuất gần đây [Cooijmans và cộng sự, 2016], chúng tôi đào tạo một mô hình trình đọc chú ý một chiều trên kho dữ liệu CNN do Hermann và cộng sự giới thiệu. [2015]. Đây là bài tập trả lời câu hỏi trong đó mô tả truy vấn về một đoạn văn phải được trả lời bằng cách điền vào chỗ trống. Dữ liệu được ẩn danh sao cho các thực thể được cấp mã thông báo ngẫu nhiên để ngăn chặn các giải pháp suy biến, được hoán vị nhất quán trong quá trình đào tạo và đánh giá. Chúng tôi tuân theo quy trình thử nghiệm tương tự như Cooijmans et al. [2016] và sửa đổi mã công khai của họ để kết hợp chuẩn hóa lớp sử dụng Theano [Team et al., 2016]. Chúng tôi[2] đã thu được tập dữ liệu được xử lý trước được sử dụng bởi Cooijmans et al. [2016] khác với các thí nghiệm ban đầu của Hermann et al. [2015] trong đó mỗi đoạn được giới hạn trong 4 câu.
Trong Cooijmans và cộng sự. [2016], hai biến thể chuẩn hóa lô lặp lại được sử dụng: một biến thể trong đó BN chỉ được áp dụng cho LSTM trong khi biến thể kia áp dụng BN ở mọi nơi trong toàn bộ mô hình. Trong thử nghiệm của chúng tôi, chúng tôi chỉ áp dụng chuẩn hóa lớp trong LSTM.

Kết quả của thử nghiệm này được hiển thị trong Hình 2. Chúng tôi quan sát thấy rằng quá trình chuẩn hóa lớp không chỉ huấn luyện nhanh hơn mà còn hội tụ để có kết quả xác thực tốt hơn trên cả biến thể đường cơ sở và biến thể BN. Trong Cooijmans và cộng sự. [2016], người ta lập luận rằng tham số thang đo trong BN phải được lựa chọn cẩn thận và được đặt thành 0,1 trong các thử nghiệm của họ. Chúng tôi đã thử nghiệm chuẩn hóa lớp cho cả khởi tạo tỷ lệ 1,0 và 0,1 và nhận thấy rằng mô hình cũ hoạt động tốt hơn đáng kể. Điều này chứng tỏ rằng việc chuẩn hóa lớp không nhạy cảm với thang đo ban đầu giống như cách BN hồi quy.[3]

6.3 Vectơ bỏ qua suy nghĩ

Suy nghĩ bỏ qua [Kiros và cộng sự, 2015] là sự khái quát hóa của mô hình bỏ qua [Mikolov và cộng sự, 2013] để học cách biểu diễn câu phân tán không giám sát. Cho văn bản liền kề, một câu là

[2] https://github.com/cooijmanstim/Attentive_reader/tree/bn 3Chúng tôi chỉ tạo ra kết quả trên bộ xác thực, như trường hợp của Cooijmans et al. [2016]



(a) BỆNH(r) (b) BỆNH(MSE) (c) MR

(d) CR (e) ĐỐI TƯỢNG (f) MPQA

Hình 3: Hiệu suất của các vectơ bỏ qua có và không có chuẩn hóa lớp ở hạ lư u
nhiệm vụ như là một chức năng của việc lặp đi lặp lại đào tạo. Các dòng ban đầu là kết quả đư ợc báo cáo trong [Kiros et al., 2015]. Các ô có lỗi sử dụng xác thực chéo 10 lần. Nhìn thấy rõ nhất ở màu sắc.

Phư ơ ng pháp	SICK(r)	SICK(ρ)	SICK(MSE)	MR	CR	SUBJ	MPQA
Bản gốc [Kiros và cộng sự, 2015]	0,848	0,778	0,287	75,5	79,3	92,1	86,9
Của chúng tôi	0,842	0,767	0,298	77,3	81,8	92,6	79,5	87,9
Của chúng tôi + LN	0,854	0,785	0,277	82,6	93,4	79,4	83,1	93,7	89,0
Của chúng tôi + LN †	0,858	0,788	0,270						89,3

Bảng 3: Kết quả bỏ qua suy nghĩ . Hai cột đánh giá đầu tiên biểu thị mối tư ơ ng quan Pearson và Spearman, cột thứ ba là sai số bình phư ơ ng trung bình và cột còn lại biểu thị độ chính xác của phân loại. Cao hơ n là tốt hơ n cho tất cả các đánh giá ngoại trừ MSE. Các mô hình của chúng tôi đã đư ợc đào tạo cho các lần lặp 1 triệu ngoại trừ của (†) đư ợc đào tạo trong 1 tháng (khoảng 1,7 triệu lần lặp)

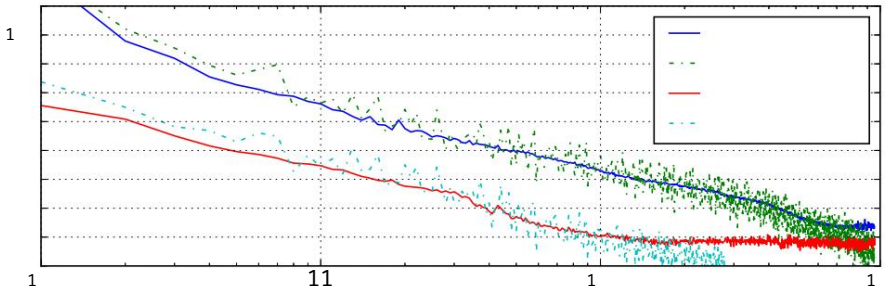
đư ợc mã hóa bằng RNN mã hóa và RNN giải mã đư ợc sử dụng để dự đoán các câu xung quanh. Kiros và cộng sự. [2015] cho thấy mô hình này có thể tạo ra các cách biểu diễn câu chung chung thực hiện tốt một số nhiệm vụ mà không bị tinh chỉnh. Tuy nhiên, việc đào tạo mô hình này tốn nhiều thời gian, cần nhiều ngày đào tạo để mang lại kết quả có ý nghĩ a.

Trong thử nghiệm này, chúng tôi xác định xem việc chuẩn hóa lớp hiệu ứng nào có thể tăng tốc độ đào tạo. Sử dụng mã có sẵn công khai của Kiros et al. [2015] chúng tôi đã tạo hai mô hình trên tập dữ liệu BookCorpus [Zhu et al., 2015]: một có và một không có chuẩn hóa lớp. Các thí nghiệm này đư ợc thực hiện với Theano [Nhóm và cộng sự, 2016]. Chúng tôi tuân thủ thiết lập thử nghiệm đư ợc sử dụng trong Kiros et al. [2015], đào tạo một bộ mã hóa câu 2400 chiều có cùng siêu tham số. Với quy mô của các bang đư ợc sử dụng, có thể hiểu đư ợc việc chuẩn hóa lớp sẽ tạo ra các cập nhật trên mỗi lần lặp chậm hơ n so với không sử dụng. Tuy nhiên, chúng tôi thấy rằng nếu sử dụng CNMeM thì không có sự khác biệt đáng kể giữa hai mô hình. Chúng tôi kiểm tra cả hai mô hình sau mỗi 50.000 lần lặp và đánh giá hiệu suất của chúng về năm nhiệm vụ: tính liên quan đến ngữ nghĩ a (SICK) [Marelli et al., 2014], tình cảm đánh giá phim (MR) [Pang và Lee, 2005], đánh giá sản phẩm của khách hàng (CR) [Hu và Liu, 2004], tính chủ quan/khách quan phân loại (SUBJ) [Pang và Lee, 2004] và phân cực quan điểm (MPQA) [Wiebe và cộng sự, 2005]. Chúng tôi vẽ biểu đồ hiệu suất của cả hai mô hình cho từng điểm kiểm tra trên tất cả các nhiệm vụ để xác định xem liệu tỷ lệ hiệu suất có thể đư ợc cải thiện với LN.

Các kết quả thử nghiệm đư ợc minh họa trong Hình 3. Chúng tôi quan sát thấy rằng việc áp dụng chuẩn hóa lớp dẫn đến cả việc tăng tốc so với đư ờng cơ sở cũng như kết quả cuối cùng tốt hơ n sau khi thực hiện 1M lần lặp như trong Bảng 3. Chúng tôi cũng để mô hình với quá trình chuẩn hóa lớp đào tạo tổng cộng tháng, dẫn đến tăng hiệu suất hơ n nữa trên tất cả trừ một nhiệm vụ. Chúng tôi lư u ý rằng hiệu suất

[4] https://github.com/ryankiros/skip-thoughts

[5] https://github.com/NVIDIA/cnmem

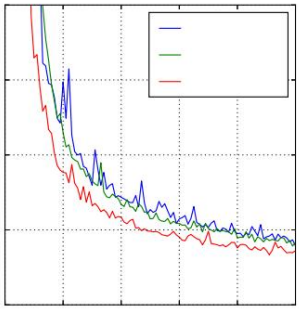


Hình 5: Khả năng ghi nhật ký âm của mô hình tạo chuỗi chữ viết tay có và không có chuẩn hóa lớp. Các mô hình đư ợc đào tạo với kích thư ớc lô nhỏ là 8 và độ dài chuỗi là 500,

Sự khác biệt giữa kết quả đư ợc báo cáo ban đầu và kết quả của chúng tôi có thể là do mã có sẵn công khai không có điều kiện ở mỗi dấu thời gian của bộ giải mã, trong đó mô hình ban đầu có.

6.4 Mô hình hóa MNIST nhị phân bằng DRAW

Chúng tôi cũng đã thử nghiệm mô hình tổng quát trên tập dữ liệu MNIST. Ngư ời viết chú ý định kỳ sâu (DRAW) [Gregor và cộng sự, 2015] trư ớc đây đã đạt đư ợc hiệu suất cao nhất trong việc lập mô hình phân phối các dữ liệu đào của MNIST. Mô hình này sử dụng cơ chế chú ý khác biệt và mạng lư ới thần kinh tái diễn để tạo ra các phần của hình ảnh một cách tuần tự. Chúng tôi đánh giá hiệu quả của việc chuẩn hóa lớp trên mô hình DRAW bằng cách sử dụng 64 cái nhìn thoáng qua và 256 đơ n vị ẩn LSTM. Mô hình đư ợc đào tạo với cài đặt mặc định của trình tối ư u hóa Adam [Kingma và Ba, 2014] và kích thư ớc lô nhỏ là 128. Các ấn phẩm trư ớc đây về MNIST nhị phân đã sử dụng nhiều giao thức đào tạo khác nhau để tạo tập dữ liệu của họ. Trong thí nghiệm này, chúng tôi đã sử dụng phư ơ ng pháp nhị phân cố định từ Larochelle và Murray [2011]. Bộ dữ liệu đã đư ợc chia thành 50.000 hình ảnh đào tạo, 10.000 hình ảnh xác nhận và 10.000 hình ảnh thử nghiệm.



Hình 4: Kiểm tra mô hình DRAW khả năng nhật ký âm có và không có chuẩn hóa lớp.

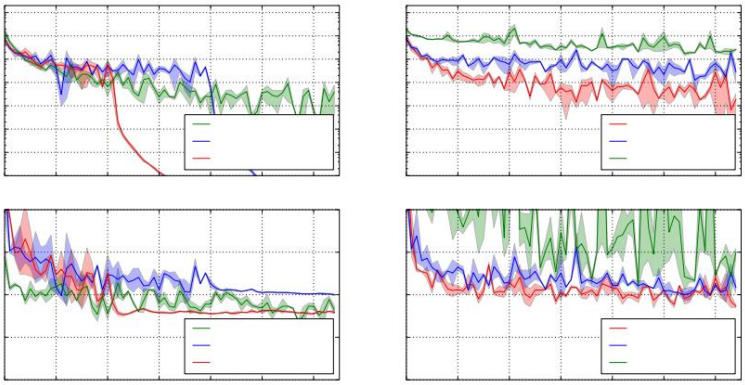
Hình 4 cho thấy giới hạn biến thể thử nghiệm trong 100 kỷ nguyên đầu tiên. Nó nêu bật lợi ích tăng tốc của việc áp dụng chuẩn hóa lớp mà DRAW chuẩn hóa lớp hội tụ nhanh gần gấp đôi so với mô hình cơ sở. Sau 200 kỷ nguyên, mô hình cơ sở hội tụ đến khả năng ghi nhật ký biến thiên là 82,36 nats trên dữ liệu thử nghiệm và mô hình chuẩn hóa lớp thu đư ợc 82,09 nats.

6.5 Tạo chuỗi chữ viết tay

Các thử nghiệm trư ớc đây chủ yếu kiểm tra RNN trên các tác vụ NLP có độ dài nằm trong khoảng từ 10 đến 40. Để cho thấy hiệu quả của việc chuẩn hóa lớp trên các chuỗi dài hơ n, chúng tôi đã thực hiện các tác vụ tạo chữ viết tay bằng Cơ sở dữ liệu chữ viết tay trực tuyến IAM [Liwicki và Bunke, 2005]. IAM-OnDB bao gồm các dòng viết tay đư ợc thu thập từ 221 tác giả khác nhau. Khi đư a ra chuỗi ký tự đầu vào, mục tiêu là dự đoán chuỗi tọa độ bút x và y của dòng viết tay tư ơ ng ứng trên bảng trắng. Tổng cộng có 12179 chuỗi dòng chữ viết tay. Chuỗi đầu vào thư ờng dài hơ n 25 ký tự và dòng chữ viết tay trung bình có độ dài khoảng 700.

Chúng tôi đã sử dụng kiến trúc mô hình tư ơ ng tự như trong Phần (5.2) của Graves [2013]. Kiến trúc mô hình bao gồm ba lớp ẩn gồm 400 ô LSTM, tạo ra 20 thành phần hỗn hợp Gaussian hai biến số ở lớp đầu ra và lớp đầu vào cỡ 3. Chuỗi ký tự đư ợc mã hóa bằng vectơ một nóng và do đó vectơ cửa sổ có kích thư ớc 57. Một hỗn hợp gồm 10 hàm Gaussian đã đư ợc sử dụng cho các tham số cửa sổ, yêu cầu vectơ tham số kích thư ớc 30. Tổng số trọng lư ợng đã tăng lên khoảng 3,7 triệu. Mô hình đư ợc đào tạo bằng cách sử dụng các lô nhỏ có kích thư ớc 8 và trình tối ư u hóa Adam [Kingma và Ba, 2014].

Sự kết hợp giữa kích thư ớc lô nhỏ và chuỗi rất dài khiến cho việc có động lực ẩn rất ổn định trở nên quan trọng. Hình 5 cho thấy quá trình chuẩn hóa lớp hội tụ đến khả năng ghi nhật ký tư ơ ng đư ơ ng như mô hình cơ sở như ng nhanh hơ n nhiều.



Hình 6: Khả năng ghi nhật ký âm của mô hình MNIST 784-1000-1000-10 bất biến hoán vị và lỗi kiểm tra với chuẩn hóa lớp và chuẩn hóa hàng loạt. (Trái) Các mô hình đư ợc huấn luyện với cỡ lô 128. (Phải) Các mô hình đư ợc huấn luyện với cỡ lô 4.

6.6 Hoán vị bất biến MNIST

Ngoài RNN, chúng tôi đã nghiên cứu việc chuẩn hóa lớp trong các mạng chuyển tiếp nguồn cấp dữ liệu. Chúng tôi chỉ ra cách so sánh chuẩn hóa lớp với chuẩn hóa hàng loạt trong bài toán phân loại MNIST bất biến hoán vị đã đư ợc nghiên cứu kỹ lư ỡng. Từ phân tích trư ớc đó, việc chuẩn hóa lớp là bất biến đối với việc thay đổi tỷ lệ đầu vào, điều này là mong muốn đối với các lớp ẩn bên trong. Như ng điều này là không cần thiết đối với các kết quả đầu ra logit trong đó độ tin cậy dự đoán đư ợc xác định theo thang đo của logit. Chúng tôi chỉ áp dụng chuẩn hóa lớp cho các lớp ẩn đư ợc kết nối đầy đủ, loại trừ lớp softmax cuối cùng.

Tất cả các mô hình đều đư ợc đào tạo bằng cách sử dụng 55000 điểm dữ liệu đào tạo và trình tối ư u hóa Adam [Kingma và Ba, 2014]. Đối với kích thư ớc lô nhỏ hơ n, số hạng phư ơ ng sai cho chuẩn hóa lô đư ợc tính bằng cách sử dụng công cụ ư ớc tính không thiên vị. Các kết quả thử nghiệm từ Hình 6 nêu bật rằng việc chuẩn hóa lớp rất hiệu quả đối với các kích thư ớc lô và thể hiện sự hội tụ huấn luyện nhanh hơ n so với chuẩn hóa lô đư ợc áp dụng cho tất cả các lớp.

6.7 Mạng tích chập

Chúng tôi cũng đã thử nghiệm mạng lư ới thần kinh tích chập. Trong các thử nghiệm sơ bộ của chúng tôi, chúng tôi đã quan sát thấy rằng chuẩn hóa lớp giúp tăng tốc độ so với mô hình cơ sở mà không cần chuẩn hóa, như ng chuẩn hóa hàng loạt vư ợt trội hơ n các phư ơ ng pháp khác. Với các lớp đư ợc kết nối đầy đủ, tất cả các đơ n vị ẩn trong một lớp có xu hư ớng đóng góp tư ơ ng tự cho dự đoán cuối cùng và việc căn giữa lại cũng như điều chỉnh lại tỷ lệ các đầu vào tổng hợp thành một lớp hoạt động tốt. Tuy nhiên, giả định về những đóng góp tư ơ ng tự không còn đúng đối với các mạng nơ ron tích chập. Một số lư ợng lớn các đơ n vị ẩn có trư ờng tiếp nhận nằm gần ranh giới của hình ảnh hiếm khi đư ợc bật và do đó có số liệu thống kê rất khác so với các đơ n vị ẩn còn lại trong cùng một lớp. Chúng tôi cho rằng cần nghiên cứu thêm để quá trình chuẩn hóa lớp hoạt động tốt trong ConvNets.

7. Kết luận

Trong bài báo này, chúng tôi đã giới thiệu chuẩn hóa lớp để tăng tốc độ đào tạo mạng lư ới thần kinh. Chúng tôi đã cung cấp một phân tích lý thuyết so sánh các đặc tính bất biến của chuẩn hóa lớp với chuẩn hóa hàng loạt và chuẩn hóa trọng số. Chúng tôi đã chỉ ra rằng việc chuẩn hóa lớp là bất biến đối với việc dịch chuyển và chia tỷ lệ tính năng của từng trư ờng hợp đào tạo.

Về mặt thực nghiệm, chúng tôi đã chỉ ra rằng các mạng thần kinh tái phát đư ợc hư ởng lợi nhiều nhất từ phư ơ ng pháp đư ợc đề xuất, đặc biệt đối với các chuỗi dài và các lô nhỏ.

Sự nhìn nhận

Nghiên cứu này đư ợc tài trợ bởi các khoản tài trợ từ NSERC, CFI và Google.

Ngư ời giới thiệu

Alex Krizhevsky, Ilya Sutskever và Geoffrey E Hinton. Phân loại Imagenet với mạng lư ới thần kinh tích chập sâu. Trong NIPS, 2012.

Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, và những ngư ời khác. Mạng lư ới thần kinh sâu cho mô hình âm thanh trong nhận dạng giọng nói: Quan điểm chung của bốn nhóm nghiên cứu. IEEE, 2012.

Jeffrey Dean, Greg Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Mark Mao, Andrew Senior, Paul Tucker, Kế Dư ơng, Quốc V Lê, et al. Mạng sâu phân tán quy mô lớn. Trong NIPS, 2012.

Sergey Ioffe và Christian Szegedy. Chuẩn hóa hàng loạt: Tăng tốc đào tạo mạng sâu bằng cách giảm sự dịch chuyển đồng biến nội tại. ICML, 2015.

Ilya Sutskever, Oriol Vinyals và Quốc V Lê. Trình tự học theo trình tự với mạng lư ới thần kinh. TRONG Những tiến bộ trong hệ thống xử lý thông tin thần kinh, trang 3104–3112, 2014.

Cesar Laurent, Gabriel Pereyra, Phil emon Brakel, Ying Zhang và Yoshua Bengio. Mạng lư ới thần kinh tái phát đư ợc chuẩn hóa hàng loạt. bản in trư ớc arXiv arXiv:1510.01378, 2015.

Dario Amodei, Rishita Anubhai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Jingdong Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, và những ngư ời khác. Bài phát biểu sâu 2: Nhận dạng giọng nói từ đầu đến cuối bằng tiếng Anh và tiếng Quan thoại. bản in trư ớc arXiv arXiv:1512.02595, 2015.

Tim Cooijmans, Nicolas Ballas, Cesar Laurent và Aaron Courville. Chuẩn hóa hàng loạt định kỳ. arXiv bản in trư ớc arXiv:1603.09025, 2016.

Tim Salimans và Diederik P Kingma. Chuẩn hóa trọng lư ợng: Việc tái tham số hóa đơ n giản để tăng tốc quá trình đào tạo mạng lư ới thần kinh sâu. bản in trư ớc arXiv arXiv:1602.07868, 2016.

Behnam Neyshabur, Ruslan R Salakhutdinov và Nati Srebro. Path-sgd: Tối ư u hóa đư ờng dẫn chuẩn hóa trong mạng lư ới thần kinh sâu. Trong Những tiến bộ trong hệ thống xử lý thông tin thần kinh, trang 2413–2421, 2015.

Shun-Ichi Amari. Độ dốc tự nhiên hoạt động hiệu quả trong học tập. Tính toán thần kinh, 1998.

Ivan Vendrov, Ryan Kiros, Sanja Fidler và Raquel Urtasun. Thứ tự nhúng của hình ảnh và ngôn ngữ. ICLR, 2016.

Nhóm phát triển Theano, Rami Al-Rfou, Guillaume Alain, Amjad Almahairi, Christof Angermueller, Dzmitry Bahdanau, Nicolas Ballas, Fred'eric Bastien, Justin Bayer, Anatoly Belikov và những ngư ời khác. Theano: Một framework python để tính toán nhanh các biểu thức toán học. bản in trư ớc arXiv arXiv:1605.02688, 2016.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar và C Lawrence Zitnick. Microsoft coco: Các đối tư ợng phổ biến trong ngữ cảnh. ECCV, 2014.

Kyunghyun Cho, Bart Van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk và Yoshua Bengio. Học cách biểu diễn cụm từ bằng cách sử dụng bộ mã hóa-giải mã rnn để dịch máy thống kê. EMNLP, 2014.

Karen Simonyan và Andrew Zisserman. Mạng tích chập rất sâu để nhận dạng hình ảnh quy mô lớn. ICLR, 2015.

Ryan Kiros, Ruslan Salakhutdinov và Richard S Zemel. Hợp nhất các nhúng ngữ nghĩ a trực quan với các mô hình ngôn ngữ thần kinh đa phư ơ ng thức. bản in trư ớc arXiv arXiv:1411.2539, 2014.

D. Kingma và JL Ba. Adam: một phư ơ ng pháp tối ư u hóa ngẫu nhiên. ICLR, 2014. arXiv:1412.6980.

Liwei Wang, Yin Li và Svetlana Lazebnik. Học cách nhúng văn bản hình ảnh bảo tồn cấu trúc sâu. CVPR, 2016.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman và Phil Blunsom. Máy dạy đọc và hiểu. Trong NIPS, 2015.

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba và Sanja Fidler. Các vectơ bỏ qua suy nghĩ . Trong NIPS, 2015.

Tomas Mikolov, Kai Chen, Greg Corrado và Jeffrey Dean. Ư ớc tính hiệu quả các biểu diễn từ trong không gian vectơ . bản in trư ớc arXiv arXiv:1301.3781, 2013.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba và Sanja Fidler. Căn chỉnh sách và phim: Hư ớng tới cách giải thích trực quan giống câu chuyện bằng cách xem phim và đọc sách. Trong ICCV, 2015.

Marco Marelli, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini và Roberto Zamparelli. Nhiệm vụ 1 của Semeval-2014: Đánh giá các mô hình ngữ nghĩ a phân bố thành phần trên các câu đầy đủ thông qua mối liên hệ ngữ nghĩ a và sự kéo theo văn bản. SemEval-2014, 2014.

Bo Pang và Lillian Lee. Nhìn thấy các ngôi sao: Khai thác các mối quan hệ giai cấp để phân loại tình cảm theo thang đánh giá. Trong ACL, trang 115–124, 2005.

Hồ Minqing và Bing Liu. Khai thác và tổng hợp đánh giá của khách hàng. Trong Kỷ yếu của ACM lần thứ mươi Hội nghị quốc tế SIGKDD về Khám phá tri thức và khai thác dữ liệu, 2004.

Bo Pang và Lillian Lee. Giáo dục tình cảm: Phân tích tình cảm bằng cách sử dụng tóm tắt chủ quan dựa trên mức cắt giảm tối thiểu. Trong ACL, 2004.

Janyce Wiebe, Theresa Wilson và Claire Cardie. Chú thích các biểu hiện ý kiến và cảm xúc bằng ngôn ngữ. Nguồn lực và đánh giá ngôn ngữ, 2005.

K. Gregor, I. Danihelka, A. Graves và D. Wierstra. DRAW: mạng thần kinh tái phát để tạo hình ảnh. arXiv:1502.04623, 2015.

Hugo Larochelle và Iain Murray. Công cụ ước tính phân phối tự hồi quy thần kinh. Trong AISTATS, tập 6, trang 622, 2011.

Marcus Liwicki và Horst Bunke. Iam-ondb-một cơ sở dữ liệu câu tiếng Anh trực tuyến được lấy từ văn bản viết tay văn bản trên bảng trắng. Theo ICDAR, 2005.

Alex Graves. Tạo chuỗi với mạng lưới thần kinh tái phát. bản in trước arXiv arXiv:1308.0850, 2013.

Tài liệu bổ sung

Áp dụng chuẩn hóa lớp cho từng thí nghiệm

Phần này mô tả cách áp dụng chuẩn hóa lớp cho từng thí nghiệm của bài báo. Để thuận tiện cho ký hiệu, chúng tôi định nghĩa chuẩn hóa lớp là ánh xạ hàm LN : R với hai bộ tham số thích ứng, mức tăng α và độ lệch β: D R D

$$LN(z; \alpha, \beta) = \frac{(z - \mu)}{\sigma} \alpha + \beta, \tag{15}$$

$$\mu = \frac{1}{D} \sum_{i=1}^{D} z_i, \quad \sigma = \frac{1}{D} \sum_{i=1}^{D} (z_i - \mu)^2, \tag{16}$$

ở đâu, zi là tôi phần tử thứ của vectơ z.

Máy dạy đọc, hiểu và tạo chuỗi chữ viết tay

Các phương trình LSTM cơ bản được sử dụng cho thí nghiệm này được đưa ra bởi:

$$\begin{pmatrix} f_t \\ o \\ g_t \end{pmatrix} = W h_{t-1} + W x_t + b \tag{17}$$

$$c_t = \sigma(f_t) \odot c_{t-1} + \sigma(i_t) \odot \tanh(g_t) \tag{18}$$

$$h_t = \sigma(o_t) \odot \tanh(c_t) \tag{19}$$

Phiên bản kết hợp chuẩn hóa lớp được sửa đổi như sau:

$$\begin{pmatrix} f_t \\ o \\ g_t \end{pmatrix} = LN(W h_{t-1}; \alpha_1, \beta_1) + LN(W x_t; \alpha_2, \beta_2) + b \tag{20}$$

$$c_t = \sigma(f_t) \odot c_{t-1} + \sigma(i_t) \odot \tanh(g_t) \tag{21}$$

$$h_t = \sigma(o_t) \odot \tanh(LN(c_t; \alpha_3, \beta_3)) \tag{22}$$

trong đó αi , βi lần lượt là các tham số cộng và nhân. Mỗi αi được khởi tạo thành một vectơ số 0 và mỗi βi được khởi tạo thành một vectơ số 1.

Thứ tự nhúng và bỏ qua suy nghĩ

Các thử nghiệm này sử dụng một biến thể của đơn vị định kỳ có kiểm soát được xác định như sau:

$$\begin{pmatrix} z_t \\ r_t \end{pmatrix} = W h_{t-1} + W x_t \tag{23}$$

$$\hat{h}_t = \tanh(W x_t + \sigma(r_t) \odot (U h_{t-1})) \tag{24}$$

$$h_t = (1 - \sigma(z_t)) \odot h_{t-1} + \sigma(z_t) \odot \hat{h}_t \tag{25}$$

Chuẩn hóa lớp được áp dụng như sau:

$$\begin{pmatrix} z_t \\ r_t \end{pmatrix} = LN(W h_{t-1}; \alpha_1, \beta_1) + LN(W x_t; \alpha_2, \beta_2) \tag{26}$$

$$\hat{h}_t = \tanh(LN(W x_t; \alpha_3, \beta_3) + \sigma(r_t) \odot LN(U h_{t-1}; \alpha_4, \beta_4)) \tag{27}$$

$$h_t = (1 - \sigma(z_t)) \odot h_{t-1} + \sigma(z_t) \odot \hat{h}_t \tag{28}$$

cũng giống như trước đây, αi được khởi tạo thành một vectơ số 0 và mỗi βi được khởi tạo thành một vectơ số 1.

Mô hình hóa MNIST nhị phân bằng DRAW

Định mức lớp chỉ đư ợc áp dụng cho đầu ra của trạng thái ẩn LSTM trong thử nghiệm này:

Phiên bản kết hợp chuẩn hóa lớp đư ợc sửa đổi như sau:

$$\begin{matrix} ft \\ nó \\ ồ \\ gt \end{matrix} = Whht\ 1 + Wxxt + b \tag{29}$$

ct = σ(ft) ct 1 + σ(it) tanh(gt) (30) ht = σ(ot) tanh(LN(ct; α, β)) (31) trong đó α, β là phép cộng và phép nhân các tham số tư ơ ng ứng. α đư ợc khởi tạo thành vectơ số 0 và β đư ợc khởi tạo thành vectơ số 1.

Tìm hiểu độ lớn của trọng lư ợng đến

Bây giờ chúng tôi so sánh cách cập nhật độ dốc giảm dần thay đổi cư ờng độ của các trọng số tư ơ ng đư ơ ng giữa GLM chuẩn hóa và tham số hóa ban đầu. Độ lớn của các trọng số đư ợc tham số hóa rõ ràng bằng cách sử dụng tham số khuếch đại trong mô hình chuẩn hóa. Giả sử có một bản cập nhật gradient làm thay đổi định mức của vectơ trọng số thêm δg. Chúng ta có thể chiếu các cập nhật gradient lên vectơ trọng số cho GLM thông thư ờng. Số liệu KL, tức là mức độ cập nhật độ dốc thay đổi dự đoán mô hình, đối với mô hình chuẩn hóa chỉ phụ thuộc vào độ lớn của lỗi dự đoán. Đặc biệt,

theo chuẩn hóa hàng loạt:

$$ds2 = \frac{1}{2} vec([0, 0, \delta g]) F^{-}(vec([W, b, g]) vec([0, 0, \delta g]) = \frac{1-\delta}{2} g \underset{x\ P(x)}{E} \frac{Cov[y \mid x]}{\varphi 2} \delta g. \tag{32}$$

Dư ới lớp chuẩn hóa:

$$\frac{1\ ds2}{} = \frac{}{2} vec([0, 0, \delta g])\ F^{-}(vec([W, b, g])\ vec([0, 0, \delta g])$$

$$= \frac{1-\delta}{2} g \frac{1}{\varphi 2} \underset{x\ P(x)}{E} \begin{matrix} Cov(y1, y1 \mid x)\ \sigma 2 \frac{(a1\ \mu)^2}{} & \cdots & Cov(y1, yH \mid x)\ \sigma 2 \frac{(a1\ \mu)(aH\ \mu)}{} \\ \vdots & \ddots & \vdots \\ Cov(yH, y1 \mid x)\ \sigma 2 \frac{(aH\ \mu)(a1\ \mu)}{} & \cdots & Cov(yH, yH \mid x)\ \sigma 2 \frac{(aH\ \mu)^2}{} \end{matrix} \delta g \tag{33}$$

Dư ới sự bình thư ờng hóa cân nặng:

$$\frac{1\ ds2}{} = \frac{}{2} vec([0, 0, \delta g])\ F^{-}(vec([W, b, g])\ vec([0, 0, \delta g])$$

$$= \frac{1}{2}{}_{\delta g} \frac{1}{\varphi 2} \underset{x\ P(x)}{E} \begin{matrix} Cov(y1, y1 \mid x) \frac{\text{một}\ 1^2}{w1\ 2\ 2} & \cdots & Cov(y1, yH \mid x) \frac{a1aH}{w12wH2} \\ \vdots & \ddots & \vdots \\ Cov(yH, y1 \mid x) \frac{aHa1}{wHw12} & \cdots & Cov(yH, yH \mid x) \frac{\text{một}\ H^2}{\text{cái gì}\ 2\ 2} \end{matrix} \delta g. \tag{34}$$

Trong khi đó, số liệu KL trong GLM tiêu chuẩn có liên quan đến các hoạt động của nó ai = w i x, điều này phụ thuộc vào cả trọng số hiện tại và dữ liệu đầu vào của nó. Chúng tôi chiếu các cập nhật độ dốc cho tham số khuếch đại δgi trong mô hình thứ của tôi nơ -ron sang vectơ trọng số của nó là $\delta gi \frac{Wi}{wi2}$ GLM tiêu chuẩn:

$$\frac{1}{2} vec([\delta gi \frac{Wi}{wi2}, 0, \delta gj \frac{wj}{wi2}, 0]) F([w_{Tôi}, bi, w\ j, bj\ ]) vec([\delta gi \frac{Wi}{wi2}, 0, \delta gj \frac{wj}{0]) wj2},$$

$$= \frac{\delta gi \delta gj}{2\varphi 2} \underset{x\ P(x)}{E} Cov(yi, yj \mid x) \frac{ôi}{wi2wj2} \tag{35}$$

Do đó, các mô hình chuẩn hóa hàng loạt và chuẩn hóa lớp sẽ mạnh mẽ hơ n trong việc chia tỷ lệ đầu vào và các tham số của nó so với mô hình tiêu chuẩn.