Google Developers

Image Captioning

Let’s build something

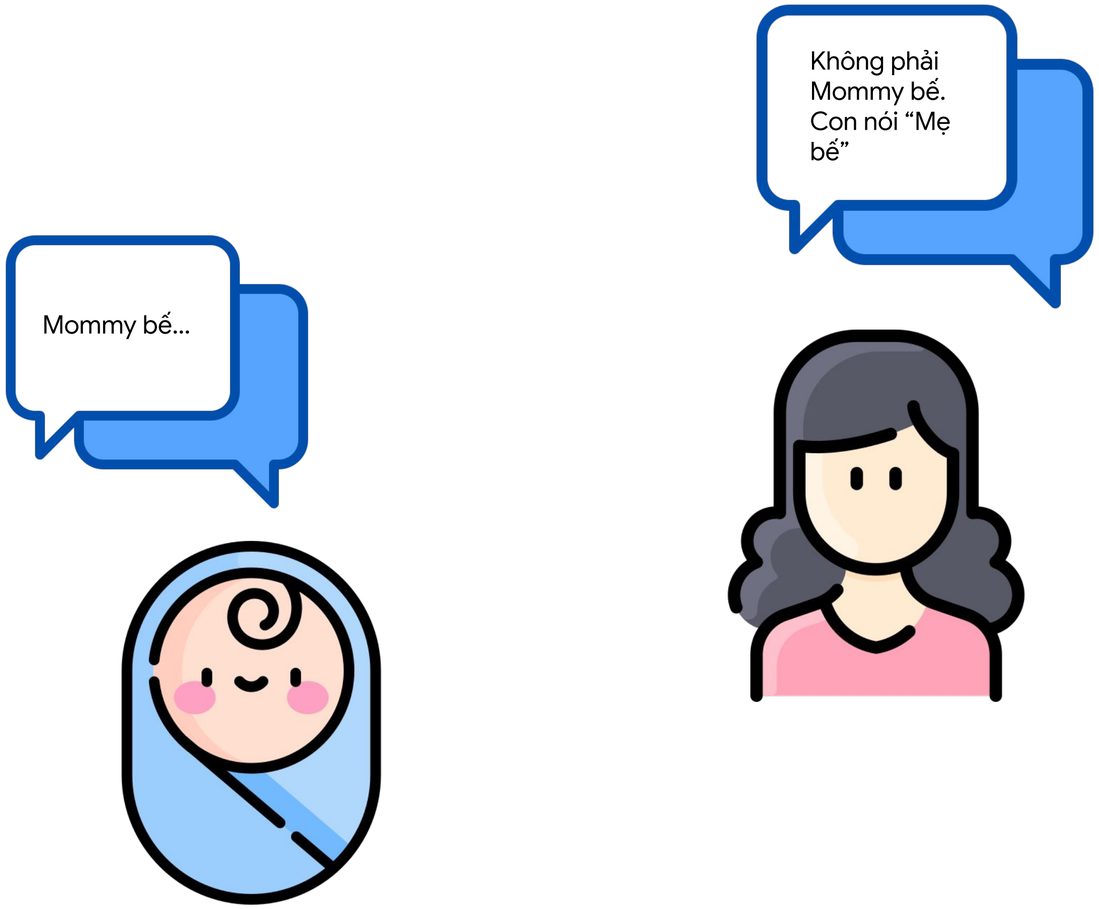
Ngoc Ba
@ProtonX @VietAI

Châm ngôn mang về

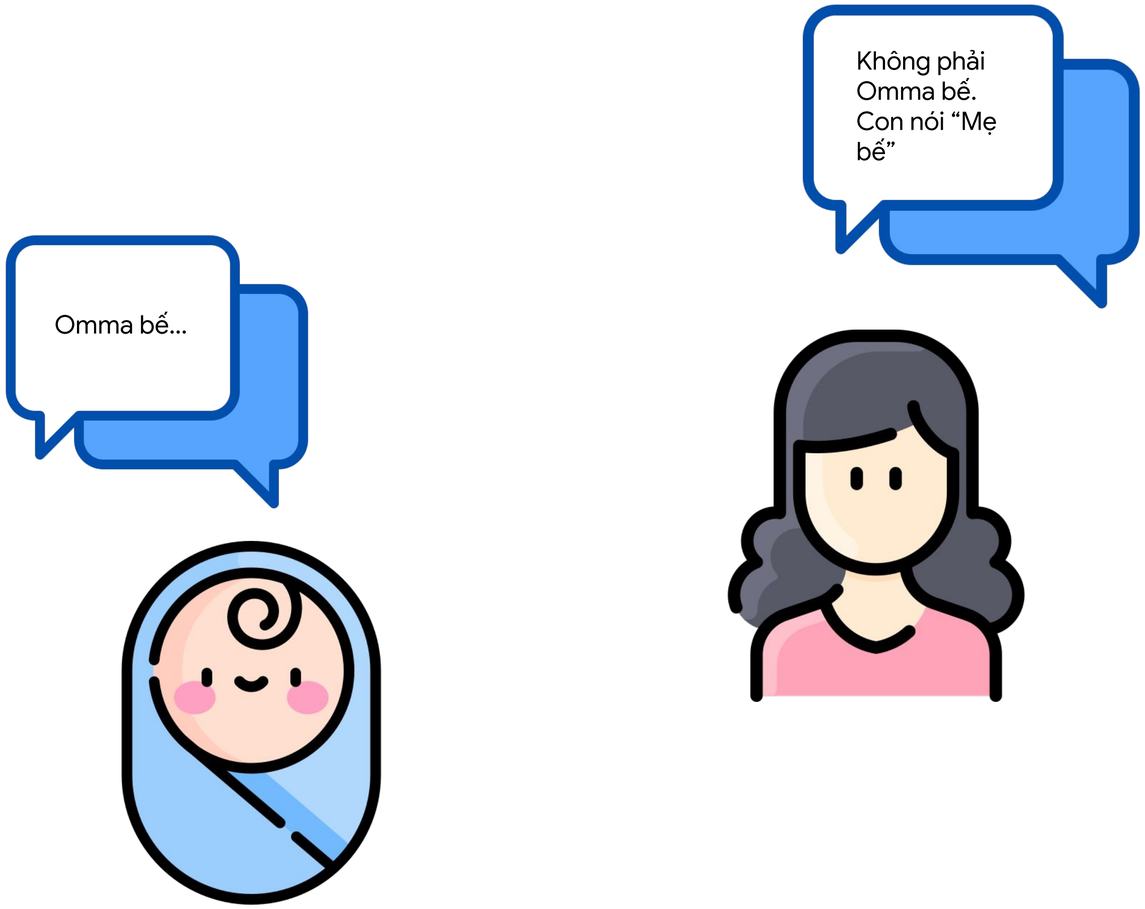
Càng **tự nhiên** bao nhiêu
Càng **hiệu quả** bấy nhiêu

Dữ liệu là chìa khoá

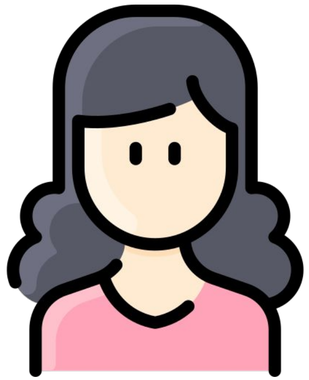
Dạy máy học như dạy trẻ con học



Dạy máy học như dạy trẻ con học



Dạy máy học như dạy trẻ con học

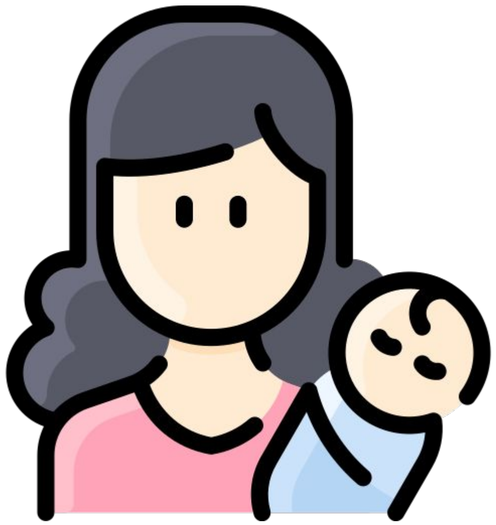


Máy học là quá trình học vô cùng tự nhiên, thử, sai và chỉnh lại.

Học cách nói “mẹ bế”

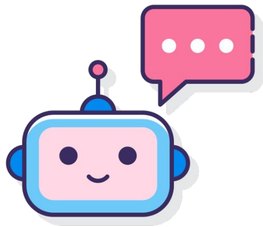
Con muốn	Con nói	Mẹ chỉnh
Con muốn Mẹ bế	Mommy bế...	Mẹ bế
Con muốn Mẹ bế	Omma bế...	Mẹ bế
Con muốn Mẹ bế	Mẹ bế...	Mẹ bế

Dạy máy học như dạy trẻ con học



>> 1.000.000 điểm dữ liệu (data point)

Thực tế chúng ta có thể mô phỏng quá trình này để đào tạo (train) cho một máy tính học được quy trình này



Quá trình training

Features	Prediction	Label	
Con muốn Mẹ bế	Mommy bế...	Mẹ bế	Sai
Con muốn Mẹ bế	Omma bế...	Mẹ bế	Sai
Con muốn Mẹ bế	Mẹ bế...	Mẹ bế	Đúng
...	...	...	
...	...	...	
...	...	...	

Chúng ta có thể **mô phỏng được sự sai lệch này**, và **giảm thiểu nó** để làm cho **dự đoán sau tốt hơn**

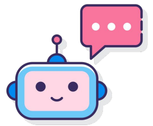
Mô hình (Model)

Features	Prediction	Label
Con muốn Mẹ bế	Mommy bế...	Mẹ bế
Con muốn Mẹ bế	Omma bế...	Mẹ bế
Con muốn Mẹ bế	Mẹ bế...	Mẹ bế
...	...	...
...	...	...
...	...	...

>> 1.000.000 điểm dữ liệu (data point)

Features
Con muốn Mẹ bế
Con muốn Mẹ bế
Con muốn Mẹ bế
...
...
...

Mô hình (Model)



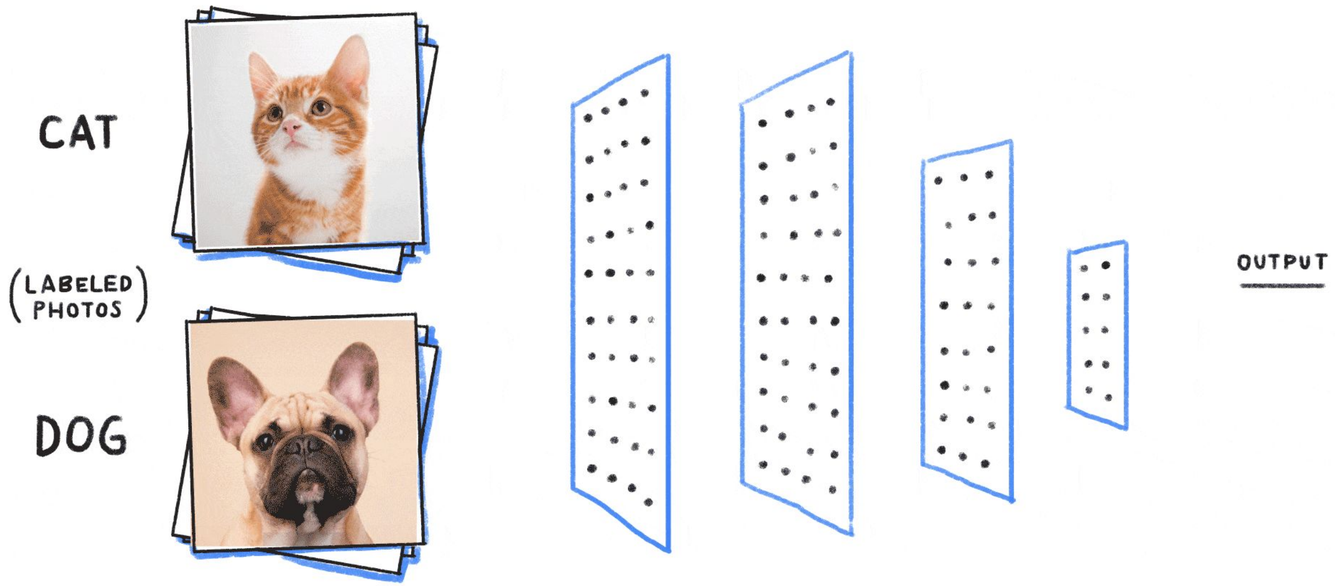
Prediction
Mommy bế...
Omma bế...
Mẹ bế...
...
...
...

Hàm mất mát (Cost Function)

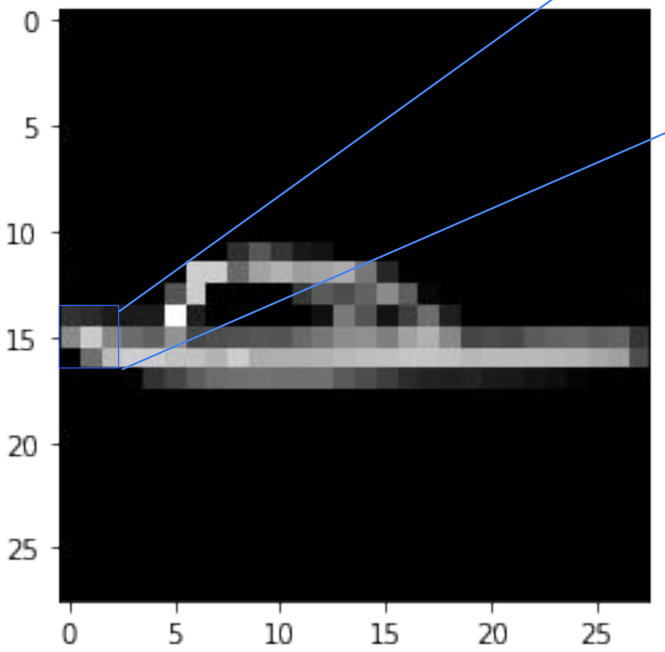
Label
Mẹ bế
Mẹ bế
Mẹ bế
...
...
...

Đọc hình ảnh

Từ một hình ảnh, mô hình sẽ trích xuất ra thông tin quan trọng.

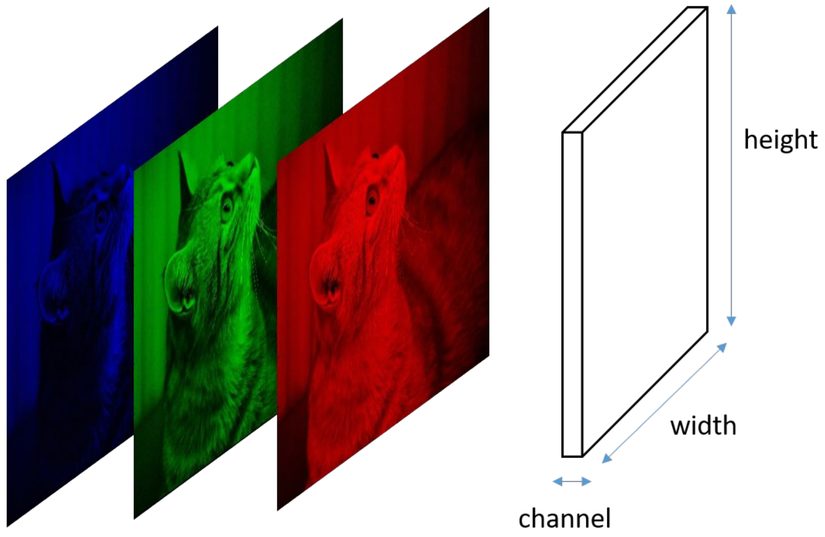


Ảnh



51	43	27
137	202	133
0	110	186

Ảnh được cấu thành từ các pixel có giá trị trong khoảng **[0, 255]**

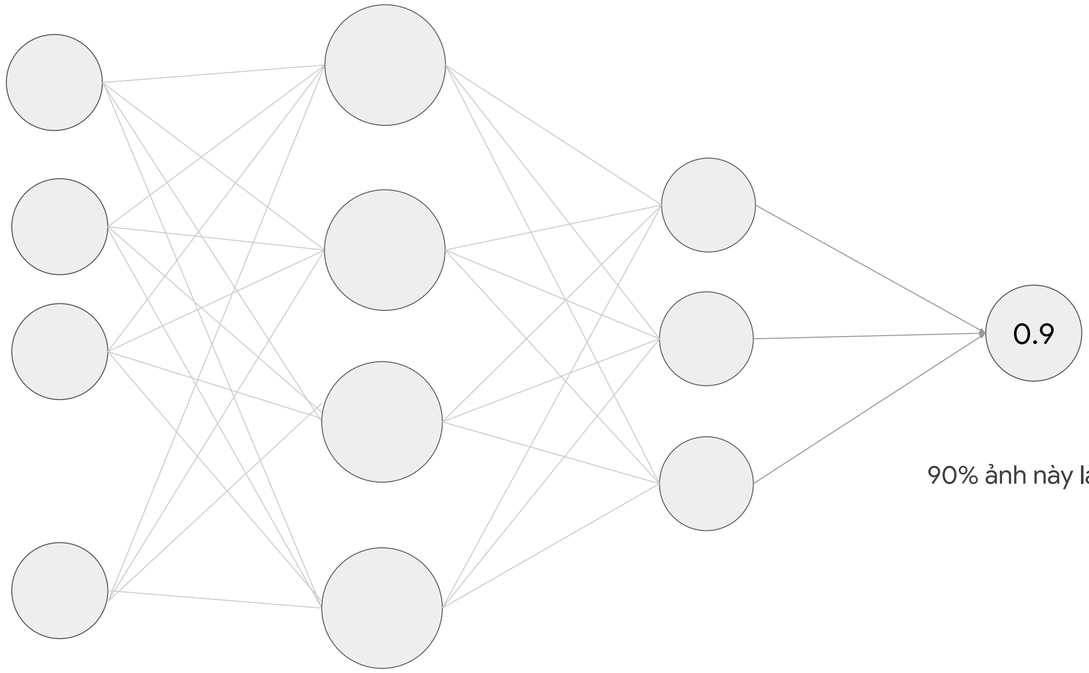


Ảnh màu có 3 channels: **xanh dương**, **xanh lục**, **đỏ**.
Mỗi channel là một ma trận 2 chiều.

Mô hình phân loại ảnh (Image Classification)

Duỗi ảnh

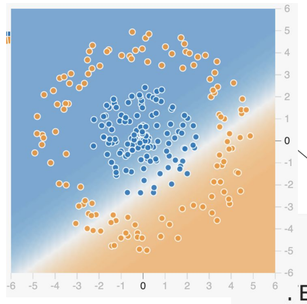
Mạng Nơ ron

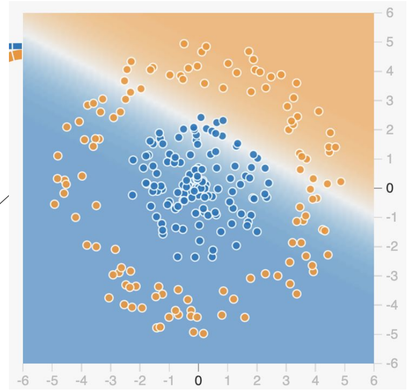


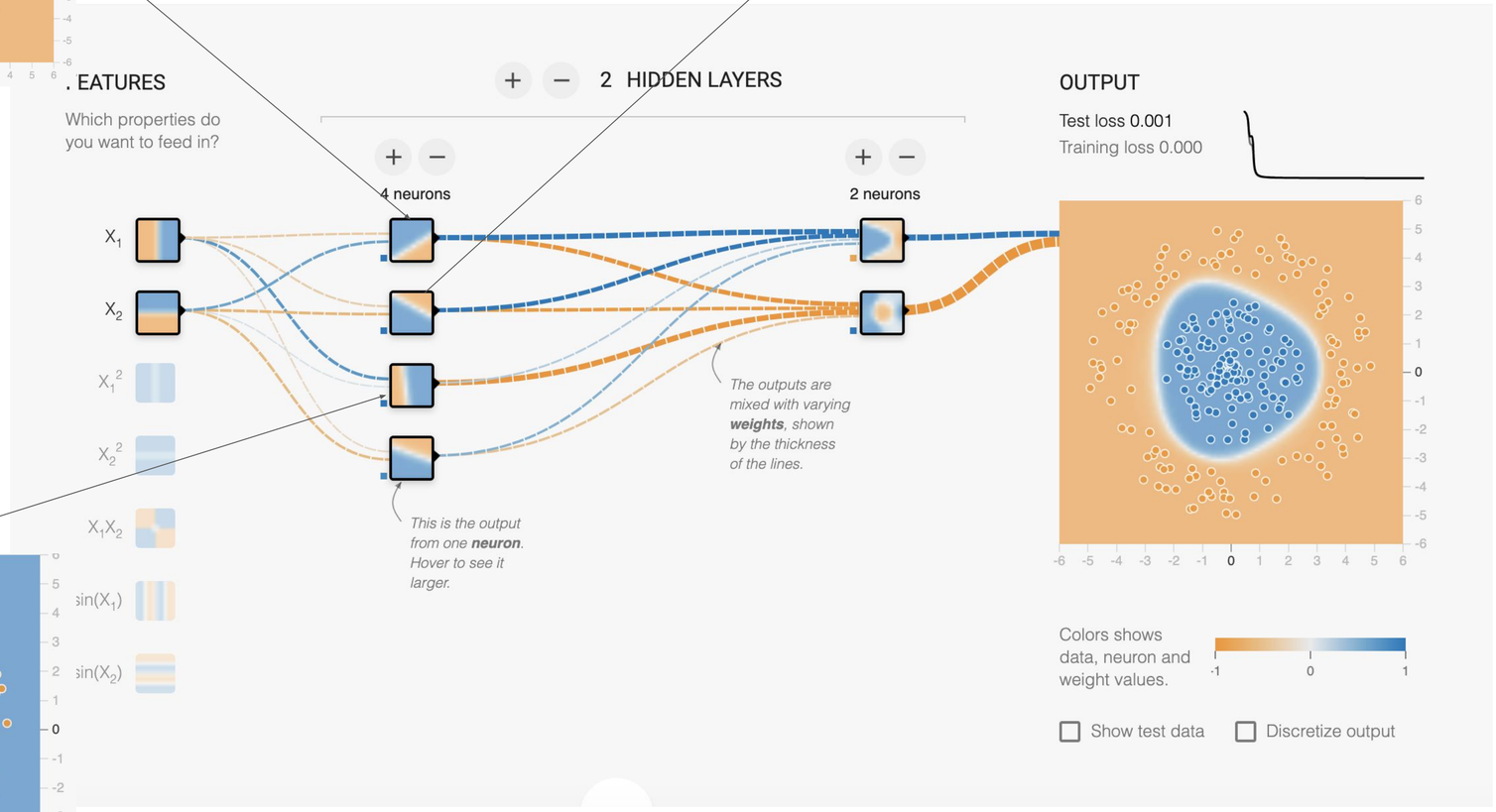
90% ảnh này là **con chó**

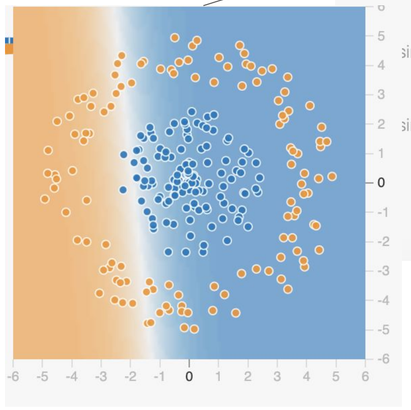
Các phép biến đổi tuyến tính/phi tuyến để trích xuất thông tin hữu ích cho mục đích phân loại

Mạng nơ ron





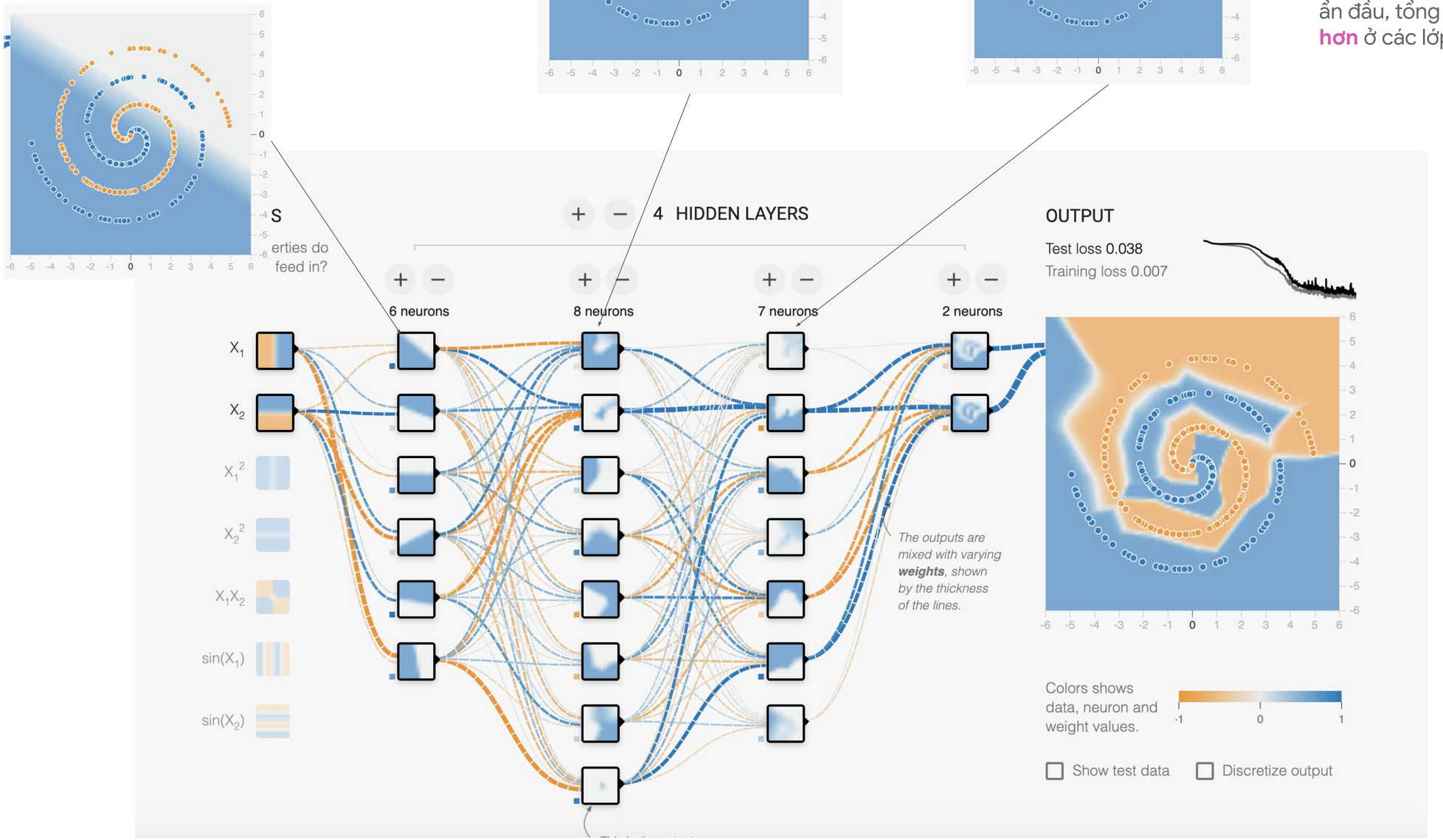




Các nơ ron ở tầng ẩn **thực hiện khám phá** một feature cụ thể

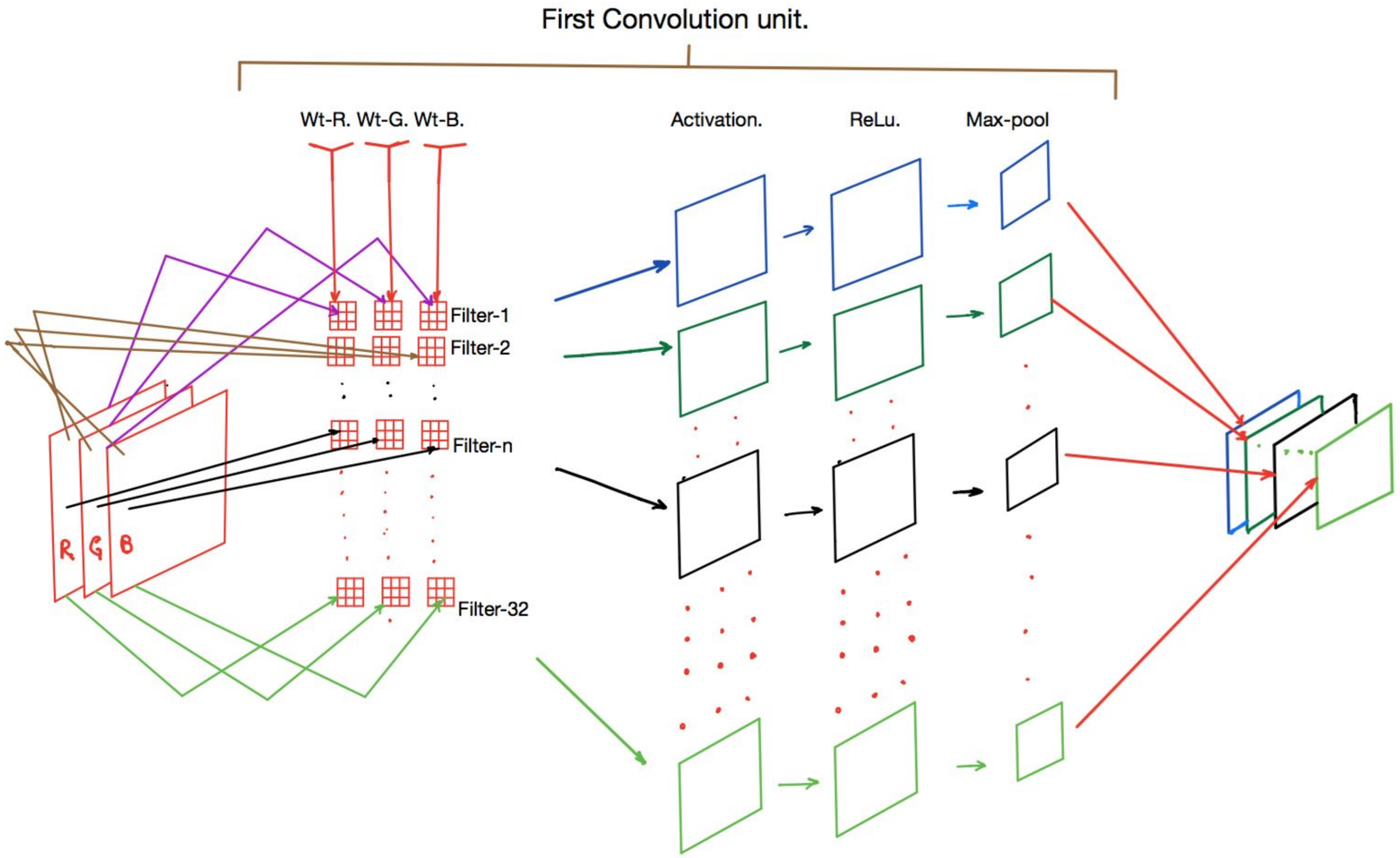
[Tensorflow Playground Demo]

Mạng nơ ron



Sự khám phá này **đơn giản** ở các lớp ẩn đầu, tổng hợp lại và **phức tạp hơn** ở các lớp ẩn tiếp theo

Học cách xử lý ảnh (Mạng CNN)



Convolutional Neural Networks (LeCun, 1989) là một loại **Neural Network đặc biệt** để xử lý dữ liệu dạng lưới (grid-like topology), ví dụ là ảnh.

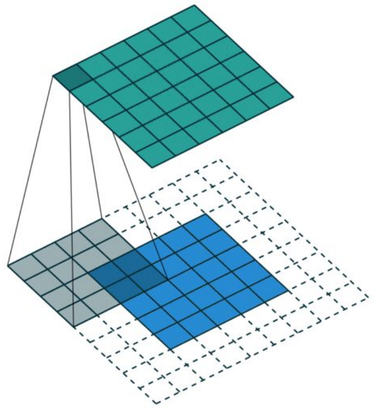
Ảnh thường có thể coi là lưới 2D các pixels

Yann LeCun
Turing Award

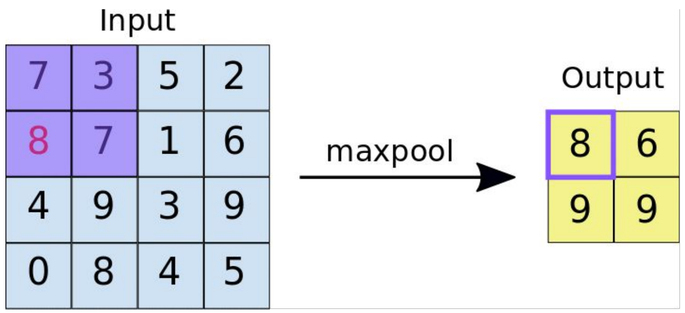
Mạng CNN được bổ sung thêm khả năng học xử lý ảnh sử dụng các lớp **Convolution** + **Pooling**

https://medium.com/@apiltamang/a-gentle-dive-into-the-anatomy-of-a-convolution-layer-6f1024339aca

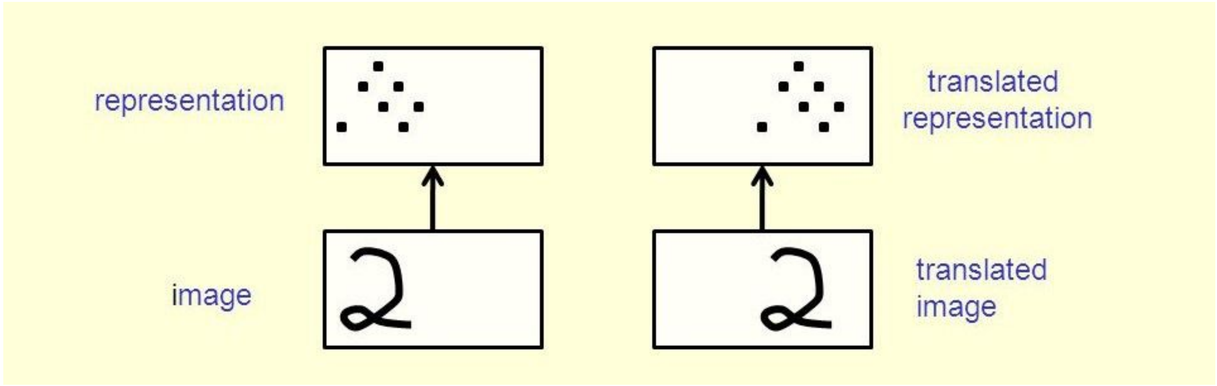
Lớp Convolution và Pooling



Lớp **Convolution**



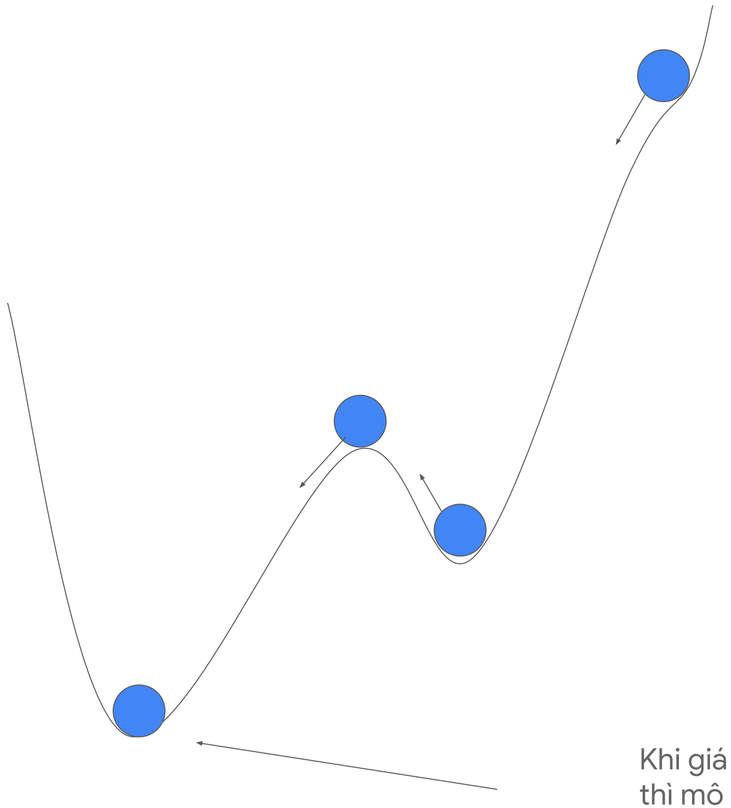
Lớp **Pooling**



Tại sao phải sử dụng 2 lớp này ???

Tạo ra **tính kháng dịch chuyển** cho mô hình, cùng một đối tượng trong không gian nhưng đặt ở vị trí khác nhau, mạng CNN vẫn có thể trích xuất được đúng.

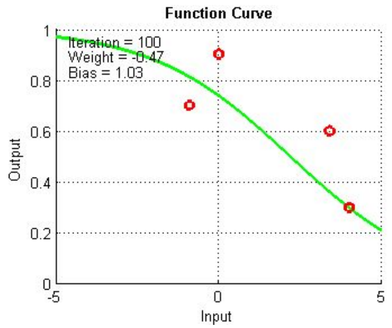
Hàm mất mát (Cost Function)

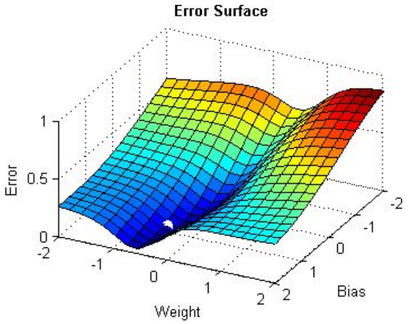


Khi giá trị **sai lệch nhỏ**
thì mô hình **càng chính xác**

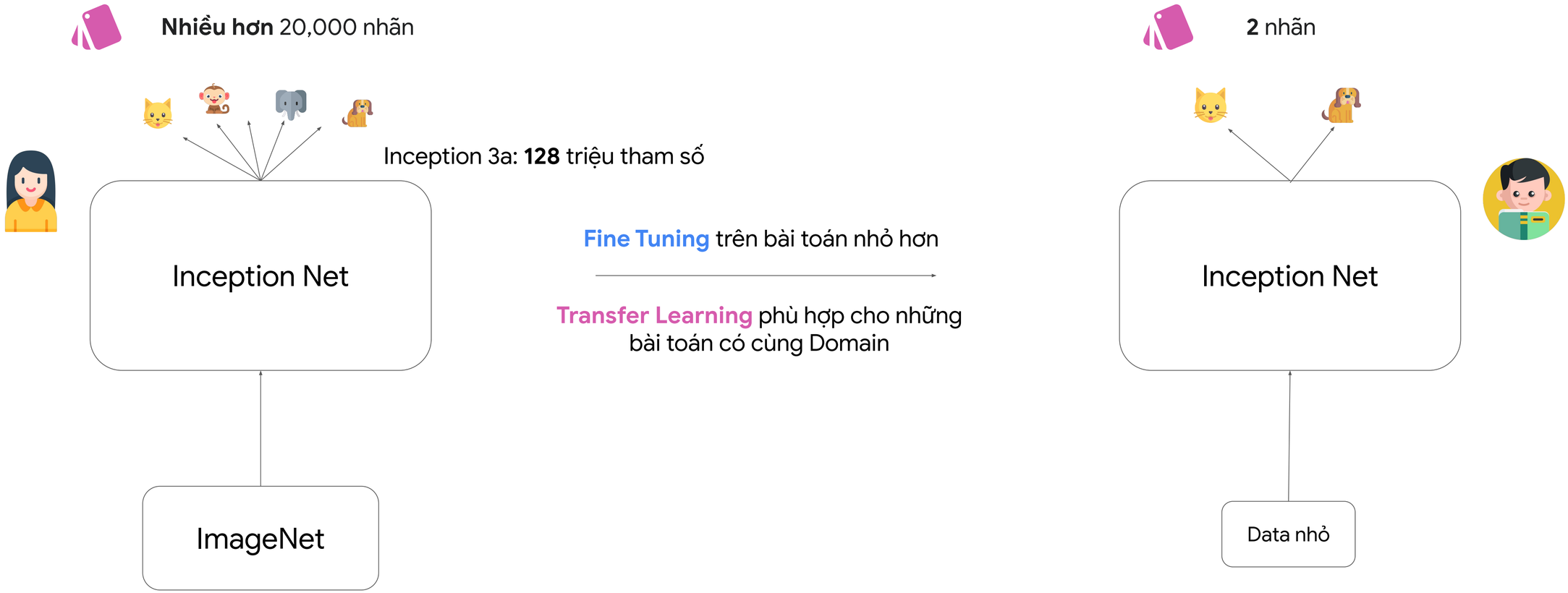
Các thuật toán tối ưu

- Gradient Descent
- Stochastic Gradient Descent
- Adadelta
- RMSprop
- Adam
- AdaMax
- Nadam
- AMSGrad





Transfer Learning trong thị giác máy tính

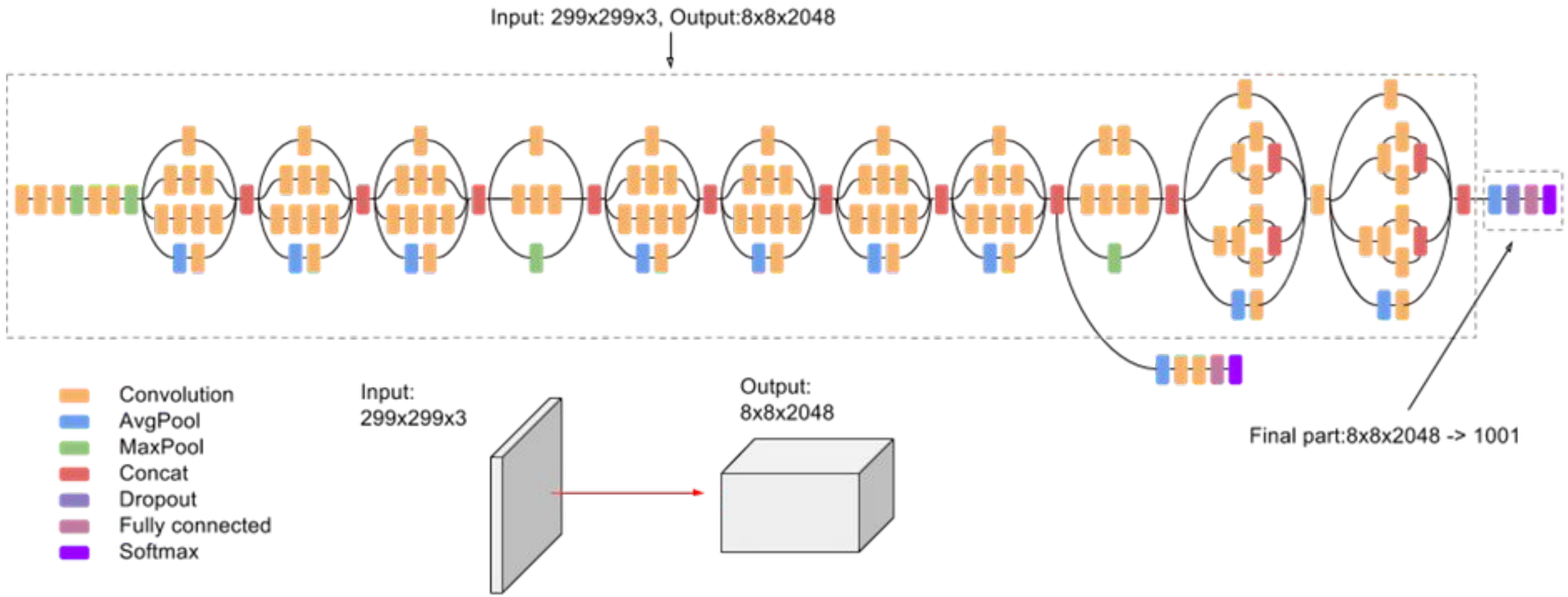


https://arxiv.org/pdf/1409.4842.pdf

Link bài thực hành

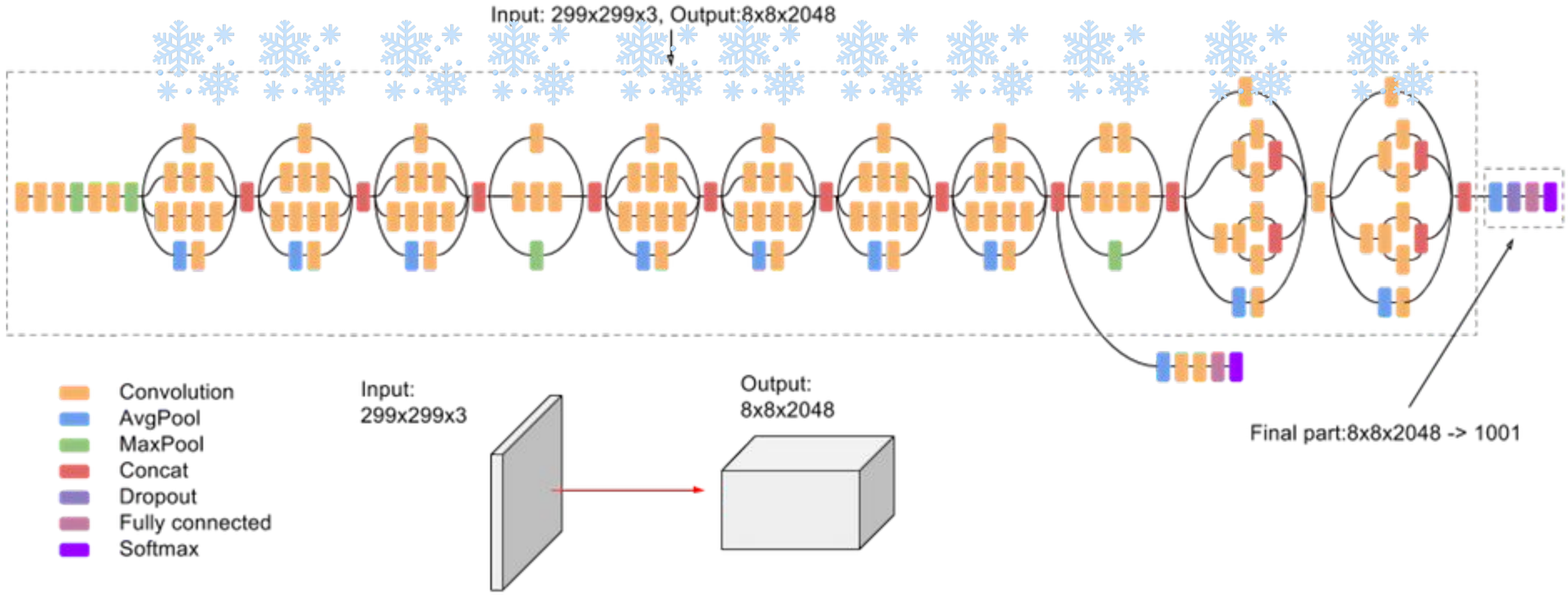
Transfer Learning đang rất gần

Inception V3



https://arxiv.org/pdf/1409.4842.pdf

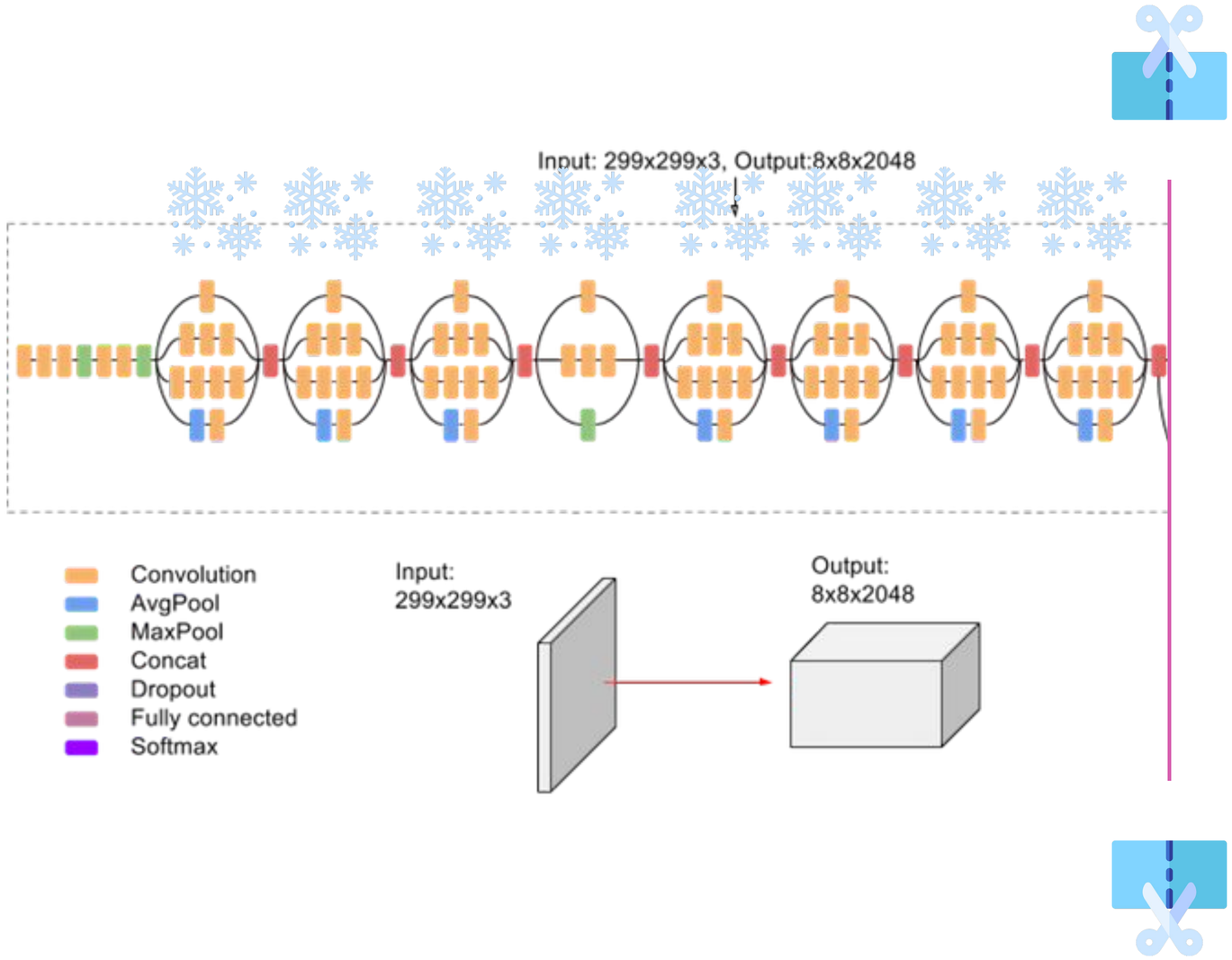
Đóng băng tham số



https://arxiv.org/pdf/1409.4842.pdf

Đóng băng và sử dụng mô hình như một hàm

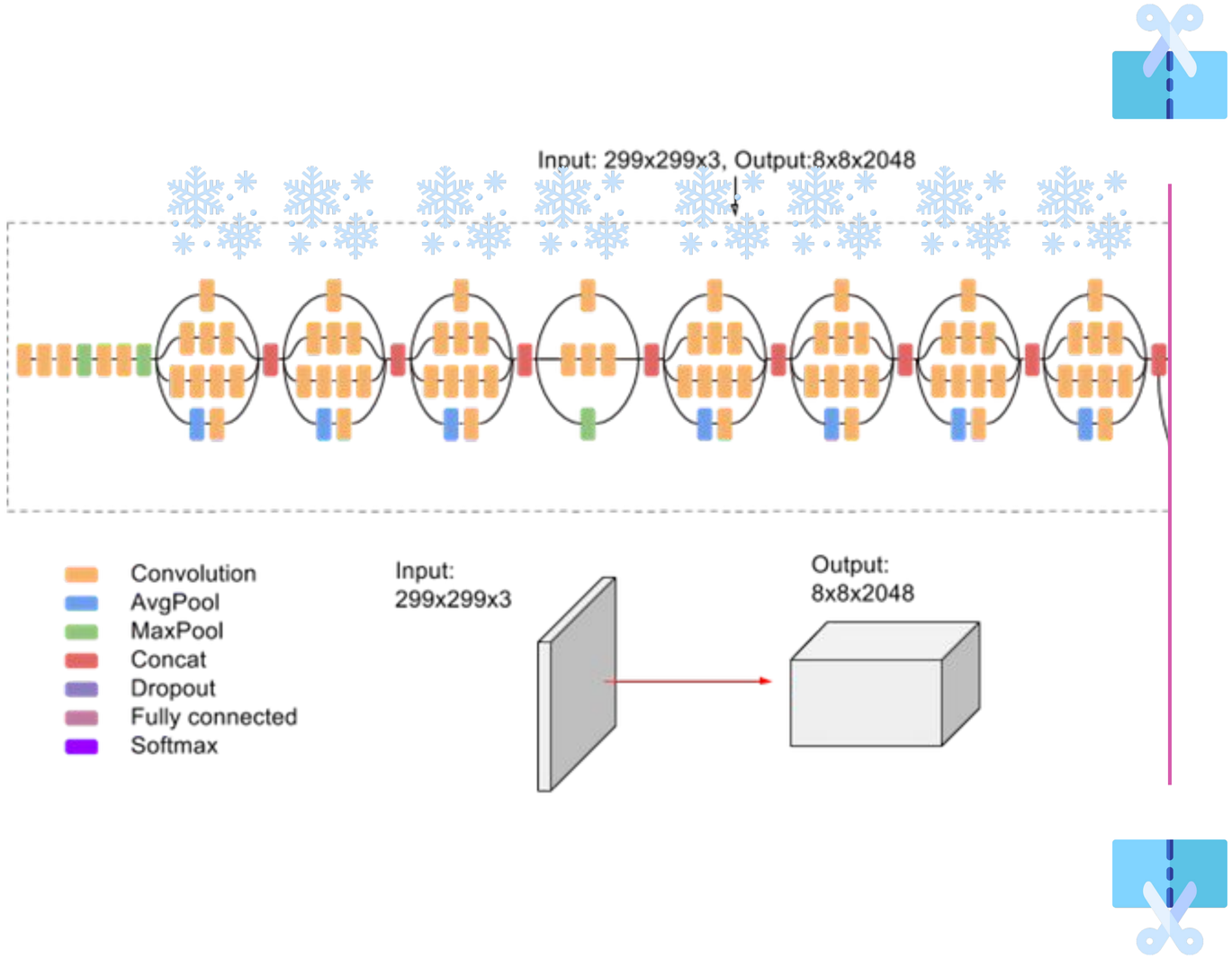
Thay đổi lớp phân loại



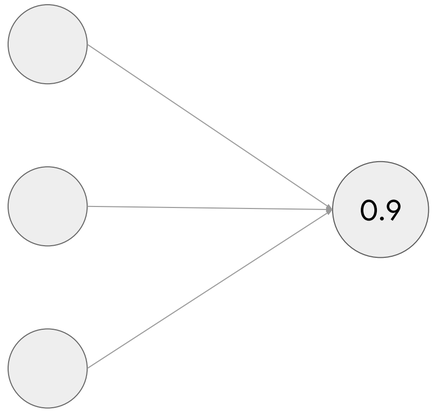
https://arxiv.org/pdf/1409.4842.pdf

Đóng băng và sử dụng mô hình như một hàm

Thêm lớp phân loại



NgocNet



https://arxiv.org/pdf/1409.4842.pdf

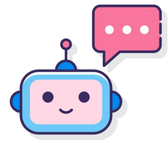
Miêu tả hình ảnh (Image Captioning)

Features

Label

A group of bikers driving down a curvy road.

Làm sao để mô phỏng được mô hình này?



Bộ dữ liệu

MS COCO

82783 hình ảnh + miêu tả

https://cocodataset.org/#captions-2015

Experts

Đọc dữ liệu (Loading Data)

Thực hành đọc dữ liệu

5000 hình ảnh + miêu tả

```
print(train_captions[0])
Image.open(img_name_vector[0])
```

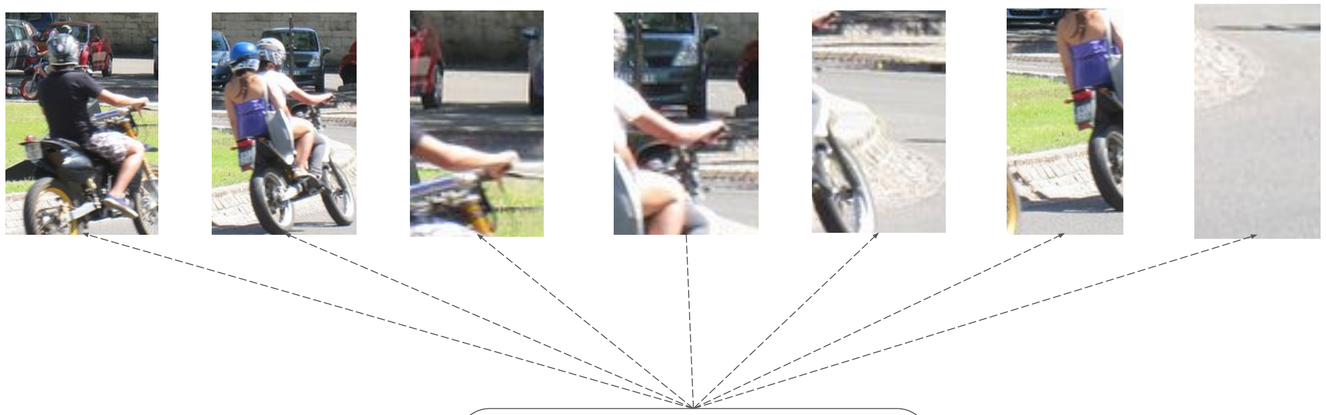
```
<start> A doughnut with a giant doughnut over it's face. <end>
```

Trong thực tế có thể gặp trường hợp nhãn chưa miêu tả hoàn toàn đúng

Cách khắc phục: Sử dụng **Label Smoothing**.

Link: https://arxiv.org/pdf/1906.02629.pdf

Trích xuất đặc trưng



(..., 64, 2048)

CNN Model

Trích xuất ra thông tin quan trọng trên bức ảnh bằng việc đưa ảnh qua các pretrained model. E.g. Inception

Inception Net

Attention is all you need

https://arxiv.org/abs/1706.03762

Tách từ

Câu cho trước

```
sentences = [
  'I am Vietnamese',
  'Vietnamese people are pretty friendly',
]
```

Từ điển

<unk>	1
'vietnamese'	2
'i'	3
'am'	4
'people'	5
'are'	6
'pretty'	7
'friendly'	8

Thứ tự càng nhỏ số lần xuất hiện trong các câu càng lớn.

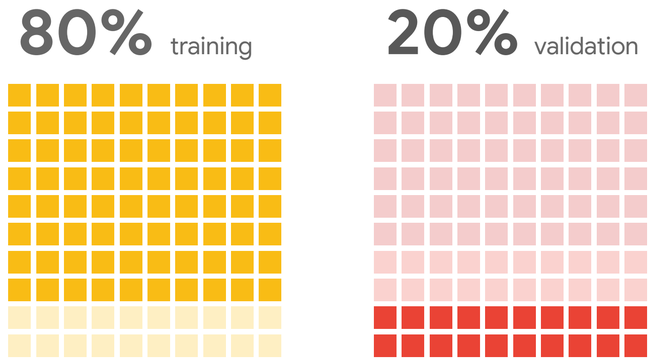
Câu mới

I	am	pretty

`texts_to_sequences` →

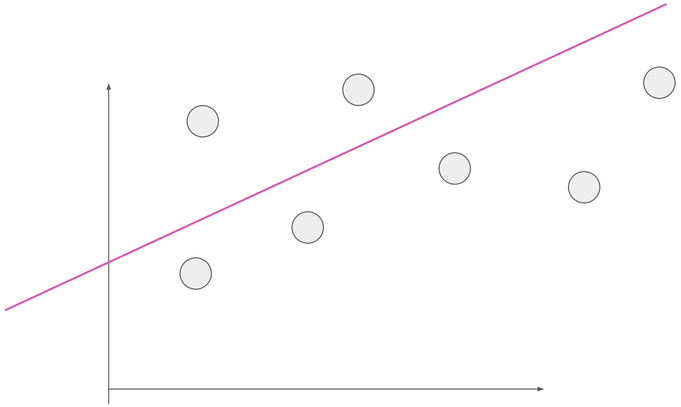
3	4	7

Chia dữ liệu (Splitting Data)

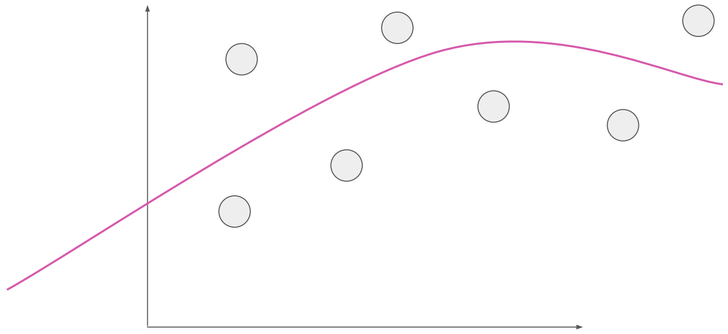


Overfitting - High Variance là hiện tượng mô hình có độ chính xác **cao** trên tập dữ liệu này tuy nhiên lại **thấp** trên tập dự dữ liệu khác

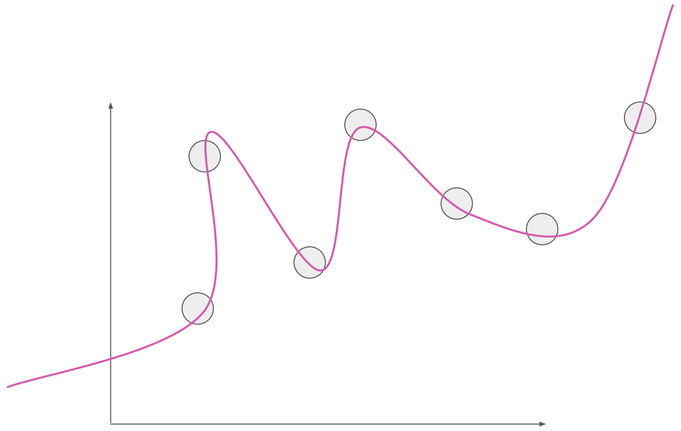
Tập validation để kiểm tra độ Variance của mô hình hiện tại



Mất mát lớn -
Underfitting



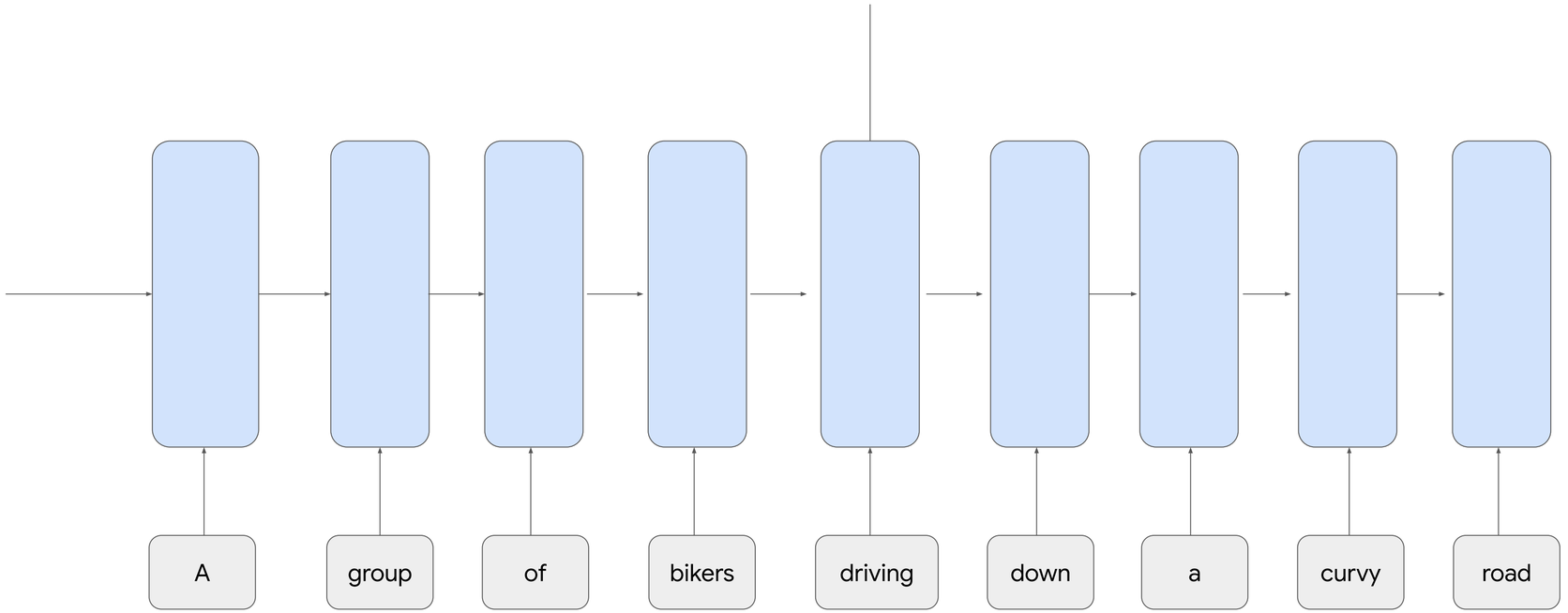
Hợp lý - đảm bảo tính tổng quát



Mất mát xấp xỉ 0
Overfitting

Đọc văn bản

Thông tin lịch sử từ đầu câu đến vị trí hiện tại (từ “driving”)



Dữ liệu ngôn ngữ có đặc tính **thứ tự theo thời gian** cho nên cần một mô hình phù hợp để đọc hiểu.

Kết hợp mô hình

A group of bikers driving down a curvy road.

RNN Model

Trích xuất ra thông tin quan trọng từ câu miêu tả

A

CNN Model

Trích xuất ra thông tin quan trọng trên bức ảnh

Attention Model

Kết nối những thành phần liên quan tới nhau

Attention is all you need

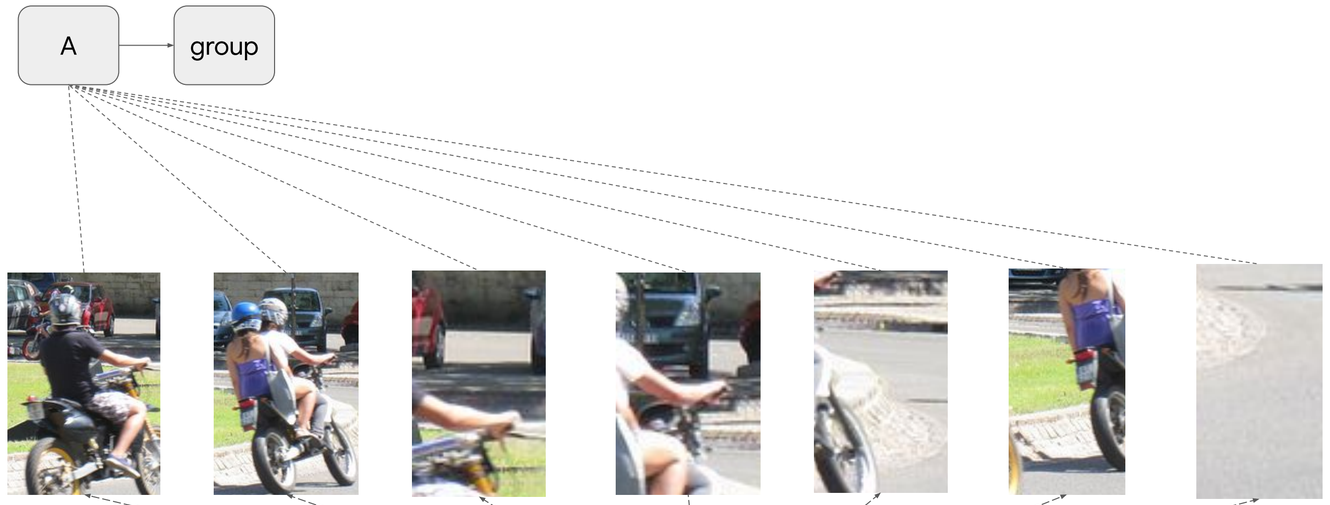
https://arxiv.org/abs/1706.03762

Kết hợp mô hình

A group of bikers driving down a curvy road.

RNN Model

Trích xuất ra thông tin
quan trọng từ câu miêu tả



CNN Model

Trích xuất ra thông tin quan trọng
trên bức ảnh

Attention Model

Kết nối những thành phần
liên quan tới nhau

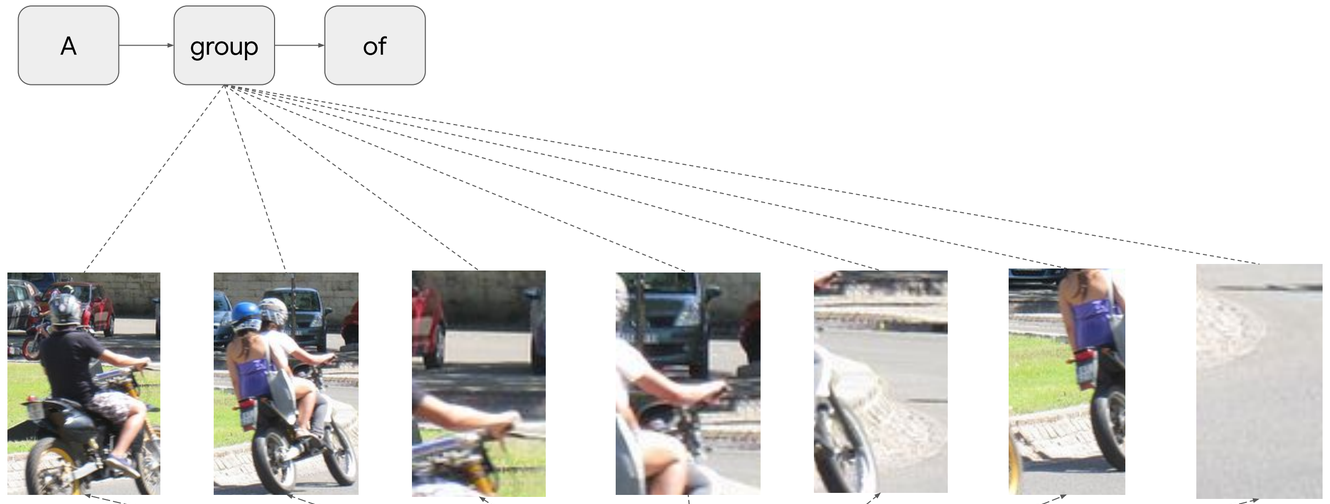
Attention is all you need
https://arxiv.org/abs/1706.03762

Kết hợp mô hình

A group of bikers driving down a curvy road.

RNN Model

Trích xuất ra thông tin
quan trọng từ câu miêu tả



Attention Model

Kết nối những thành phần
liên quan tới nhau

CNN Model

Trích xuất ra thông tin quan trọng
trên bức ảnh

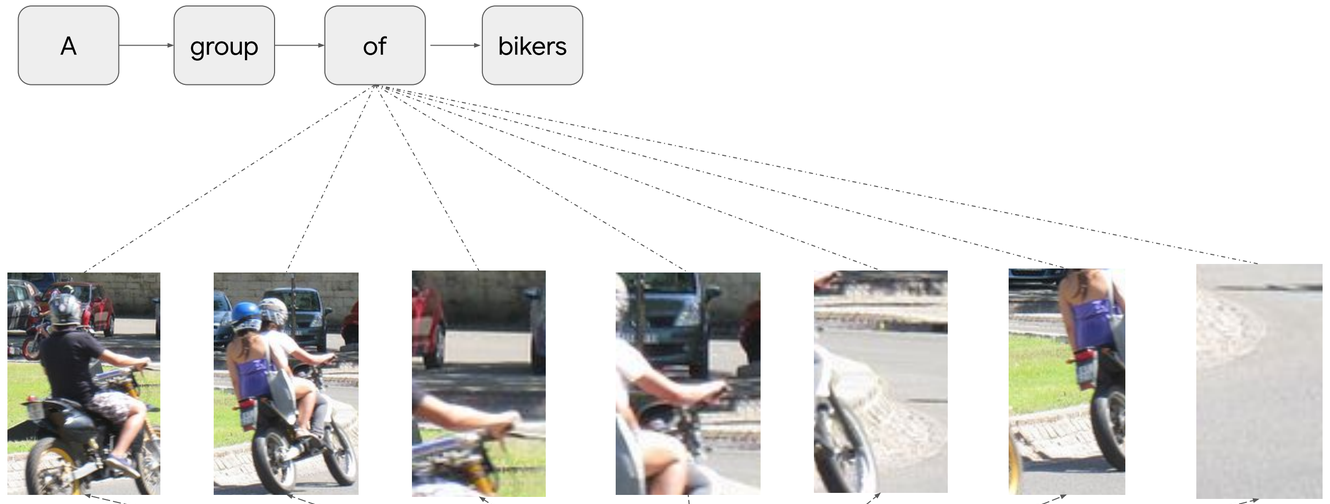
Attention is all you need
https://arxiv.org/abs/1706.03762

Kết hợp mô hình

A group of bikers driving down a curvy road.

RNN Model

Trích xuất ra thông tin
quan trọng từ câu miêu tả



Attention Model

Kết nối những thành phần
liên quan tới nhau

CNN Model

Trích xuất ra thông tin quan trọng
trên bức ảnh

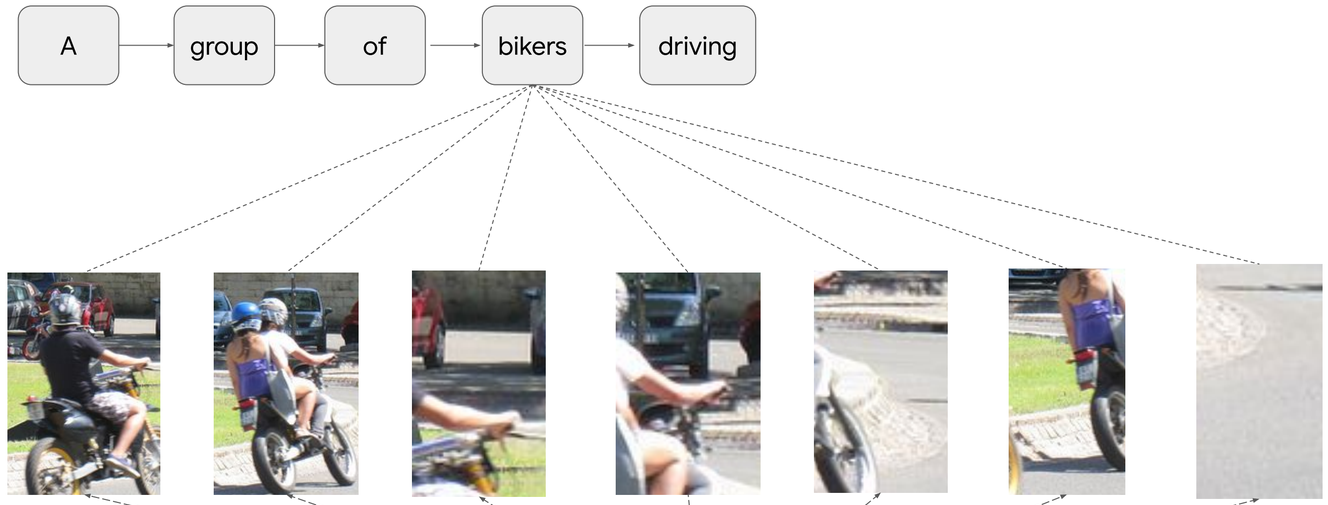
Attention is all you need
https://arxiv.org/abs/1706.03762

Kết hợp mô hình

A group of bikers driving down a curvy road.

RNN Model

Trích xuất ra thông tin
quan trọng từ câu miêu tả



Attention Model

Kết nối những thành phần
liên quan tới nhau

CNN Model

Trích xuất ra thông tin quan trọng
trên bức ảnh

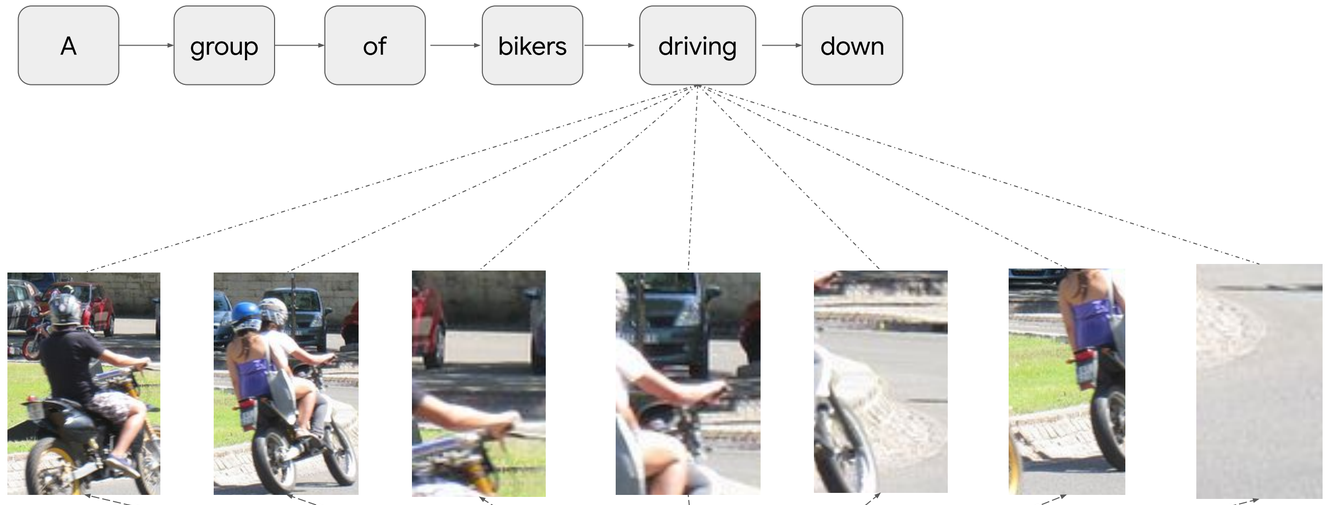
Attention is all you need
https://arxiv.org/abs/1706.03762

Kết hợp mô hình

A group of bikers driving down a curvy road.

RNN Model

Trích xuất ra thông tin
quan trọng từ câu miêu tả



Attention Model

Kết nối những thành phần
liên quan tới nhau

CNN Model

Trích xuất ra thông tin quan trọng
trên bức ảnh

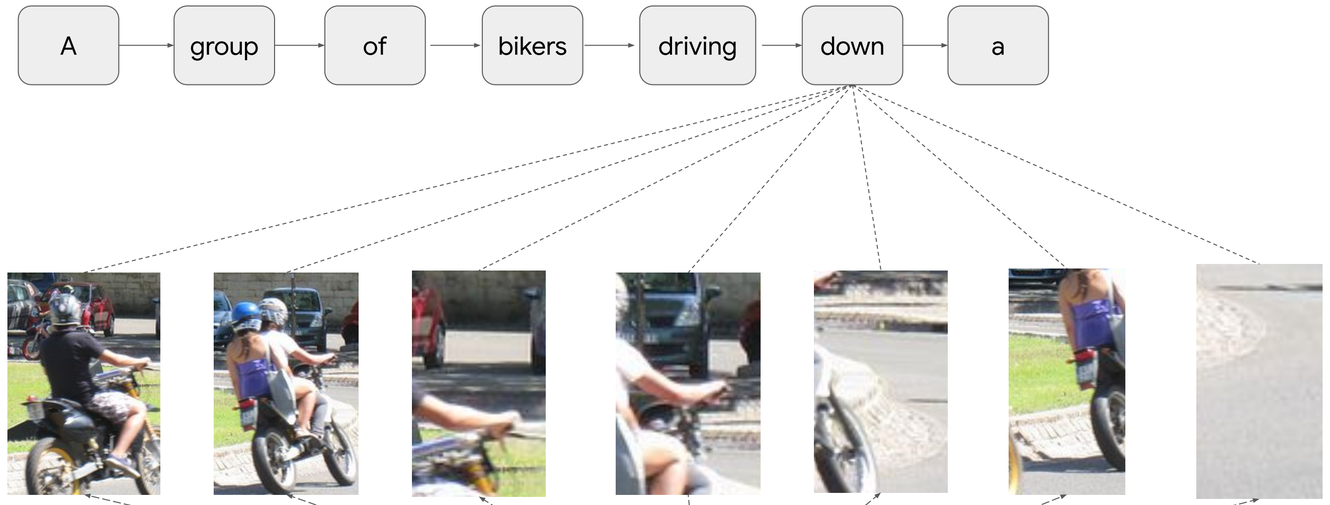
Attention is all you need
https://arxiv.org/abs/1706.03762

Kết hợp mô hình

A group of bikers driving down a curvy road.

RNN Model

Trích xuất ra thông tin
quan trọng từ câu miêu tả



Attention Model

Kết nối những thành phần
liên quan tới nhau

CNN Model

Trích xuất ra thông tin quan trọng
trên bức ảnh

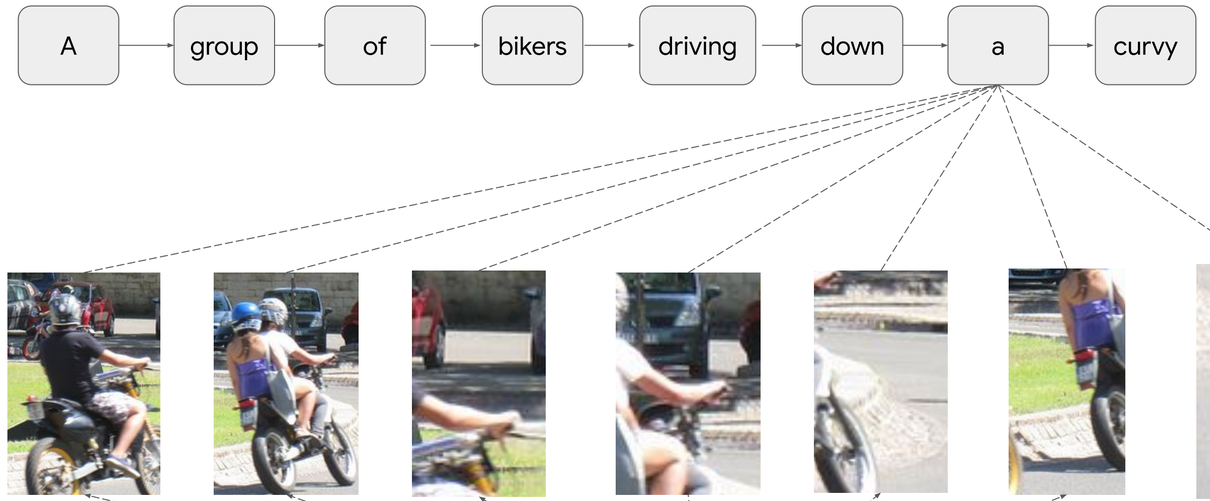
Attention is all you need
https://arxiv.org/abs/1706.03762

Kết hợp mô hình

A group of bikers driving down a curvy road.

RNN Model

Trích xuất ra thông tin
quan trọng từ câu miêu tả



Attention Model

Kết nối những thành phần
liên quan tới nhau

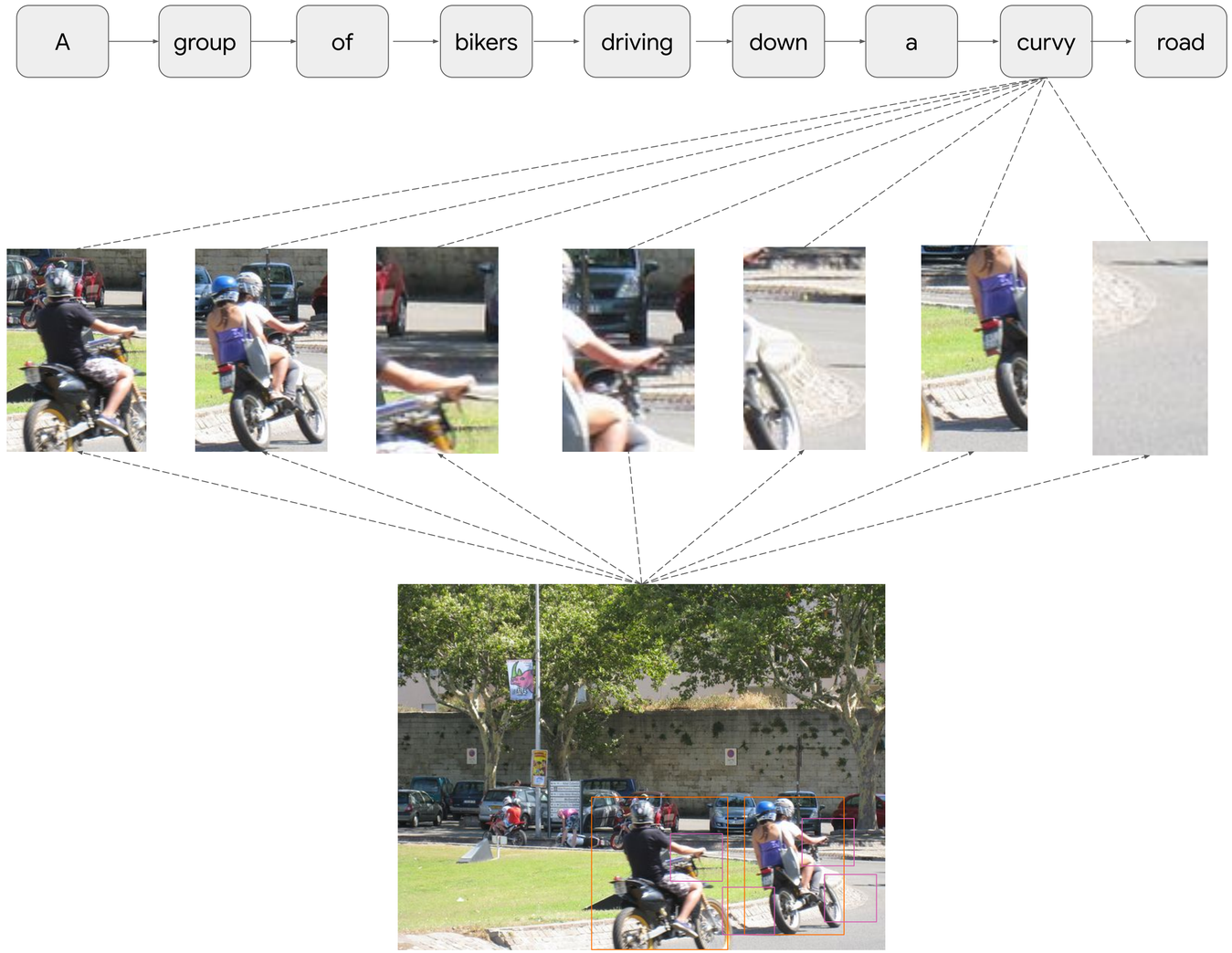
CNN Model

Trích xuất ra thông tin quan trọng
trên bức ảnh

Attention is all you need
https://arxiv.org/abs/1706.03762

Kết hợp mô hình

A group of bikers driving down a curvy road.



RNN Model

Trích xuất ra thông tin quan trọng từ câu miêu tả

CNN Model

Trích xuất ra thông tin quan trọng trên bức ảnh

Attention Model

Kết nối những thành phần liên quan tới nhau

Attention is all you need

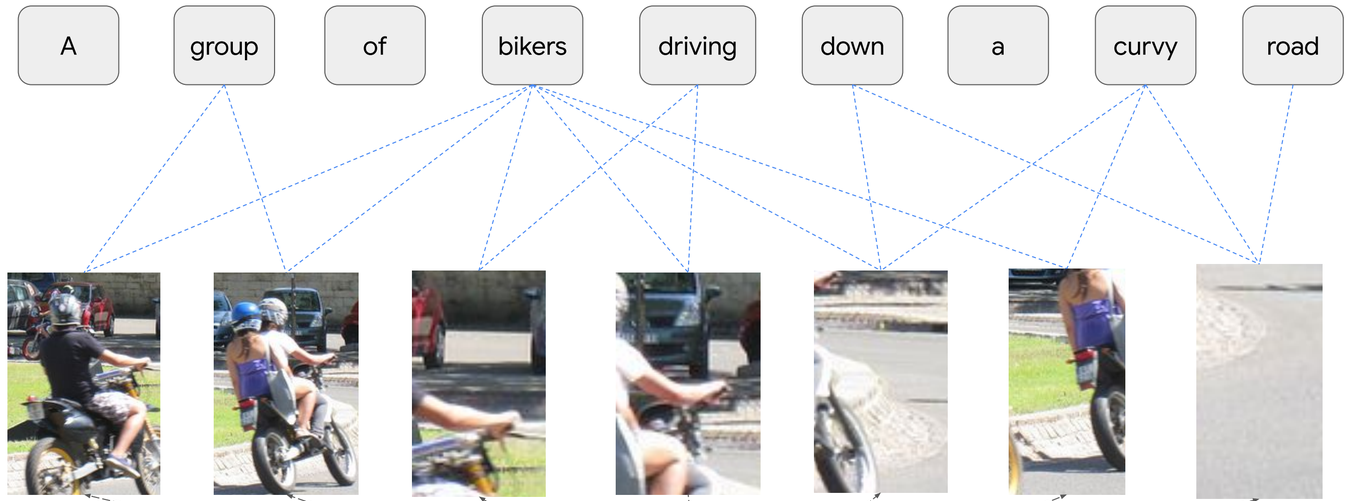
https://arxiv.org/abs/1706.03762

Những kết nối quan trọng để đưa quyết định

A group of bikers driving down a curvy road.

RNN Model

Trích xuất ra thông tin quan trọng
từ câu miêu tả



CNN Model

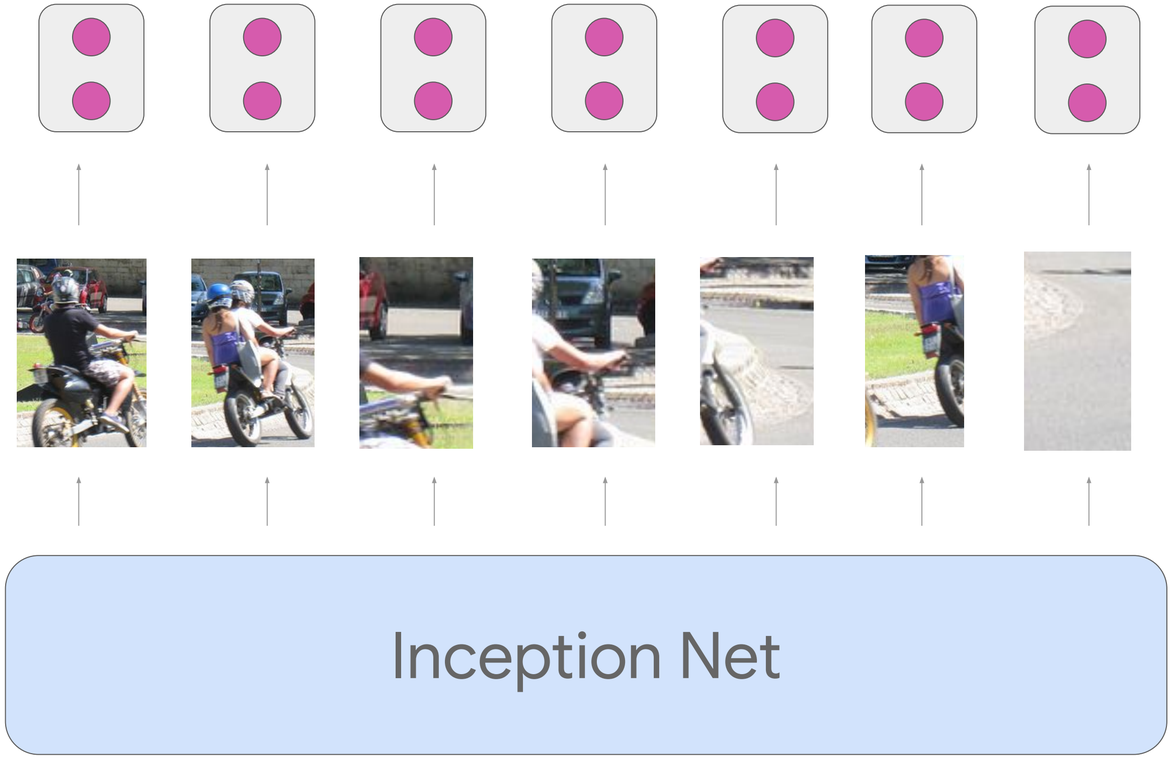
Trích xuất ra thông tin quan trọng
trên bức ảnh

Attention Model

Kết nối những thành phần
liên quan tới nhau

Attention is all you need
https://arxiv.org/abs/1706.03762

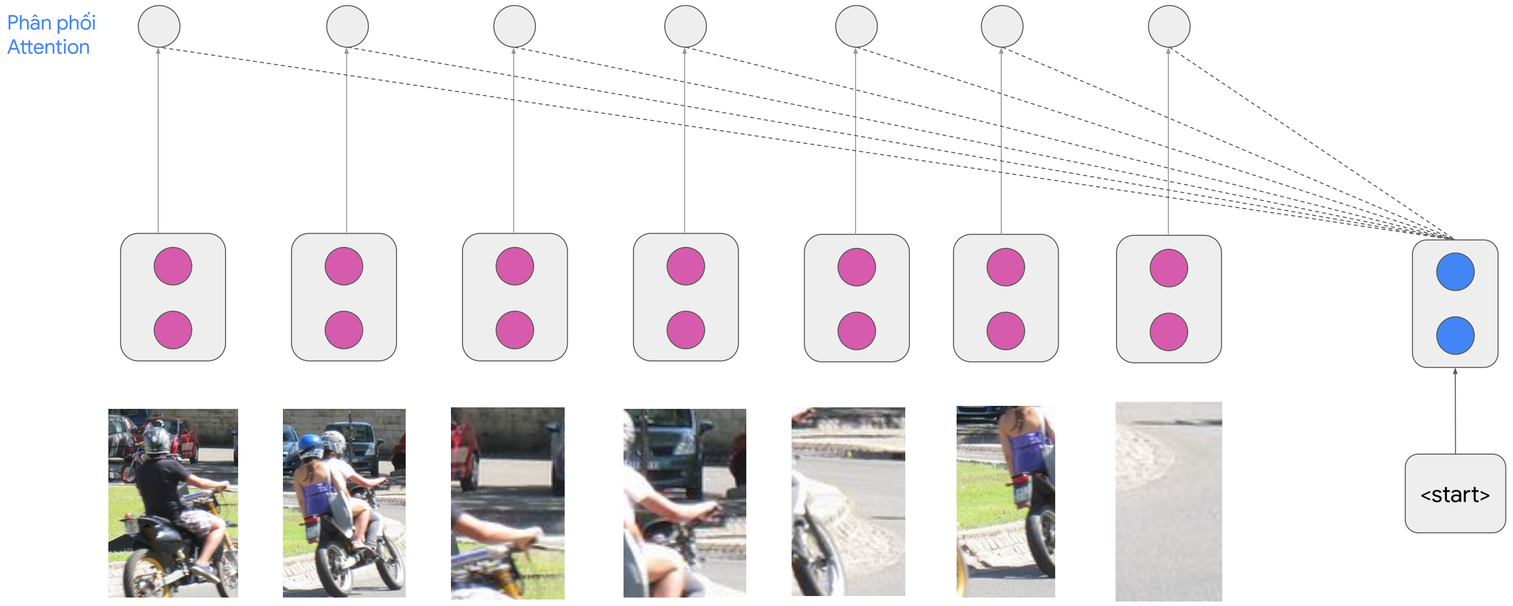
Chi tiết Attention



Ảnh đưa qua Inception Net

Khởi nguồn: Bài toán dịch máy - https://arxiv.org/abs/1409.0473

Chi tiết Attention

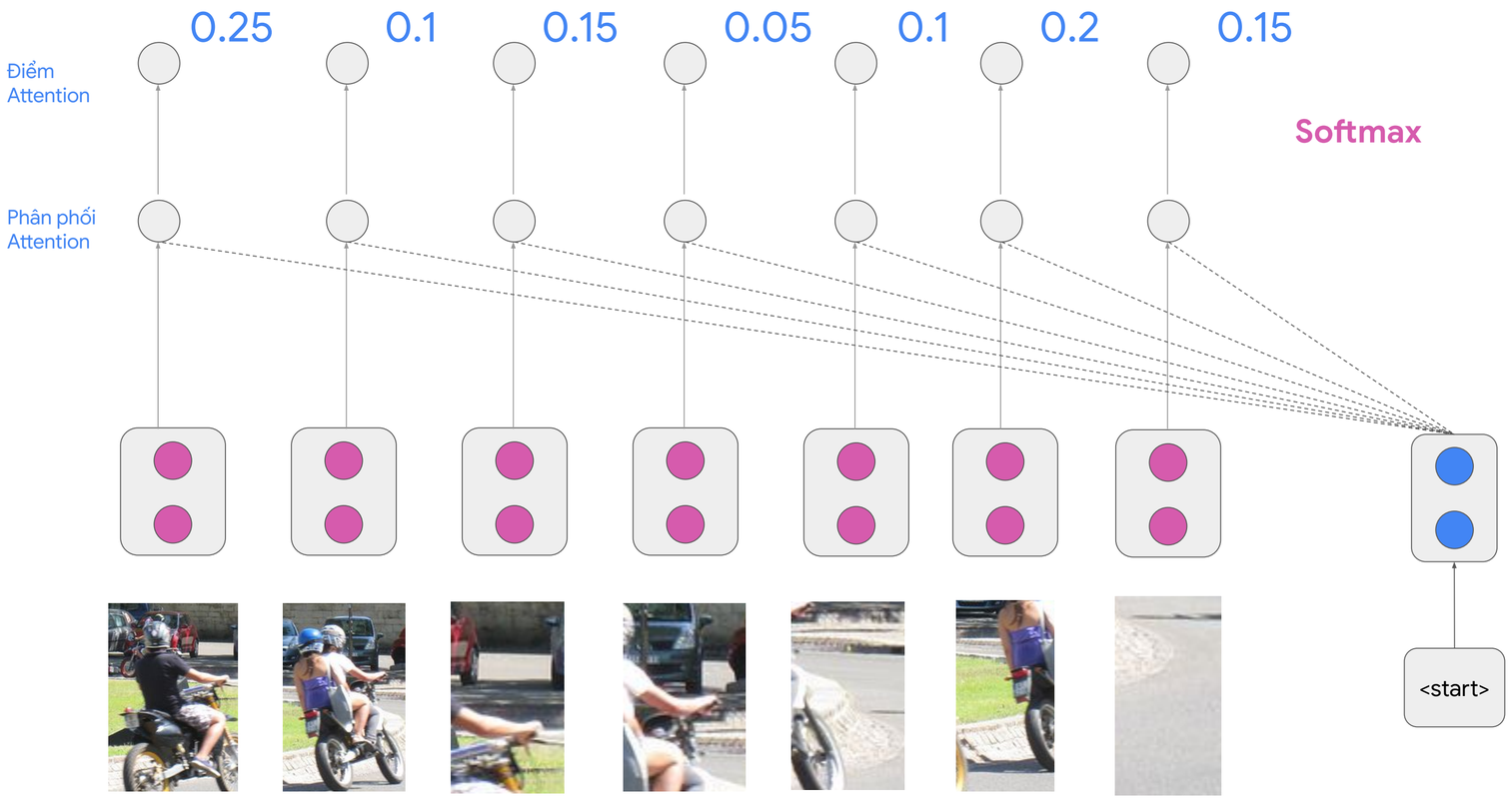


Liên kết hidden state của từ với các vùng ảnh

Trong những mô hình đơn giản thì có thể
sử dụng phép nhân vô hướng 2 vector

Khởi nguồn: Bài toán dịch máy - https://arxiv.org/abs/1409.0473

Chi tiết Attention

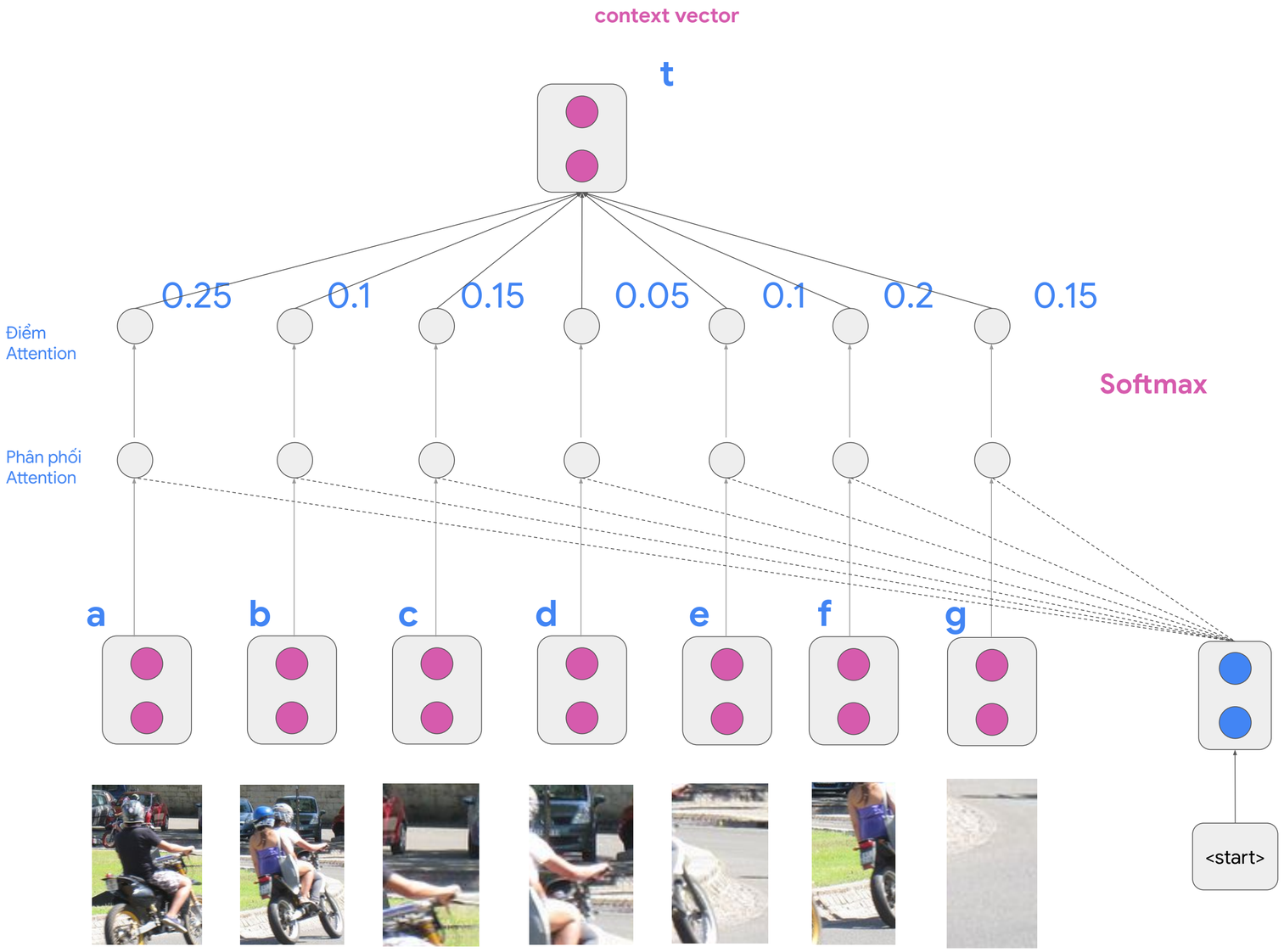


Mô hình hóa dưới dạng xác suất

Các giá trị quan hệ này được mô phỏng thành các giá trị đại diện cho một phân bố với **tổng các gía trị bằng 1**

Khởi nguồn: Bài toán dịch máy - https://arxiv.org/abs/1409.0473

Chi tiết Attention

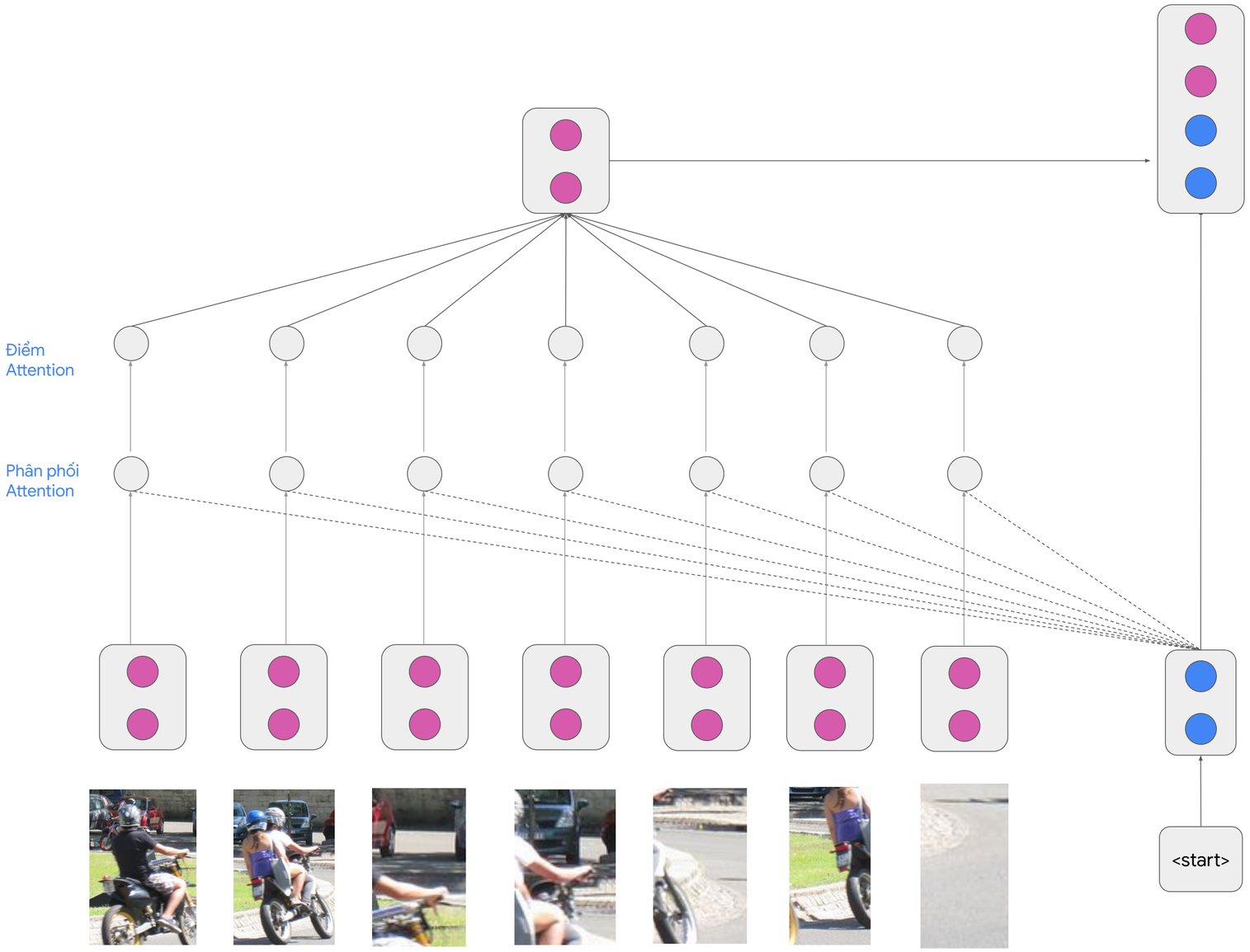


Tổng hợp mối quan hệ thành vector đại diện

t = 0.25 * a + 0.1 * b + 0.15 * c + 0.05 * d + 0.1 * e + 0.2 * f + 0.15 * g

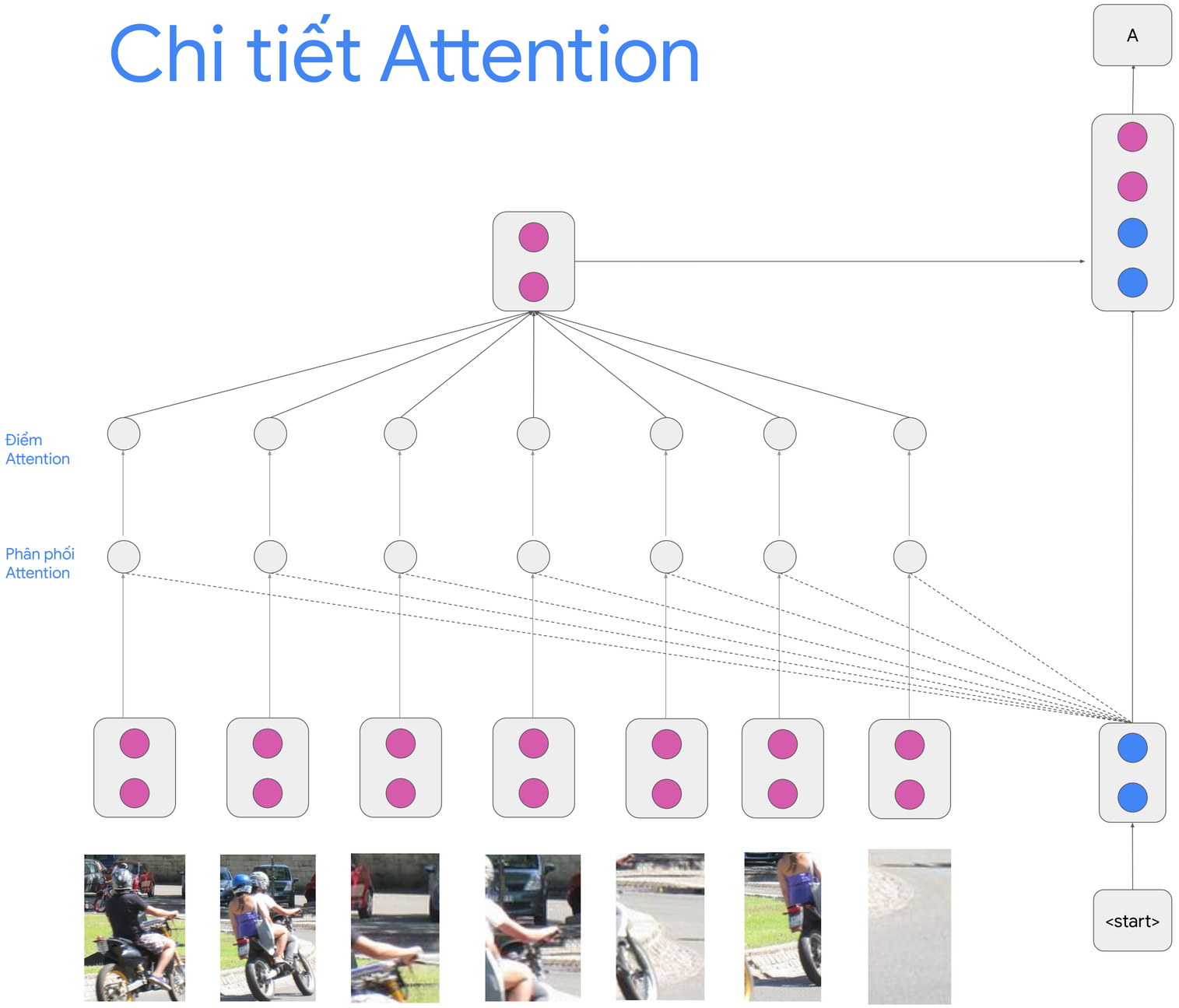
Khởi nguồn: Bài toán dịch máy - https://arxiv.org/abs/1409.0473

Chi tiết Attention



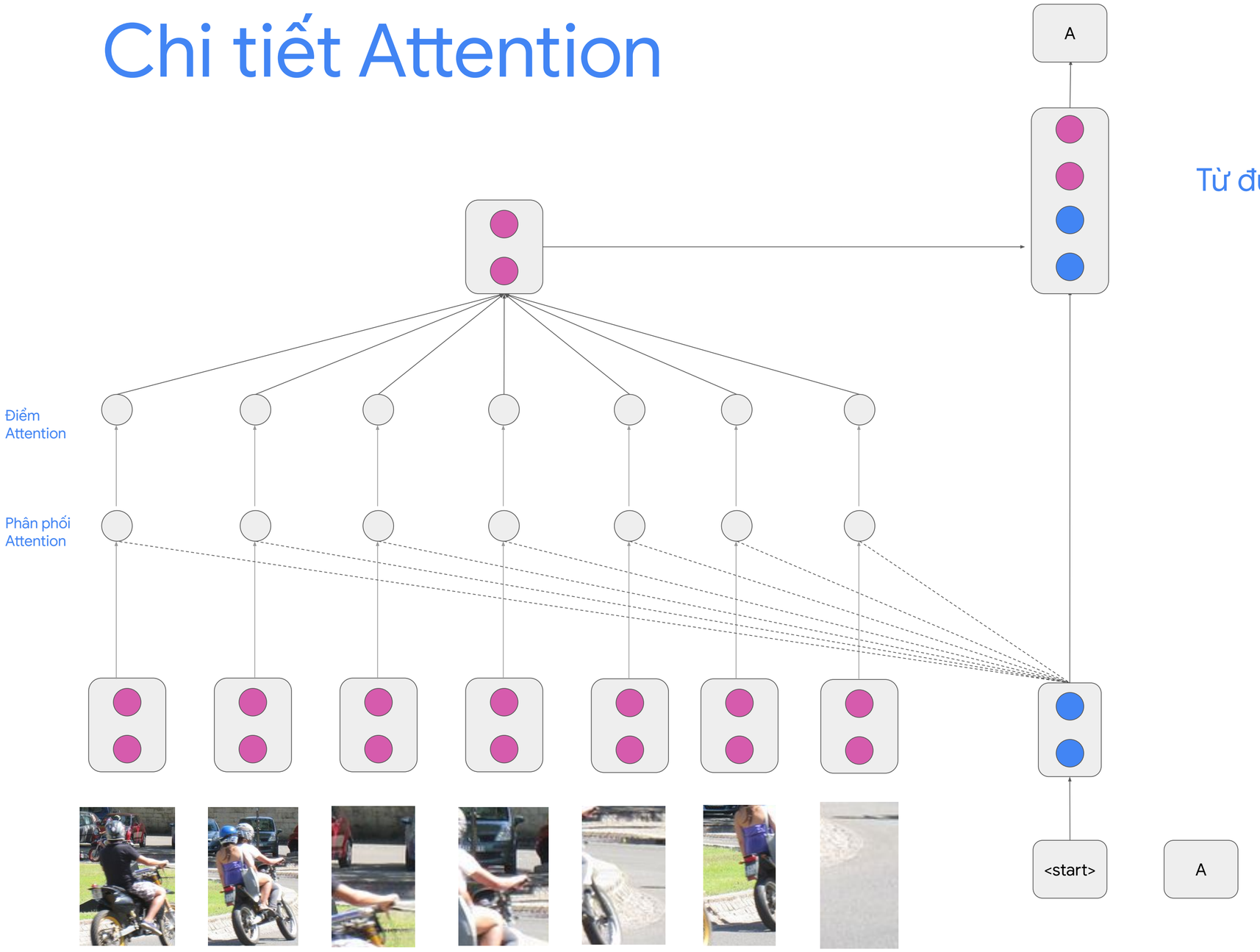
Kết nối mối quan hệ với từ hiện tại

Chi tiết Attention



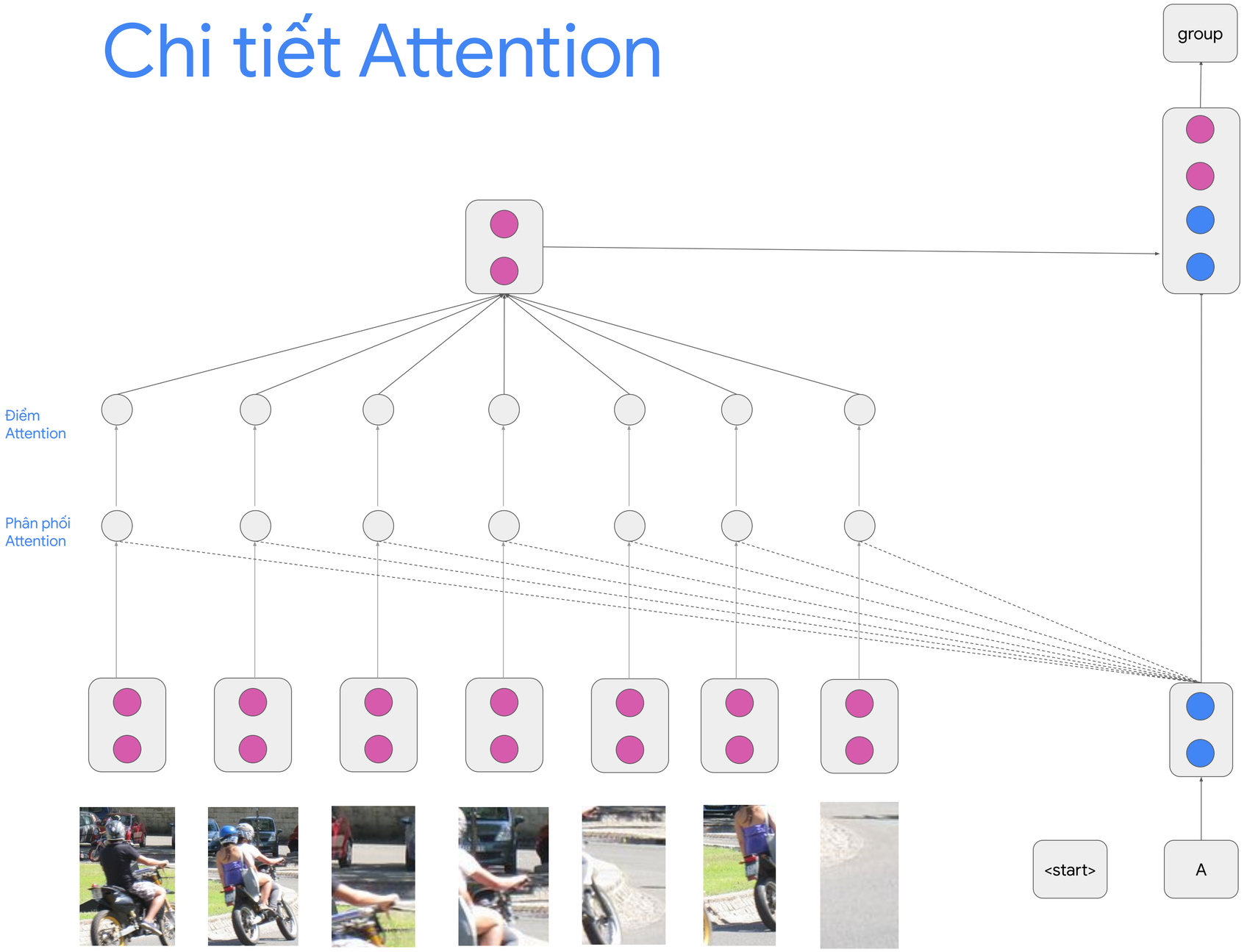
Tiến hành dự đoán

Chi tiết Attention



Từ được dự đoán tiếp tục được sử dụng

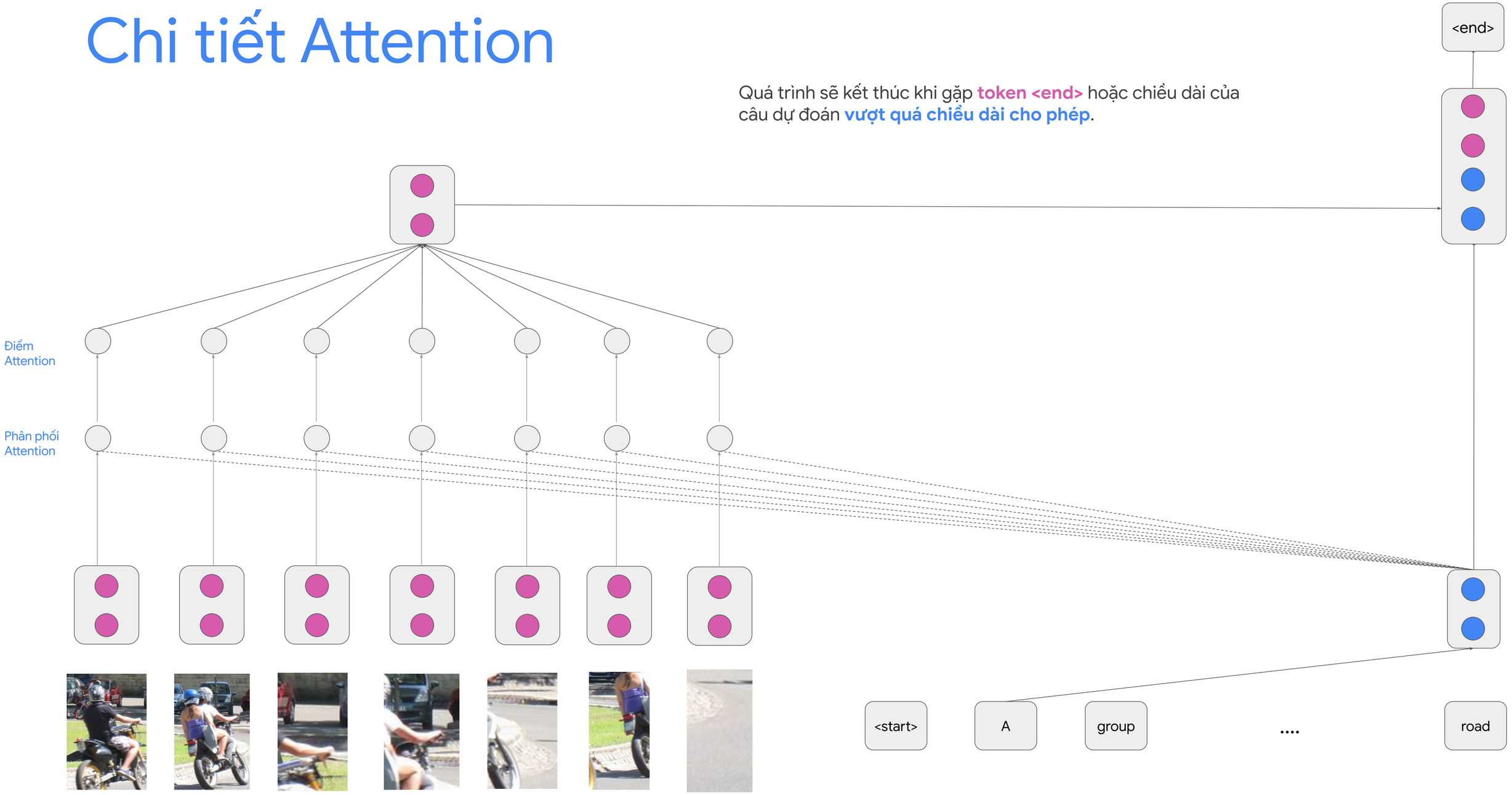
Chi tiết Attention



Từ được dự đoán tiếp tục được sử dụng

Chi tiết Attention

Quá trình sẽ kết thúc khi gặp **token <end>** hoặc chiều dài của câu dự đoán **vượt quá chiều dài cho phép**.



Bước chân vào ngành

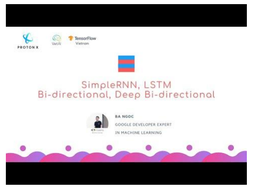
Khóa học

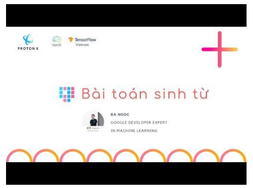
Đại số	https://www.khanacademy.org/math/algebra
Xác suất	https://www.khanacademy.org/math/statistics-probability
Đạo hàm	https://www.khanacademy.org/math/statistics-probability
ML Cơ bản	https://youtu.be/PPLop4L2eGk?list=PLLssT5z_DsK-h9vYZkQkYNWcItqhlRJLN
Thị giác máy tính - CS231	http://cs231n.stanford.edu/
Xử lý ngôn ngữ tự nhiên	http://web.stanford.edu/class/cs224n/

Sách

Deep Learning	https://www.deeplearningbook.org/
Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow	https://www.khanacademy.org/math/statistics-probability
Speech and Language Processing	https://web.stanford.edu/~jurafsky/slp3/

Chuỗi video luyện thi chứng chỉ Tensorflow





Bài báo (Thị giác máy tính) - Nâng cao

Network in Network (NIN)	https://arxiv.org/pdf/1312.4400.pdf
Inception Net	https://arxiv.org/pdf/1409.4842.pdf
Resnet	https://arxiv.org/abs/1512.03385 https://arxiv.org/pdf/1603.05027.pdf
MobileNet	https://arxiv.org/pdf/1704.04861.pdf

Bài báo (Xử lý ngôn ngữ tự nhiên) - Nâng cao

A Neural Probabilistic Language Model	https://www.jmlr.org/papers/volume3/bengio03a/bengio03a.pdf
Xác suất	https://arxiv.org/pdf/1408.3456v1.pdf
CBOW + SkipGram	https://arxiv.org/pdf/1301.3781.pdf
ELMO	https://arxiv.org/abs/1802.05365
BERT	https://arxiv.org/pdf/1810.04805.pdf

Bài báo (thuật toán huấn luyện) - Nâng cao

AdaGrad	https://www.jmlr.org/papers/volume3/bengio03a/bengio03a.pdf
AdaDelta	https://arxiv.org/pdf/1408.3456v1.pdf
RMSProp	http://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf
Adam	https://arxiv.org/abs/1412.6980
Slanted triangular LRs	https://arxiv.org/pdf/1506.01186v6.pdf

Google Developers

Experts

Thank You!

Ngoc Ba
@ProtonX @VietAI
Email: protonxai@gmail.com