



TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN  
KHOA CÔNG NGHỆ THÔNG TIN

**NHÓM 12**

BÙI HUỲNH TRUNG NAM (19127478) - NGUYỄN CÔNG PHÚ (21C11018)  
TRẦN ANH TÚC (19127651) - CHUNG KIM KHÁNH (19127644)

**BÁO CÁO**

**VEHICLE IDENTIFY**

Học máy ứng dụng

**GIÁO VIÊN HƯỚNG DẪN**

PGS.TS LÊ HOÀNG THÁI  
THS. TRƯƠNG TẤN KHOA

Thành phố Hồ Chí Minh - 2022

## MỤC LỤC

<i>I.</i>	<i>Giới thiệu</i>	2
<i>II.</i>	<i>Động lực nghiên cứu</i>	2
1.	Ý nghĩa khoa học	2
2.	Ứng dụng thực tiễn	2
<i>III.</i>	<i>Phát biểu bài toán</i>	3
	Object detection	3
	Classify	3
	Các thách thức	3
<i>IV.</i>	<i>Mô hình đề xuất</i>	4
1.	VGG16 [1]	4
2.	Resnet50 [2]	5
3.	MobileNetV3 large [3]	6
4.	Vision Transformer [4]	6
<i>V.</i>	<i>Các nghiên cứu liên quan (Related works)</i>	7
<i>VI.</i>	<i>Thực nghiệm</i>	7
1.	Bộ dữ liệu [8]	7
2.	Kết quả	8
<i>VII.</i>	<i>Kết luận</i>	11
<i>VIII.</i>	<i>Nguồn tham khảo</i>	11

## I. Giới thiệu

Trong cuộc sống, đi trên đường, ngoài những biển báo giao thông thì còn có những bảng quảng cáo to hoặc những bảng quảng cáo led. Nếu chỉ cố định một quảng cáo nhất định thì thật đáng tiếc vì không phải quảng cáo nào cũng phù hợp với mọi người. Cụ thể, người chạy xe G63 sẽ có sự quan tâm về những quảng cáo liên quan đến môn thể thao Golf hơn.



Vì thế, một ý tưởng được đưa ra là làm cách nào để có thể tùy chỉnh quảng cáo cho phù hợp với đối tượng người lái xe. Việc những quảng cáo này có phản ánh sự quan tâm của từng người lái xe hay không là điều đang được quan tâm. Tuy nhiên, xét trên tổng thể, quá trình sau đây là cách hoạt động của quảng cáo:

1. Phân tích nhân khẩu học.
2. Xác định sở thích của họ.
3. Hiển thị quảng cáo phản ánh những sở thích này.

Trong chương này, chúng ta sẽ tìm hiểu cách xây dựng một hệ thống thị giác của biển quảng cáo thông minh thông qua đánh giá loại xe của người lái. Cụ thể hơn là ta sẽ trích xuất các đặc trưng cơ bản của từng loại xe để so sánh và phân loại.

## II. Động lực nghiên cứu

### 1. Ý nghĩa khoa học

Trên thực tế, cho đến ngày nay, có rất nhiều các hãng xe và mẫu xe ra đời hằng năm. Với một lượng lớn dữ liệu như thế, việc truy vấn càng trở nên phức tạp. Từ đó, việc phát triển bài toán nhận dạng và phân lớp dựa trên **tập dữ liệu lớn (large-scale)** là rất cần thiết.

### 2. Ứng dụng thực tiễn

Việc áp dụng hệ thống quảng cáo thông minh sẽ giúp thúc đẩy hiệu quả của các quảng cáo. Khi ta chọn đúng quảng cáo cho đúng đối tượng, **người xem sẽ quan tâm nhiều hơn về nội dung quảng cáo** và tăng cường doanh thu từ các quảng cáo ấy. Ngoài ra, khi lặp đi lặp lại việc phải xem những nội dung quảng cáo mà ta không quan tâm, ta sẽ cảm thấy khó chịu. Điều này dẫn đến việc con người sẽ cảm thấy việc xem quảng cáo là rất phiền phức. Vậy nên ta cần một hệ thống quảng cáo thông minh có thể **làm hài lòng người xem** (tệp khách hàng quan trọng). Và cuối cùng là chọn đúng đối tượng sẽ giúp ta **tối ưu chi phí quảng cáo**. Cụ thể một quảng cáo đặt ở bên đường 1 giờ sẽ là 50.000 (VND) thì sau 1 ngày phát liên tục quảng cáo đấy ta sẽ mất 1.200.000 (VND) tiền quảng cáo. Ngược lại nếu chỉ phát

quảng cáo khi gặp đúng đối tượng, ta sẽ tiết kiệm được một số tiền quảng cáo (cũng như thời gian quảng cáo) không cần thiết.

### III. Phát biểu bài toán

#### Object detection

- **Input:** Chuỗi các frame ảnh quan sát được (Trích xuất từ camera)
- **Output:** Xác định đối tượng xe trong frame ảnh

#### Classify

- **Input:** Chuỗi các frame ảnh có đối tượng xe đã được Detect
- **Output:** Loại xe, hãng xe

#### Các thách thức

Các yếu tố trong lúc vận hành thực tế:

##### 1. Xe thuê/mượn

Trong thực tế, có rất nhiều người đang chạy xe thuê/mượn từ người khác vì không đủ tiền để sở hữu một chiếc xe. Từ đây, đối tượng người lái đây sẽ rất khó để xác định chính xác sở thích cá nhân của họ thông qua việc đánh giá dựa trên xe họ chạy.

##### 2. Thay đổi bìe ngoài của xe, trang trí xe.

Do nhu cầu và sở thích cá nhân, nhiều người rất thích thiết kế lại xe của mình. Việc điều chỉnh như vậy rất khó để máy có thể xác định chính xác được loại xe thông qua những đặc điểm cơ bản đã được học.

##### 3. Yếu tố thời tiết.

Quan sát ngoài trời thì yếu tố thời tiết sẽ ảnh hưởng là điều không tránh khỏi. Cụ thể các ảnh hưởng làm mờ dữ liệu như sương mù hoặc thời tiết xấu (trời mưa lớn). Điều này xảy ra thường xuyên và ta cần phải thích nghi với nó.

##### 4. Số lượng lớn xe đang lưu thông.

Vào các giờ cao điểm, xe lưu thông trên đường rất nhiều dẫn đến việc đa dạng các loại, mẫu mã xe. Như vậy rất khó để tiếp cận được chính xác người dùng.

Các yếu tố về dữ liệu xử lý:

##### 5. Góc nhìn của xe

Một số các loại xe có nét gần giống nhau nên việc các chi tiết, đặc điểm riêng của xe bị che khuất đi sẽ gây khó khăn trong việc xác định loại xe.

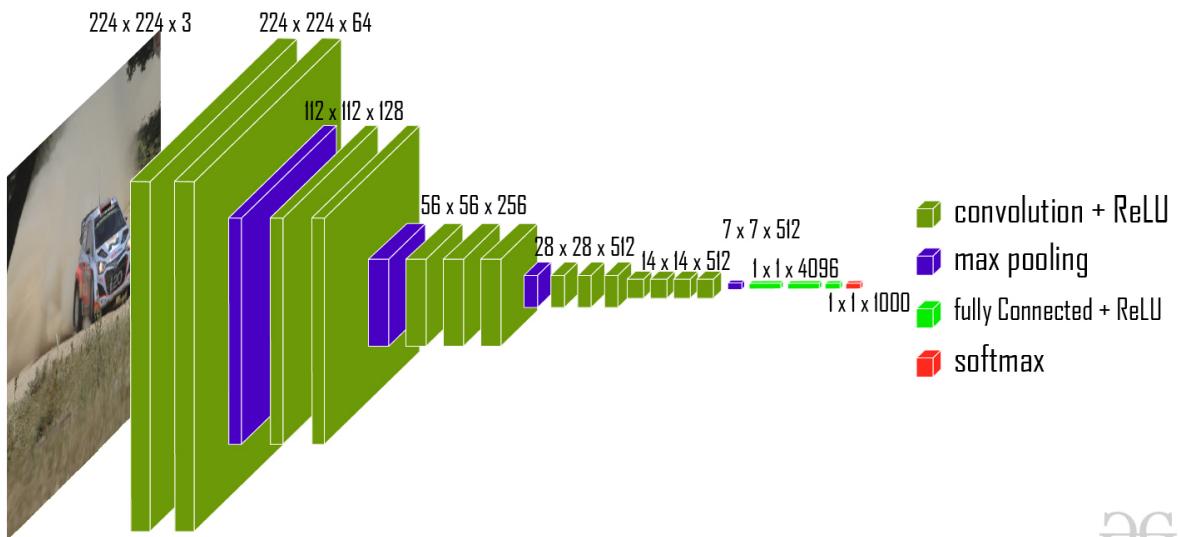
##### 6. Đa dạng về mẫu mã

Hiện tại có rất nhiều các nhãn hàng, mẫu mã, loại xe khác nhau và rất khó để ta có thể thu thập được hết các dữ liệu đầy đủ về các loại xe khác nhau để học. Việc này dẫn đến việc nguồn dữ liệu cần học rất lớn để phù hợp cho nhiều đối tượng người dùng khác nhau.

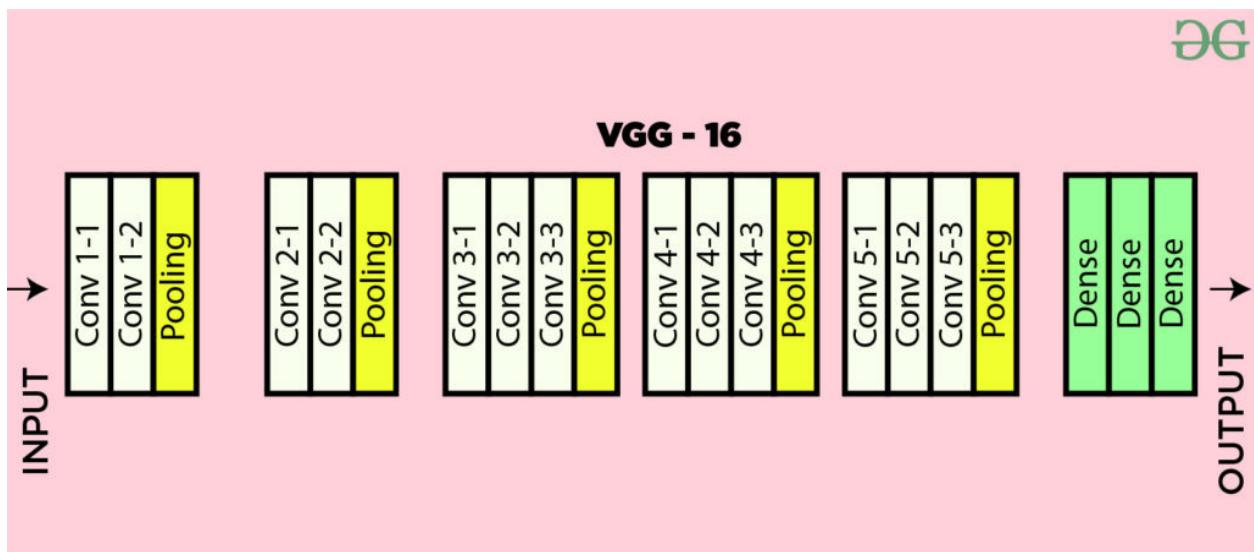
## IV. Mô hình đề xuất

### 1. VGG16 [1]

VGG 16 được đề xuất bởi Karen Simonyan và Andrew Zisserman thuộc Visual Geometry Group Lab của Đại học Oxford vào năm 2014 trong bài báo “Very Deep Convolutional Networks For Large-Scale Image Recognition” (Đã được đề cập đến ở phần IV. Related works). Mô hình này đã giành được vị trí thứ nhất và thứ hai trong các hạng mục của ILSVRC challenge năm 2014. Cụ thể là đứng thứ nhất trong nhiệm vụ phân loại với lỗi phân loại top 5 là 7,32% (chỉ sau GoogLeNet với sai số phân loại là 6,66%). Nó cũng là người chiến thắng trong nhiệm vụ localization với 25,32% localization error.



Đầu vào của mạng là một hình ảnh có kích thước  $224 \times 224$  và có màu RGB, dẫn tới input có  $(224, 224, 3)$ . Hai lớp đầu tiên có 64 channel với kích thước bộ lọc  $3 \times 3$  và phần padding giống nhau. Sau khi max-pooling layer của stride  $(2, 2)$ , hai lớp có các lớp tích chập 128 kích thước bộ lọc và kích thước bộ lọc  $(3, 3)$ . Tiếp theo là max-pooling layer của stride  $(2, 2)$  giống như lớp trước đó. Sau đó, có 2 lớp chập với kích thước bộ lọc  $(3, 3)$  và 256 bộ lọc. Sau đó, có 2 bộ 3 lớp chập và max-pooling layer. Mỗi bộ lọc có 512 bộ lọc có kích thước  $(3, 3)$  với cùng một padding. Hình ảnh này sau đó được chuyển đến ngăn xếp của hai lớp tích chập. Trong các lớp tích hợp và max-pooling layer này, các bộ lọc chúng tôi sử dụng có kích thước  $3 \times 3$  thay vì  $11 \times 11$  trong AlexNet và  $7 \times 7$  trong ZF-Net. Trong một số lớp, nó cũng sử dụng pixel  $1 \times 1$  được sử dụng để thao tác số lượng kênh đầu vào. Có một phần đệm 1 pixel (đệm giống nhau) được thực hiện sau mỗi lớp tích chập để ngăn chặn đặc điểm không gian của hình ảnh.



Sau lớp tích lũy và max-pooling layer, chúng tôi có một feature map (7, 7, 512) . Feature map này được làm phẳng thành một vectơ đặc trưng (1, 25088) . Sau đó, có 3 layers được kết nối đầy đủ (fully connected), layer đầu tiên lấy đầu vào từ vectơ đặc trưng cuối cùng và xuất ra một vectơ (1, 4096) , layer thứ hai cũng xuất ra một vectơ có kích thước (1, 4096), layer thứ ba xuất ra vector kết quả cho 1000 loại của Challenge ILSVRC (tức là lớp thứ 3 được kết nối đầy đủ được sử dụng để thực hiện chức năng softmax để phân loại 1000 loại). Tất cả các lớp ẩn sử dụng ReLU làm chức năng kích hoạt của nó. ReLU hiệu quả hơn về mặt tính toán vì nó giúp học tập nhanh hơn và nó cũng làm giảm khả năng biến mất các vấn đề về độ dốc.

Hạn chế của VGG 16:

- Rất chậm để đào tạo (mô hình VGG ban đầu được đào tạo trên GPU Nvidia Titan trong 2-3 tuần).
- Kích thước của trọng lượng imageNet được đào tạo VGG-16 là 528 MB. Vì vậy, nó chiếm khá nhiều dung lượng ổ đĩa và băng thông khiến nó hoạt động kém hiệu quả.
- 138 triệu tham số dẫn đến vấn đề độ dốc bùng nổ.

## 2. Resnet50 [2]

ResNet (Residual Network) được giới thiệu đến công chúng vào năm 2015 và thậm chí đã giành được vị trí thứ 1 trong cuộc thi ILSVRC 2015 với tỉ lệ lỗi top 5 chỉ 3.57%. Hiện tại, có rất nhiều biến thể của kiến trúc ResNet với số lớp khác nhau như ResNet-18, ResNet-34, ResNet-50, ResNet-101, ResNet-152,... Với tên là ResNet theo sau là một số chỉ kiến trúc ResNet với số lớp nhất định.

Mạng ResNet (R) là một mạng CNN được thiết kế để làm việc với hàng trăm hoặc hàng nghìn lớp chập. Một vấn đề xảy ra khi xây dựng mạng CNN với nhiều lớp chập sẽ xảy ra hiện tượng Vanishing Gradient dẫn tới quá trình học tập không tốt.

### 3. MobileNetV3 large [3]

MobileNets được giới thiệu lần đầu vào năm 2017. MobileNets có thể rút gọn lại vài triệu tham số nhưng vẫn giữ được độ chính xác ổn, đó là nhờ sử dụng một cơ chế gọi là Depthwise Separable Convolutions. MobileNetV3 được cải tiến chính ở việc bổ sung Squeeze-and-Excite.

Sự khác nhau giữa V2 và V3:

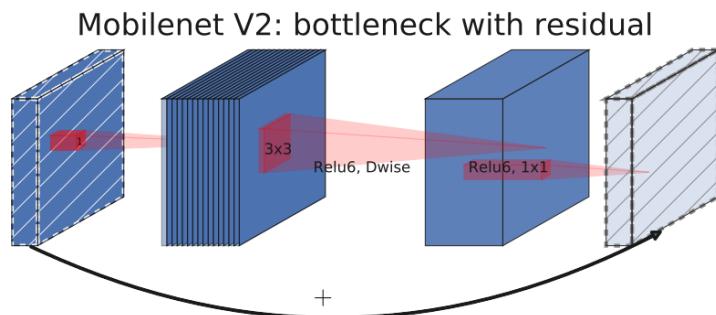


Figure 3. MobileNetV2 [39] layer (Inverted Residual and Linear Bottleneck). Each block consists of narrow input and output (bottleneck), which don't have nonlinearity, followed by expansion to a much higher-dimensional space and projection to the output. The residual connects bottleneck (rather than expansion).

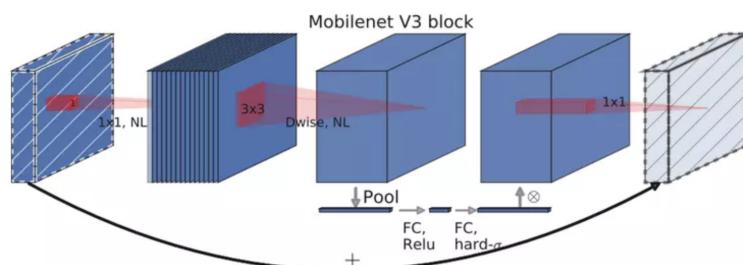


Figure 4. MobileNetV2 + Squeeze-and-Excite [20]. In contrast with [20] we apply the squeeze and excite in the residual layer. We use different nonlinearity depending on the layer, see section 5.2 for details.

### 4. Vision Transformer [4]

Vision Transformer lấy cảm hứng từ việc ứng dụng kiến trúc Transformer vào xử lý ngôn ngữ tự nhiên, các nhà nghiên cứu từ Google Research đã giới thiệu kiến trúc Vision Transformer - một phiên bản kiến trúc Transformer cho ảnh trong bài báo “An Image Is Worth 16x16 Words: Transformers For Image Recognition At Scale”.

Cách hoạt động:

1. Chia ảnh thành các phần nhỏ và duỗi thẳng.
2. Nhúng vị trí - Position embedding.
3. Transformer Encoder.

## V. Các nghiên cứu liên quan (Related works)

Trong phần này là các bài nghiên cứu liên quan để đưa ra được phương pháp tiếp cận tốt nhất cho bài toán.

Mô hình	Bài báo	Link	Mô tả	Kết quả
Vgg-16	Very Deep Convolutional Networks For Large-Scale Image Recognition	<a href="https://arxiv.org/pdf/1409.1556.pdf">https://arxiv.org/pdf/1409.1556.pdf</a>	Sử dụng kiến trúc mạng với các lớp convolution kích thước 3x3 để tăng chiều sâu cho mô hình.	Top 1 accuracy: 0.8351 Top 5 accuracy: 0.9574
Resnet-50	Deep Residual Learning for Image Recognition	<a href="https://arxiv.org/pdf/1512.03385.pdf">https://arxiv.org/pdf/1512.03385.pdf</a>	Sử dụng kết nối tắt để tránh tình trạng tiêu biên đạo hàm, từ đó có thể tăng số lớp của mô hình.	Top 1 accuracy: 0.8321 Top 5 accuracy: 0.9565
Mobile Net v3	Searching for MobileNetV3	<a href="https://arxiv.org/pdf/1905.02244.pdf">https://arxiv.org/pdf/1905.02244.pdf</a>	Được phát triển lên từ mô hình Mobile Net v2. Mô hình có ưu điểm nhỏ nhẹ, nên có thể sử dụng trên các thiết bị nhúng.	Top 1 accuracy: 0.7386 Top 5 accuracy: 0.9115
Vision Transformer	An Image Is Worth 16x16 Words: Transformers For Image Recognition At Scale	<a href="https://arxiv.org/pdf/2010.11929.pdf">https://arxiv.org/pdf/2010.11929.pdf</a>	Áp dụng kiến trúc Transformer nổi tiếng bên NLP vào Computer Vision. Cơ chế attention giúp cho mô hình có thể tập trung vào những phần quan trọng của bức ảnh.	Top 1 accuracy: 0.8175 Top 5 accuracy: 0.9517

## VI. Thực nghiệm

Đây là kết quả thực nghiệm trên mô hình VGG-16.

### 1. Bộ dữ liệu [8]

Kết quả được thực nghiệm trên “Bộ dữ liệu ô tô Stanford”. Đây là tập hợp 16.185 hình ảnh của 196 ô tô (Hình bên dưới), do Krause et al. trong án phẩm năm 2013 của họ “3D Object Representation for Fine-Grained Classification”.



Tải xuống bản lưu trữ của tập dữ liệu tại đây: <http://pyimg.co/9s9mx>

Sau khi tải xuống, sử dụng lệnh sau để giải nén tập dữ liệu:

```
$ tar -xvf car_ims.tar.gz
```

## 2. Kết quả

Prediction: Jeep Compass SUV 2012  
Ground Truth: Jeep Compass SUV 2012

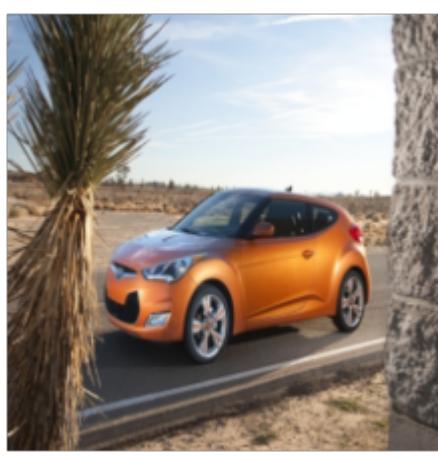


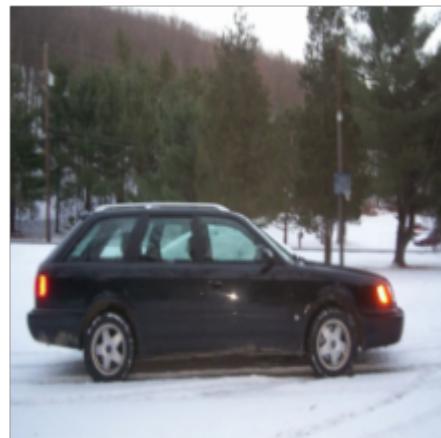
Prediction: Buick Rainier SUV 2007  
Ground Truth: Ford Focus Sedan 2007



Prediction: BMW M6 Convertible 2010  
Ground Truth: Aston Martin V8 Vantage Convertible 2012

Prediction: BMW X6 SUV 2012  
Ground Truth: BMW X6 SUV 2012

	
<p>Prediction: Ford Ranger SuperCab 2011          Ground Truth: Ford Ranger SuperCab 2011</p> 	<p>Prediction: Audi S5 Coupe 2012          Ground Truth: Audi S4 Sedan 2012</p> 
<p>Prediction: Fisker Karma Sedan 2012          Ground Truth: Tesla Model S Sedan 2012</p> 	<p>Prediction: Hyundai Veloster Hatchback 2012          Ground Truth: Hyundai Veloster Hatchback 2012</p> 
<p>Prediction: Volvo 240 Sedan 1993          Ground Truth: Volvo 240 Sedan 1993</p>	<p>Prediction: Audi 100 Wagon 1994          Ground Truth: Audi 100 Wagon 1994</p>



Prediction: Ferrari California Convertible 2012  
Ground Truth: Ferrari 458 Italia Coupe 2012



Prediction: Nissan Juke Hatchback 2012  
Ground Truth: Nissan Juke Hatchback 2012



Prediction: Bugatti Veyron 16.4 Convertible 2009  
Ground Truth: Bugatti Veyron 16.4 Coupe 2009

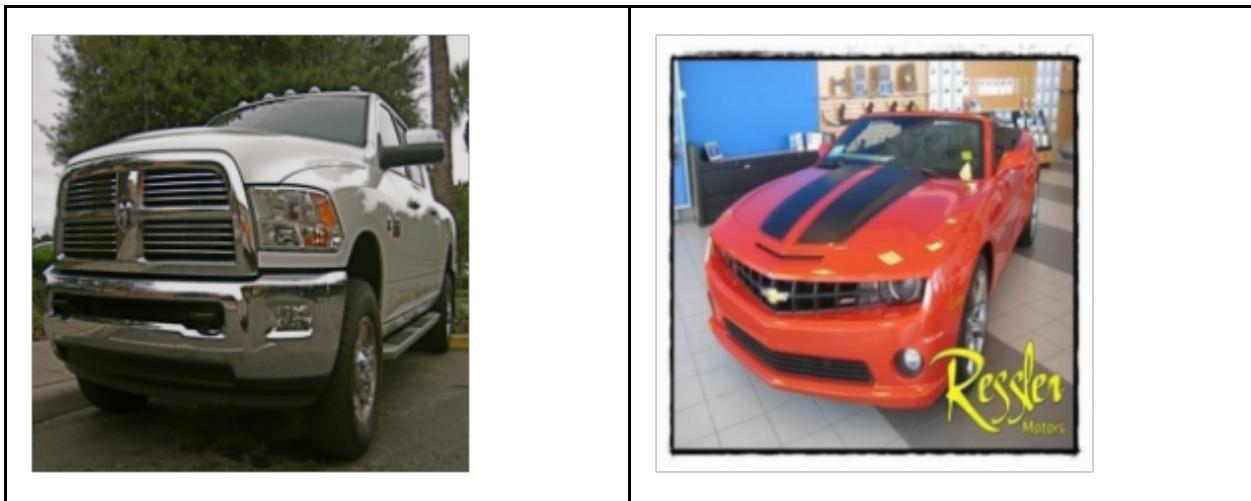


Prediction: Audi S6 Sedan 2011  
Ground Truth: Audi S6 Sedan 2011



Prediction: Dodge Ram Pickup 3500 Crew Cab 2010  
Ground Truth: Dodge Ram Pickup 3500 Crew Cab 2010

Prediction: Chevrolet Camaro Convertible 2012  
Ground Truth: Chevrolet Camaro Convertible 2012



## VII. Kết luận

Dựa trên 16 hình ảnh dữ liệu Test thì ta thấy có 6 kết quả sai (màu đỏ) và 10 kết quả đúng (màu đen), tỷ lệ chính xác của mô hình này là 62.5% (Đối với kết quả thực nghiệm ta thu được).

Tuy nhiên, do có:

1. Sự mất cân bằng về số lượng cực kỳ nghiêm trọng trong tập dữ liệu. Cụ thể một số dòng xe và mẫu xe được thu thập số lượng quá mức (ví dụ: Audi và BMW mỗi hãng có hơn 1.000 điểm dữ liệu trong khi Tesla chỉ có 77 ví dụ).
2. Dữ liệu rất ít, ngay cả đối với các lớp lớn.

Dr. Adrian Rosebrock [8] quyết định loại bỏ nhãn năm và thay vào đó phân loại các hình ảnh một cách chắt chẽ dựa trên kiểu dáng và kiểu dáng của chúng. Ngay cả khi thực hiện phân loại theo cách này, vẫn có những trường hợp ít hơn 100 hình ảnh cho mỗi lớp. Điều này khiến việc tinh chỉnh mạng CNN học sâu trên tập dữ liệu này rất khó khăn. Tuy nhiên, với phương pháp tinh chỉnh phù hợp vẫn có thể đạt được độ chính xác phân loại lớn hơn 95%. Sau khi loại bỏ nhãn năm, tập dữ liệu còn lại 164 tổng số loại xe và mô hình để nhận biết. Mục tiêu của việc làm này là tinh chỉnh VGG16 để xác định từng lớp trong số 164 lớp.

## VIII. Nguồn tham khảo

- [1] [VGG-16 | CNN model, geeksforgeeks, 29 Jun, 2022](#)
- [2] [Giới thiệu mạng ResNet, viblo, ToThang, 14 Jan, 2020](#)
- [3] [\[CNN Architecture series #1\] MobileNets - Mô hình gọn nhẹ cho mobile applications, viblo, huyennguyenthanh, 16 Sep, 2021](#)
- [4] [Vision Transformer - An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale, viblo, buiquangmanh, 21 Nov, 2021](#)
- [5] [When using VGG, why is 3 x 3 often used, and not 2 x 2, as the size of a convolution filter? - Quora](#)
- [6] [\(65\) Explained VGG-16 With Keras on Custom Dataset | Convolutional Neural Network | Deep Learning - YouTube](#)
- [7] [\(65\) 9. VGG16 architecture and implementation - YouTube](#)

[8] Deep Learning for Computer Vision with Python, Practitioner Bundle, Dr. Adrian Rosebrock,  
1st Edition (1.2.1)