專題報告

利用動作辨識破解 reCAPTCHA

指導老師：余執彰　副教授

學生：高偉承、張任宏

專題報告目錄

二、摘要

隨著人工智慧的成熟發展，現有的 reCAPTCHA 圖片驗證已被破解，因此本專題欲提出一解決方案，將物件辨識更為動作辨識。並嘗試站在攻擊者的角度訓練影像辨識模型，也站在防禦者角度假想不同情境驗證其防禦性。

核心精神是利用 Grad-CAM 技術得知模型對於圖片的關注區域，重複進行裁切之圖片並優化模型，探討裁切至何種程度的圖片足以區分機器與人類。

三、報告內容

● 研究動機

隨著人工智慧的成熟發展，原本防止網路機器人濫用服務的驗證機制 reCAPTCHA 已經被影像辨識技術破解[1]，本專題預測 reCAPTCHA 版本之升級，將原本為辨識圖片中之「物體」改為辨識圖片中之「主體動作」，我們認為這樣的更動對於機器來說難度會大幅增加，另外想知道機器面對這種類型之圖片的表現如何。

● 工具介紹

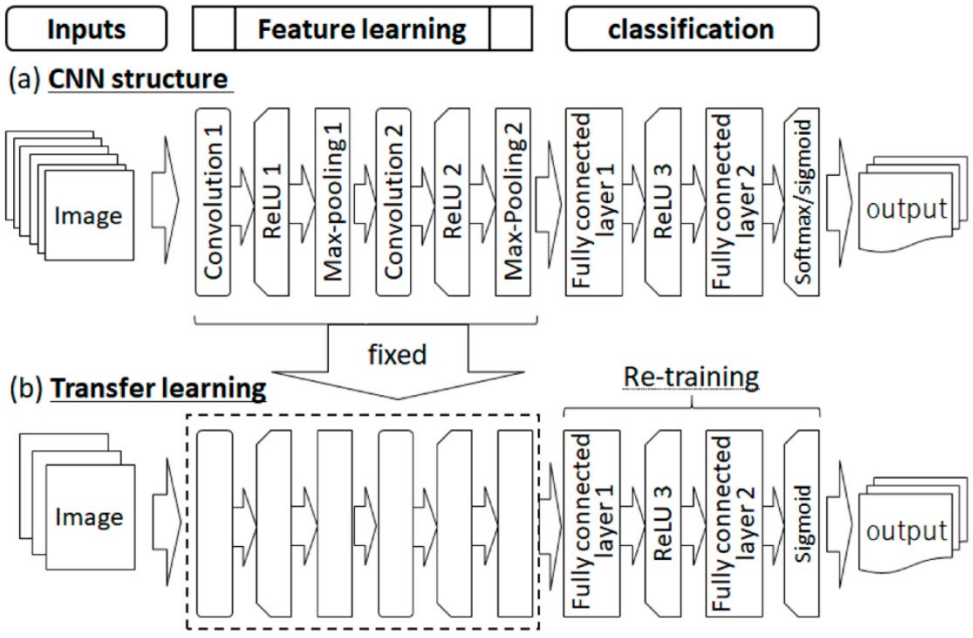
1. 資料集來源與製作

我們選用「籃球動作」作為本專題實驗的對象，由於網路上找不到專門製作籃球動作的資料集，於是從 Getty image 圖庫網，下載蒐集三種籃球動作的圖片，分別是「運球」、「灌籃」、「投籃」三類，且為了探討不同裁切比例的圖片會為模型帶來的影響如何，因此必須要為圖片進行加工，比如遮擋、裁切，作為本專題的資料集。

▲ 圖 1:由左而右分別為 運球、灌籃、投籃

2. Transfer learning

本專題採用了 Transfer learning[2]，並使用 Resnet152 作為 Pretrained model 來訓練模型。其實做的概念是先固定神經網路架構中間的卷積層，使模型為圖片作特徵提取，比如圖片的邊邊角角、輪廓，而當模型具備能獲得圖片底層結構的基礎訊息的能力之後，接著重新訓練分類器，再以較低的學習率幫整個模型做微調(Fine-tuning)，來適應新的數據，並為模型帶來更佳的學習成效。

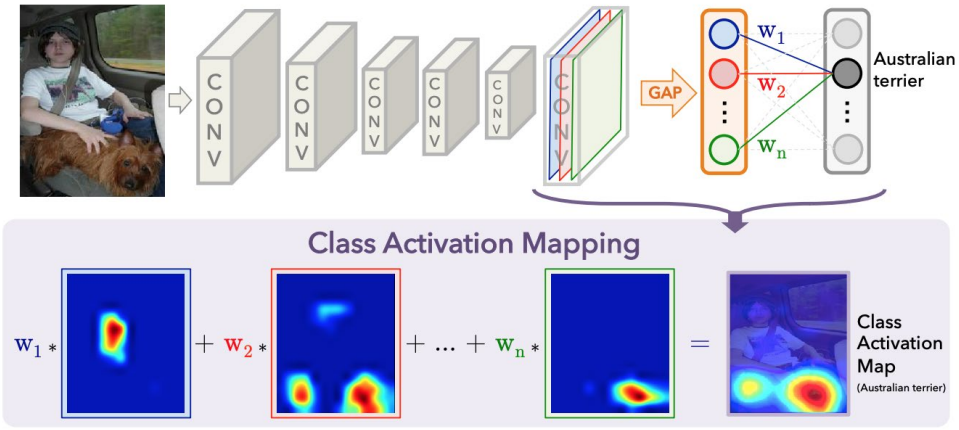


▲ 圖 2:Transfer Learning 示意圖

使用 Ttransfer learning 的好處是不需要蒐集大量的資料就能訓練模型，這一點對於專題前期沒有大量資料集的階段，依然能使模型有不錯的辨認率；另外也不須花大量的時間從頭訓練，這一點對專題後期不斷優化模型的訓練模式節省了大量的訓練時間。

3. Grad-CAM

我們使用基於 CAM[3]技術所產生的 Grad-CAM[4]技術，用於幫助了解模型專注的區域，其運作原理是利用類似反向傳播(back-propagation)的概念，使得到圖片中各個區域的像素在每一層做運算時的權重，並把這些權重在其對應的像素做堆疊，最後能得到關注區域熱力圖(如下圖)。



▲ 圖 3:Grad-CAM 技術示意圖

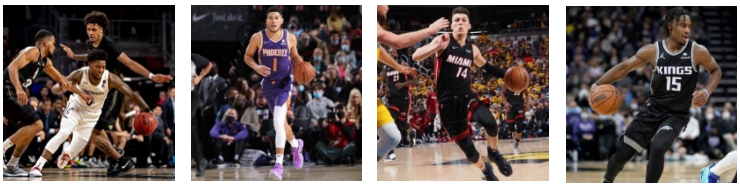
而從熱力圖中就能很直觀的判斷該區域對於預測結果的重要程度，本專題會依據這些熱力圖來判斷模型訓練的結果是否符合關注區域的預期，另外也能根據熱力圖來衡量裁切圖片的依據，以用於後續的訓練。

● 實驗過程

第一部分、完整圖片識別模型

1-1 模型簡介

本階段將訓練最基本的模型，欲測試經由大量圖片訓練後之學習成效，並觀察模型經由此類資料集訓練後關注的區域，作為後續資料集及模型之參考。使用的圖片為網路上隨機下載的圖片，同一種類別的圖片有不同的拍攝角度及姿勢，以及單一主體及多位主體的差別。



▲ 圖 4：運球

▲ 圖 5：灌籃

▲ 圖 6：投籃

	運球(張)	灌籃(張)	投籃(張)
訓練集	1500	1500	1500
驗證集	500	500	500
測試集	500	500	500

▲ 表 1：訓練資料種類及數目

1-2 模型訓練結果與討論

A. 測試集資料辨識之結果

運球	89 %
灌籃	95.8 %
投籃	95.6 %

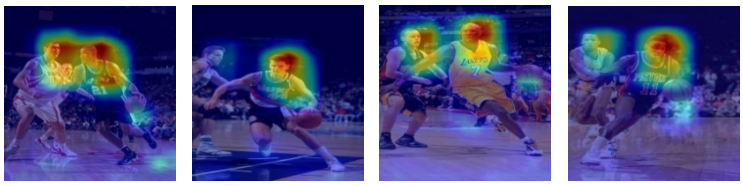
B. 測試集資料生成 Grad-CAM Picture

以下為根據 Grad-CAM Picture 歸納之特徵：

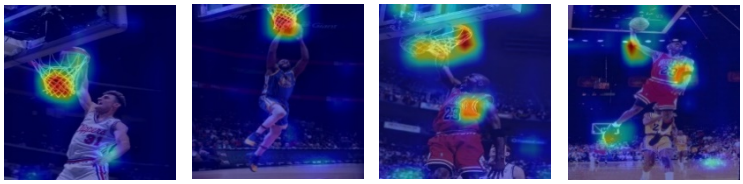
運球：人頭、球、非運球的人

灌籃：籃框、球、背景燈光

投籃：手肘、手掌＋球



▲ 圖 7：運球



▲ 圖 8：灌籃

▲ 圖 9：投籃

C. 階段討論

本階段辨識的準確率很高，模型皆能正確判斷圖片所屬的類別，但是我們發現模型極度仰賴類別的專屬特徵，比如「灌籃」類別的「籃框」、「運球」類別的「球員的頭部」，而非圖片中人物的動作，此與「動作辨識」相違背，因此視為模型的表現不符合預期之結果。

改善的模型的方向應著重在去除類別的專屬特徵，並把主角單一化，減少模型學習成效不佳的潛在因素，因此發展下一階段的模型訓練。

第二部分、遮擋重點區塊識別模型

2-1 模型簡介

本階段模型訓練的目標為，「去除類別的專屬特徵」，另外再使「主體單一化」，使模型只專注在單一主體，以減少潛在的變因，也迫使模型關注的區域更改為主體的身體的部位。

而本階段使用灰色方框來遮擋原模型所關注的區域，方法是來自於論文[5]，因灰色方框來降低模型對該區域的關注程度，希望藉此能迫使模型關注的區域改為人物主體的動作。

▲ 圖 10：運球

▲ 圖 11：灌籃

▲ 圖 12：投籃

	運球(張)	灌籃(張)	投籃(張)
訓練集	400	400	400
驗證集	50	50	50
測試集	50	50	50

▲ 表 2：訓練資料種類及數目

2-2 模型訓練結果與討論

A. 測試集資料辨識之結果

運球	100 %
灌籃	98%
投籃	96%

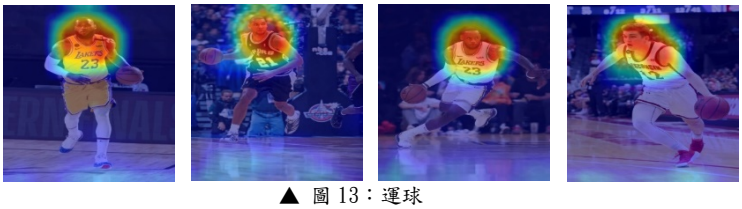
B. 測試集資料生成 Grad-CAM Picture

以下為根據 Grad-CAM Picture 歸納之特徵：

運球：人頭至肩膀兩側

灌籃：較為分散(如場地、燈光等)、手腕

投籃：手肘、腋下



▲ 圖 13：運球

▲ 圖 14：灌籃

▲ 圖 15：投籃

C. 階段討論

本階段如同預期，模型在辨識時較不會因各類別的專屬特徵進行預測，也能關注主體人物動作，而且其測試結果的準確率很高。就結果而言雖然已將各類別的專屬特徵去除，且每一類別都平均含有灰框使其一般化，但仔細觀察仔細觀察測試資料所產生的 Grad-CAM，模型會學到灰框的些許部分，以及部份的圖片仍然會關注籃球進行判斷。

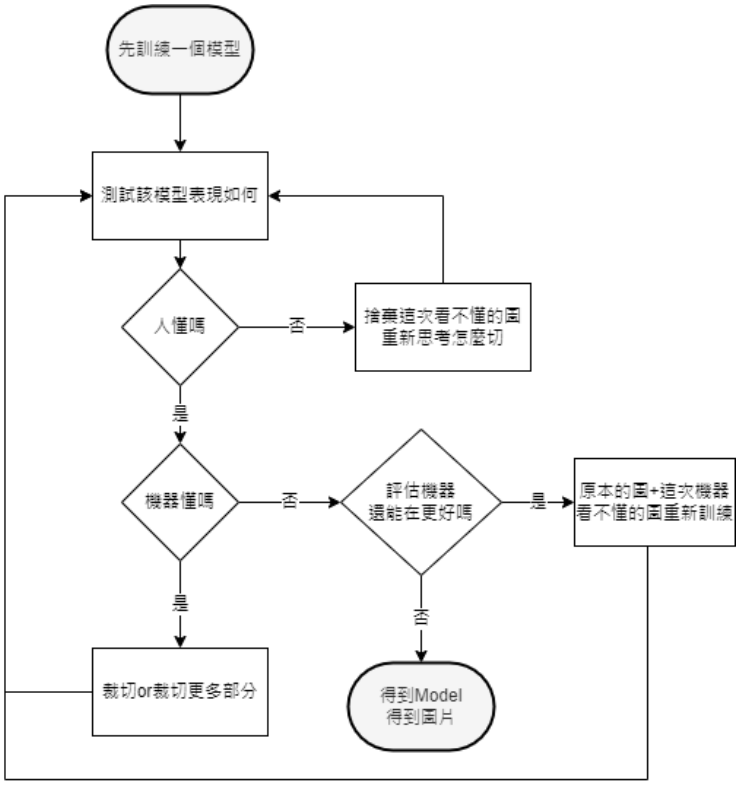
如前述，我們認為本階段目標「主體單一化」能有效的改善模型的辨識，但另一目標「使用灰框遮擋類別的專屬特徵」，如上述並無成功。因此

下一階段僅沿用主體單一化，以其他方式削去圖片中的專屬特徵，迫使模型關注在主體的身體動作，因而發展下一階段。

第三部分、裁切重點部位識別模型

3-1 模型訓練流程

本階段模型訓練的目標為，延續上一階段將主體單一化，且設法使模型關注主體身體動作，因此本階段改用裁剪圖片的方式進行實驗，並依照下列流程圖持續訓練模型，嘗試使模型逼近至人類仍能辨識下的辨認極限，且希望找出足以區分機器與人類的圖片裁切程度，而此程度的裁切於本專題中將以機器與人類辨識的「甜蜜點」代稱。

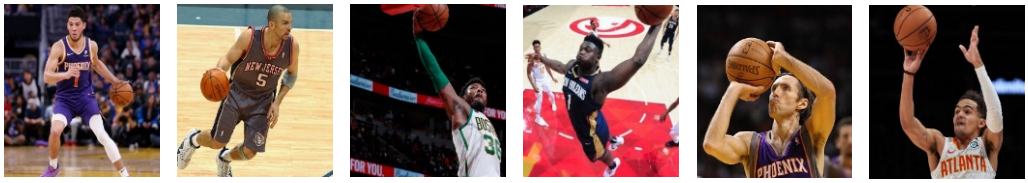


因此本階段會遵循流程圖發展多個版本，直到找出「甜蜜點」。其中，每一版本的測試圖片會參考該版本所產生的 Grad-CAM 圖，歸納機器關注區域所在位置，並作為裁切的依據，目的為測試裁去部分模型關注的區域後，模型是否仍可以辨認成功。以下分別說明進行流程圖的各版本模型與其結果。

3-2 模型簡介

A. Basic

此版本基於把「灌籃」類別中的「籃框」特徵從訓練和驗證集中拿掉，並使籃球球體於各類別皆能存在，實現削去專屬特徵和將籃球元素一般化的效果。其預期結果希望模型以主體身體的部位做辨識依據。



	運球(張)	灌籃(張)	投籃(張)
訓練集	200	200	200
驗證集	25	25	25
測試集	300	300	300

▲ 表 3：訓練資料種類及數目

訓練完後測試之結果：

運球：100%　　灌籃：85.3%　　投籃：95.7%

以下為根據 Grad-CAM Picture 歸納之特徵：

運球：人頭至肩膀一兩側、球

灌籃：手腕、球下緣、場地+燈光

投籃：手臂、手肘、球

由結果得知，已達成專注於主體人物之目的，且模型不再關注籃框區域，因此決定採用此版本作為流程圖的開頭。接下來的版本皆要持續測試受裁切後的圖片，若正確率不佳，則將其加入訓練集重新訓練。

B. Level 1

此版本首先使用如下圖範例的資料集進行測試，該資料集已裁切掉部分模型關注區域，其結果如下表 4。可明顯發現其正確率明顯有待加強，因此加入訓練集重新訓練如下表 5。

	運球(張)	灌籃(張)	投籃(張)
測試集	300	300	300
正確率	63.3%	97.3%	12.7%

▲ 表 4：測試資料數目及正確率

	運球(張)	灌籃(張)	投籃(張)
訓練集	280	280	280
驗證集	70	70	70
測試集	100	100	100

▲ 表 5：訓練資料種類及數目

訓練完後測試之結果：

運球：100%　　灌籃：97%　　投籃：87%

以下為根據 Grad-CAM Picture 歸納之特徵：

運球：人頭至肩膀一兩側、球

灌籃：手臂、球下緣、人頭

投籃：手臂、腋下

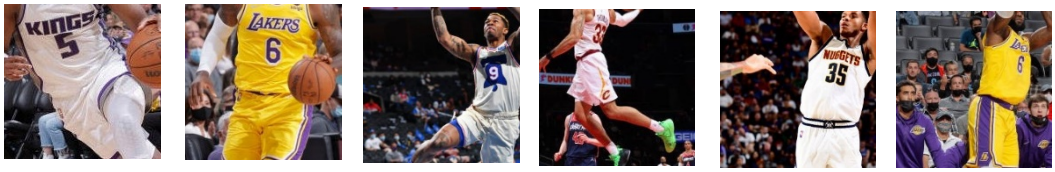
由結果得知，於此階段模型學習如上所述與上一版本比較下來仍相似，因此準確率極高。下一版本會繼續針對模型關注處裁切並進行測試。

C. Level 2

首先根據流程圖，再以各圖片的關注區域進行裁切並測試準確率。如下表 6 所示結果不佳，又發現其正確率仍有待加強，因此加入訓練集重新訓練如下表 7。

	運球(張)	灌籃(張)	投籃(張)
測試集	100	100	100
正確率	20%	97%	48%

▲ 表 6：測試資料數目及正確率



	運球(張)	灌籃(張)	投籃(張)
訓練集	360	360	360
驗證集	90	90	90
測試集	100	100	100

▲ 表 7：訓練資料種類及數目

訓練完後測試之結果：

運球：99%　　灌籃：97%　　投籃：95%

以下為根據 Grad-CAM Picture 歸納之特徵：

運球：胸前的隊名及數字部分、球、手臂

灌籃：手臂、手腕、較為分散(如場地、燈光等)

投籃：手臂、腋下

此版本測試準確率極高，可知於相同層級的圖片是能夠辨識的，其區域仍有空間進行裁切，且目前仍在人類可以辨識的範圍，因此進行下版本的實驗。

D. Level 3

根據流程圖，再繼續使用裁切關注區域後的圖片並測試準確率，其表現仍不佳如下表 8，因將再次進行重新訓練。

	運球(張)	灌籃(張)	投籃(張)
測試集	100	100	100
正確率	51%	91%	67%

▲ 表 8：測試資料數目及正確

	運球(張)	灌籃(張)	投籃(張)
訓練集	440	440	440
驗證集	110	110	110
測試集	100	100	100

▲ 表 9：訓練資料種類及數目

訓練完後測試之結果：

運球：98%　　灌籃：60%　　投籃：65%

以下為根據 Grad-CAM Picture 歸納之特徵：

運球：胸前的隊名及數字部分、手臂

灌籃：人臉、場地背景、手腕

投籃：手臂、腋下、臉

此版本測試準確率僅運球的類別極高，另外兩者皆已較前版本下降許多。此版本所用之圖片已漸漸趨於人類難以辨識的區域，若持續裁切會使得人類會難以辨別，且機器也無法透過訓練再次提升準確率，因此本版本為最終版本。待報告最後做人類辨識與機器辨識率的討論。

3-3 模型訓練結果與討論

對最後版本的結果我們認為，「灌籃」及「投籃」之「甜蜜點」就是把圖片裁切至剩下「部分手臂」及「胸部區域，而本階段所進行的實驗，是以人類可以辨識的情形之下盡可能的去增強模型的準確率，並且於實驗中，我們也同時影響關注的區域，但可以看到運球的測試準確率仍舊很高，代表著運球還是沒有到所謂最佳的「甜蜜點」。我們認為該類別的動作在裁切過後相較其他兩者有較大的差異，所以會較為容易辨認。有這樣的現象原因是本專題起初在設計問題時，並無設想到於裁切後的「灌籃」與「投籃」會有如此大的相似性。

● 實驗結果與未來展望

由上述實驗過程能模擬成以下三種情況來描述攻擊者與防禦者的關係：

情境一、

如同「實驗過程階段一」，若 reCAPTCHA 設計者使用「完整圖片」供用戶點選，而攻擊者則能利用爬蟲下載大量網路圖片作為訓練集，訓練出專門用於破解 reCAPTCHA 的模型當作攻擊手段，而設計者的應對措施則是能把圖片更改為，「去除類別專屬特徵之圖片」作為防禦措施。

情境二、

reCAPTCHA 設計者繼續使用情境一所述之防禦措施，攻擊者則能使用有經過處理後的圖片作為訓練集，重新訓練模型，則如同「實驗過程階段三」之過程，但若設計者改為使用去除大量身體部位的圖片，當機器面對此類圖片時則，依然無法做出正確的判斷。

情境三、

若攻擊者繼續使用去除大量身體部位的圖片再重新訓練模型，就如同「實驗過程階段三」之結論，除了「運球」之外，模型在辨識「灌籃」及「投籃」此兩類別時，似乎已經到了訓練的極限，同時攻擊者也必須付出大量的時間及人力成本；先分析機器關注的區域，再進行裁切及訓練並重複此流程，我們認為對於攻擊者來說代價過高且效益不大，以達到嚇阻的效果。

相對的，當人類辨識去除大量身體部位的圖片時，也無法立刻選擇出正確答案，同時正確率也有下降，我們也另外邀請了系上的同學同樣測試了 300 張

經裁切後的圖片，並與「實驗過程階段三」最終的模型一同比較，結果如下表所示。

	運球	灌籃	投籃	平均
受試人1	99%	77%	94%	90%
受試人2	100%	94%	94%	96%
受試人3	100%	95%	86%	93.67%
受試人4	100%	73%	97%	90%
受試人5	100%	80%	91%	90.33%
受試人6	97%	52%	92%	80.33%
受試人7	99%	68%	86%	84.33%
受試人8	100%	90%	95%	95%
受試人9	99%	98%	89%	95.33%
受試人10	100%	77%	91%	89.33%
受試人11	99%	54%	90%	81%
受試人12	99%	86%	97%	94%
受試人13	98%	72%	92%	87.33%
受試人14	99%	70%	84%	84.33%
受試人15	98%	90%	93%	93.67%
受試人16	100%	74%	79%	84.33%
受試人17	100%	45%	84%	76.33%
受試人18	96%	90%	93%	93%
Model_Fin	94%	64%	70%	76%

▲ 表 10：機器與人類辨認率比較及測試資料演化圖

由上述表格可看出，「灌籃」及「投籃」容易互相搞混，運球則不然。

人類為什麼還能成功辨識圖片呢?

根據組員間的討論及受試者提供的回饋，人類能辨識圖片是因為人類具有能透過分析圖片中各類資訊，比如球員的表情、衣服的皺褶、圖片拍攝的角度、球員所在場景等，推理出適當的猜想，並在腦中自行想像畫面，進而做出正確的判斷。

本專題實驗驗證，去除大量身體部位的圖片可以做為 reCAPTCHA 的防禦機制，且不易被現正流行的圖像辨識攻破。但因為本專題只有三種類別，未來若是能增加同樣是籃球運動中的不同運動，或是增加不同的運動種類，並套用如本專題一樣的實驗過程，相信一定能為這類測試更完整的解釋。

四、參考文獻

[1] Md Imran Hossen, Yazhou Tu, Md Fazle Rabby, Md Nazmul Islam, Hui Cao, Xiali Hei. (2020). An Object Detection based Solver for Google's Image reCAPTCHA v2.

[2] Karl Weiss* , Taghi M. Khoshgoftaar and DingDing Wang. (2016). A survey of transfer learning.

[3] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, Antonio Torralba. (2015). Learning Deep Features for Discriminative Localization.

[4] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra. (2016). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization.

[5] Matthew D. Zeiler, Rob Fergus. (2013). Visualizing and Understanding Convolutional Networks.