6.06.2022

Streszczenie

Uczenie maszynowe, a szczególnie sztuczne sieci neuronowe pozwoliły dokonać przełomu w analizie informacji. Obecnie są wykorzystywane we wspomaganiu decyzji nt. udzielania kredytu, jazdy autonomicznym samochodem, diagnostyce medycznej i wielu innych. Niestety, mimo ich zachwycającej skuteczności stanowią dla użytkowników czarną skrzynkę. Często jedynym wyznacznikiem jakość stworzonego modelu jest ocena poprawności udzielania odpowiedzi. W celu lepszego zrozumienia decyzji sieci neuronowych, a także ich ulepszenia, trzeba lepiej analizować generowane przez nie reprezentacje.

W poniższej pracy proponuję serię metod, które są odpowiedzią na zidentyfikowane braki w procedurach analizy i oceny reprezentacji generowanych przez Głębokie Konwolucyjne Sztuczne Sieci Neuronowe.

Pierwsza grupa metod, to te wspomagające ocenę ocenianego modelu. Oprócz zwykłej skuteczności proponuję ocenę skupienia uwagi sieci, oraz wykrywaniu potencjalnych, ukrytych cech koniecznych, ale nie wystarczających do danej klasyfikacji.

W pierwszej metodzie wprowadzam liczbę ocenę skupienia sieci na obiekcie. Pozwala ona wykryć sytuację, gdy model bierze pod uwagę tło, czyli kontekst, w którym obiekt się znajduje, a ignoruje sam podmiot klasyfikacji. W tym celu wyznaczamy obszar obiektu (tzw. ROI - Region of Interest), a następnie oceniamy stosunek ważności obszaru wewnątrz ROI do całej ilustracji. Jako ważność przyjmuje wartości z mapy ciepła uzyskanej za pomocą wybranej metody wizualizacji - także licznie opisane w tej pracy. Jeżeli uzyskana wartość średnia dla danej klasy jest zauważalnie niższa oznacza to, że klasyfikator w tym przypadku może się skupiać na kontekście, a nie na samym obiekcie.

Druga metoda do automatycznej oceny modelu służy do wykrywania sytuacji, gdy tylko część obiektu jest czynnikiem decydującym o klasyfikacji. Podobnie jak pierwsza polega na wyznaczeniu ROI, a następnie obliczeniu stosunku pola o dużej ważności do pola całego ROI. Uzyskana wartość jest kolejną liczbową oceną poprawności działania utworzonego modelu i jej niska wartość dla poszczególnych klas oznacza klasy, które należy dogłębniej przeanalizować w celu wykrycia niepoprawnej generalizacji.

W przypadku drugiej metody zwykłe techniki wizualizacji okazały się niewystarczające - były zbyt rozmyte by dostrzec skupienie sieci na cechach obiektu, zamiast całego obiektu. W tym celu opracowałem metodę Stopniowego Rozszerzania (Gradual Extrapolation - GE), która zamiast rozszerzać skokowo mapę ważności, uzyskaną w głębokiej warstwie sieci neuronowej, rozszerza się warstwa po warstwie. Dodatkowo po każdej ekstrapolacji uzyskana mapa jest mnożona przez średnie wartości aktywacji w danej warstwie. Ta procedura z pierwotnie rozmytych, niewyraźnych map, tworzy szczegółowe ilustracje, znacznie wyraźniej oddające kształty obiektu lub elementu, który zdominował klasyfikację. Ponadto metoda GE jest kompatybilna z niemal każdą techniką wizualizacji wykorzystywaną w literaturze.

W celu udowodnienia skuteczności metody GE opracowałem i zarekomendowałem w kolejnej publikacji procedure oceny metod wizualizacji. Pozwala ona liczbowo porównać obecne jak i przyszłe metody, co ma pozwolić na lepszą ich systematykę. Test opiera się na trzech kryterach: wiarygodności, interpretowalności i aplikowalności.

Wiarygodność jest obliczana jako współczynnik utraty poprawności klasyfikacji po usunięciu odpowiednio 1%, 2%, 5%, 10%, itd. najważniejszych pikseli wskazanych wg ocenianej metody. Im szybciej pewność klasyfikacji spada, tym bardziej precyzyjna jest wskazana metoda.

Szandała

Interpretowalności jest definiowana jako odpowiedź na pytanie: która metoda najbardziej zawęża informacje. Po porównaniu wejściowego obrazka z uzyskaną mapą ważności wycinamy piksele, które uzyskały ważność bliską zero. Im mniej pikseli pozostanie, tym wyżej w rankingu plasuje się analizowana metoda.

Ostatnim czynnikiem jest aplikowalności, czyli sprawdzenia, czy daną metodę wizualizacji da się zastosować do wielu modeli w skończonym czasie, przy ograniczonych zasobach obliczeniowych. W celu analizy aplikuje się tę samą procedurę wielokrotnie dla różnych modeli, szczególnie tych dobrze opisanych w literaturze. Podczas przeprowadzania testu, oprócz stosowalności metody z danym modelem należy zwrócić uwagę na czas potrzebny do uzyskania wyniki oraz niezmienność wyniku między wykonaniami dla tego samego przypadku. Warto zwrócić uwagę, że ta część testu nie jest decydująca - istnieją przypadki, gdzie metoda sprawdzi się tylko dla konkretnego modelu, albo wynik metody może się różnić między wykonaniami (np. LIME). Nie dyskwalifikuje to metody, ale wymaga by adnotacja o tym znalazła się w opisie techniki.

FIA-test (Faithfulness, Interpretability and Applicability) jest próba usystematyzowania prac dotyczących wyjaśnialnej sztucznej inteligencji, a zwłaszcza technik wizualizacji - generatorów map ważności.

Ostatnią metodą zaproponowaną w dysertacji jest Mapowanie Głównych Składowych (Principal Image Sections Mapping - PRISM). Polega ona na zastosowania Analizy Głównych Składowych dla reprezentacji wygenerowanej przez dany model. Metodę należy wykonywać jednocześnie dla grup obrazów. Uzyskana macierz Głównych Komponentów może być wykorzystana na 2 sposoby. PIerwszy polega na przypisaniu trzem pierwszym komponentom kolorów czerwony, zielony i niebieski, co pozwala na wizualne porównanie najważniejszych cech występujących na danych obrazach. Dodatkowo, połączenie PRISMa z GE pozwala uzyskać ilustracje prezentujące rozmieszczenie cech wykrytych przez model na obrazach. Procedura ta jest doskonałym uzupełnieniem metody Explanation by Example, gdzie staramy się dobrać obrazy podobne do badanego i wyciągnięciu wniosków o przyczynie danej klasyfikacji. Z PRISMem wskazanie cech wspólnych jest znacznie łatwiejsze.

Oryginalny wynik PRISMa może też posłużyć do masowej analizy wielu klas. Uzyskane Główny Składowe mogą zostać użyte jako dane wejściowe do dowolnej metody klasteryzacji. Po rozmieszczeniu poszczególnych obrazów w zadanej przestrzeni i utworzeniu klastrów można dostrzec nachodzące na siebie klastry. Takie obszary sugerują, iż wskazywane klasy posiadają podzbiór cech podobnych i potencjalnie mogą być błędnie sklasyfikowane przez badany model.

Metody i procedury zaproponowane w poniższej pracy są uzupełnieniem obecnie wykorzystywanych technik. Dodaję nowe kryteria oceny jakości modelu, oraz pozwalają wykryć newralgiczne klasy, które wymagają szczególnej uwagi. Mam nadzieję, że choć część z nich trafi do standardowych narzędzi wykorzystywanych przez inżynierów sztucznych sieci neuronowych.

Szandała

6.06.2022

Abstract

Machine learning and artificial neural networks have created an unprecedented breakthrough in data analysis. Nowadays they are used as decision reinforcements in loan decisions, autonomous cars driving, medicine and many more. Unfortunately they often are just a blackbox to their users. The main criteria for a model's acceptance is its accuracy in solving a given problem. In order to better understand their reasoning and thus improve performance a new trend arose: Explainable Artificial Intelligence. For Deep Convolutional Neural Network (DCNN) it focuses on studying and analyzing the representation generated by the model.

In this dissertation I am proposing several methods that are an answer to the gaps identified in interpretation and validation of representation generated by DCNNs.

The first group of methods aim to reinforce the model's evaluation. Apart from general accuracy of the classifying model I propose to evaluate its attention focus, whether it is in spurious correlation with context or focuses on the latent feature of the actual object.

The first approach I am introducing is a new quantitative metric of how much the model focuses on the object. It allows to identify circumstances where DCNN concentrates on the context of the object, instead of the classified item. We start by locating the Region of Interest(ROI) – a rectangular area around an interesting object. Next we generate a saliency map - using one of the techniques like GradCAM, also described in this paper. Finally we calculate the ratio of saliency inside ROI, divided by the sum of saliency in the entire image. If the obtained value, average for class, is significantly lower than other it may indicate that model is focusing on the context not at the object itself.

Second method to an automatic model's evaluation is the case when the model takes into consideration only a small part of the entire object - like wheels for a car. Similar to the previous one we draw the ROI around interesting objects and obtain saliency values for the image. Finally we divide saliency inside the ROI by the area of the ROI. If the value is noticeably lower for a certain class it indicates a potential latent discriminative feature is present and may incur faulty classification in future by breaking the model's generalization.

For the second method there was no visualization technique sharp enough to see the shape of the network's focus. For the case I have come up with a technique called Gradual Extrapolation (GE), which instead of directly resizing saliency from generated representation to the input image size, extrapolates layer by layer. Moreover after each extrapolation to the preceding layer we are multiplying the obtained matrix with average activation in the given layer. This procedure results in significantly sharper saliency maps from which we are usually able to recognize shapes of salient areas. Furthermore, the devised method is compatible with most of the other state-of-the-art visualization techniques.

In order to prove the value of the proposed GE method I have defined and recommended in another publication a set of tests for visualization techniques. It allows us to quantitatively compare any method that results in a saliency map image. It is based on three criteria: faithfulness, interpretability and applicability.

Faithfulness is calculated as a factor of accuracy drop after removal of 1%, 2%, 5%, 10% etc. of the most important pixels highlighted by the analyzed method. The faster the accuracy/confidence of classification drops, the more precise the method is.

Interpretability is defined as the answer to a question: which method reduces the amount of information the most. After comparison of the initial image with the saliency map we are removing

Scondola

pixels which had saliency below the chosen threshold. The less pixels are left, the more effective the analyzed method is.

Final factor is the applicability, which means evaluation of whether a given method is applicable to the most of the state-of-the-art models in a finite time, with finite computing resources. To perform this test one has to perform several analyzed method on a set of different models. After that a time used for computation should be identified as well as the invariance of the result between tests for the same instance. Note that this test is not deterministic - there are methods that are applicable only to specific cases as well there are valuable methods that are slightly volatile, like LIME. Failure in the test does not disqualify the method but requires a clear statement of the method's limitations.

FIA-test (Faithfulness, Interpretability and Applicability) is an attempt to order works about visualization, mainly the techniques that result in a saliency map.

The final method from this dissertation is a Principal Image Sections Mapping (PRISM). This technique relies on performing Principal Component Analysis on a layer that generates the final representation. This method should be executed for a batch of images. Obtained matrix of Principal Components (PC) can be utilized in two ways. First is to focus only on 3 (usually) first PCs and assign them colors, respectively: red, green and blue. This allows users to visually compare features present in an image identified by the model. Moreover, the combination of PRISM and GE results in illustrations that can be used even by a non-technical user to identify exclusive and inclusive features. This method is an excellent addendum to the Explanation by Example, where we try to find pictures similar to the analyzed ones and then make conclusions on the model's reasoning. With PRISM it is much simpler.

Additionally the PRISM's original output may serve as an input for multiple class analysis for chosen clustering methods like Self Organizing Maps. After depicting each instance in a finite space and assigning them to a cluster we can identify overlapping clusters. These clusters indicate that respective classes have subset of similar features and thus can be misclassified. Similar to the two first methods: they require closer insight.

Methods and procedures proposed in the paper are supplementary to the state-of-the-art techniques. I am adding new criteria to DCNNs evaluation as well as tools to identify classes that require deeper analysis in order to create more robust models. I hope that some of my findings will find a place in a regular DCNN practitioner's toolbox.

Szandała