Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization

備考

著者

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra

掲載

2017 IEEE International Conference on Computer Vision (ICCV), pp. 618-626, 2017.

Abstract

我々は、CNN（Convolutional Neural Network）ベースのモデルをより透明化するために、これらのモデルからの予測にとって「重要」な入力の領域を視覚化する技術、すなわち視覚的説明を提案する。

Grad-CAM (Grad-weighted Class Activation Mapping) と呼ばれる我々のアプローチは、CNNの最終畳み込み層に流入するクラス固有の勾配情報を用いて、画像中の重要な領域の粗い局在マップを生成する。Grad-CAMはClass Activation Mapping（CAM）[43]を厳密に一般化したものである。CAMが狭いクラスのCNNモデルに限定されるのに対し、Grad-CAMはあらゆるCNNベースのアーキテクチャに広く適用可能である。また、Grad-CAMを既存のピクセル空間可視化（Guided Backpropagation [38]など）と組み合わせ、高解像度クラス識別可視化（Guided Grad-CAM）を作成する方法を示す。

我々は、画像分類、画像キャプション、視覚的質問応答（VQA）モデルをより理解するために、Grad-CAMとGuided Grad-CAMの視覚的説明を生成する。また、ILSVRC-15の弱教師付きローカライゼーション課題において、画素空間勾配の視覚化（Guided Backpropagation [38] and Deconvolution

[41] ）に優る性能を示した。また、画像キャプションとVQAについては、一般的なCNN＋LSTMモデルは、下地画像-テキストペアで学習していないにもかかわらず、しばしば識別可能な入力画像領域を局所化するのに適しているという、やや意外な洞察を明らかにすることができた。

最後に、Guided Grad-CAMの説明によって、ユーザーがディープネットワークによってなされた予測に対する信頼を確立できるかどうかを測定するために、人体実験を計画し実施する。興味深いことに、我々はGuided Grad-CAMが、訓練を受けていないユーザーが、両ネットワークが同じ予測をした場合でも、単にその異なる説明に基づいて、「強い」ディープネットワークと「弱い」ネットワークを見分けることに成功することを示す。