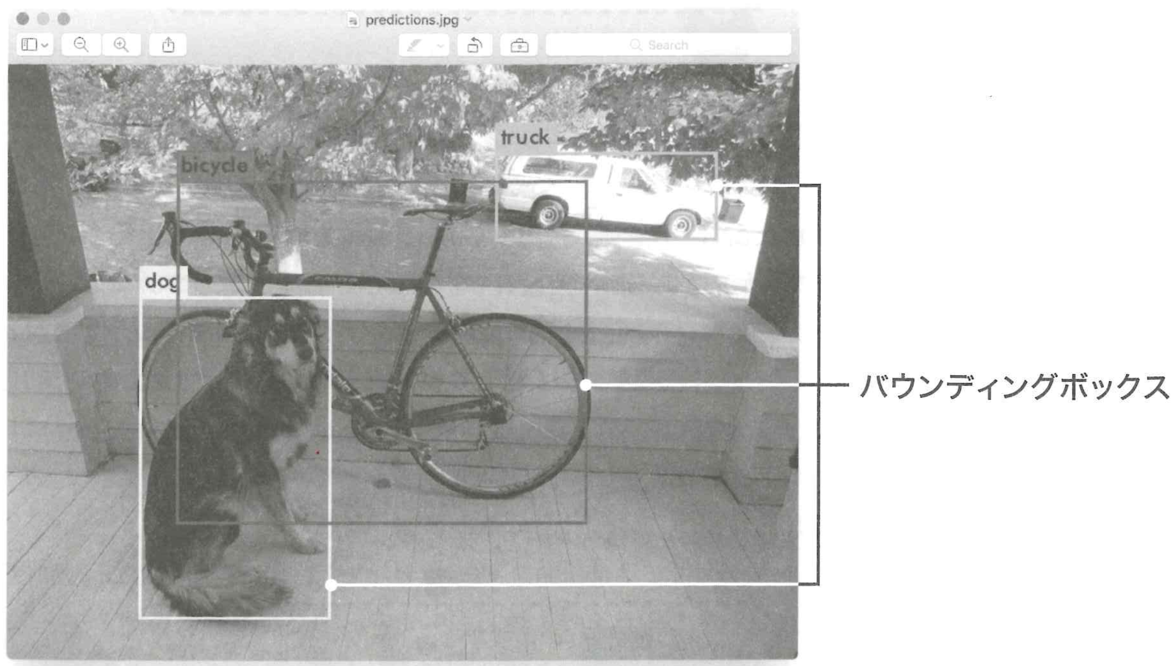
(2) 主な技術

ここでは、「眼の発見」といわれるほど認識精度を向上し、その後様々な認識処理の飛躍の元になった、ディープラーニングを使った様々な認識技術について紹介する。

物体認識

YOLO (You Only Look Once) 技術は自動運転における道・歩行者・車などの物体認識に使われている (**図2-2-4**)。YOLOは、2016年に発表された物体認識アルゴリズムで、領域判定 (物体があるらしい領域の提案) と対象分類 (物体がどのクラスに分類できるか) を同時に行うことで物体認識処理を高速化した。YOLO以前のアルゴリズムは、画像から物体を含む複数のバウンディングボックスを切り出し、それぞれに含まれる対象のクラス分類を行っていたため、バウンディングボックスの数が多いと膨大な時間を要した。YOLOでは、バウンディングボックスの切り出しと同時に、画像全体をS×Sの格子状の領域に分割し、それぞれの領域に対してオブジェクトが含まれているならばどのクラスに分類されるかの確率を計算し、これらを掛け合わせることで判定している。これにより、バウンディングボックスの切り出しとその中に含まれるクラスの分類結果を一度に計算できる。

■図2-2-4　**YOLOの物体検知例**



出典：YOLO: Real-Time Object DetectionのWebページより※3

画像全体を学習対象とするため、領域を分割して認識を行うR-CNN※4 (Regional Convolutional Neural Network) と比較すると、誤認識 (背景と見なすべき部分を物体と判断するなど) が少なくなっている。

※3 https://pjreddie.com/darknet/yolo/
※4 R-CNNはディープラーニングのネットワーク構造の一つ、空間認識に優れているとされるCNN (Convolutional Neural Network) を分割された領域ごとに適用する (詳細は2.3参照)。

行動認識

人などの行動を撮影した動画などから、何をしているのかを認識することを「行動認識」という。行動認識は、物体認識や画像認識とは違い、時間軸に沿ったデータを判定する必要があり、歩行ならば、足を上げる、手を上げる、踏み出すなどの、一連の動作を認識する必要がある。このため、動画を2次元の静止画と時間軸で考える「Spatio-temporal」(時空) を対象として認識を行う必要がある。そのためのものとして、時空を対象とした畳込みニューラルネットワーク (Spatio-temporal Convolutional Neural Network、以下Spatio-temporal ConvNet) がある。これは、動画の10フレームを畳込みニューラルネットワーク (CNN※5) に入力するものであり、行動認識を可能にした。その後登場したTwo-stream CNNは、フレーム画像を認識するニューラルネットワークであるSpatial stream ConvNetと、動きの情報を認識するニューラルネットワークであるTemporal stream ConvNetの2つのCNNを組み合わせ、行動認識で高い精度を実現した。

こうした行動認識でディープラーニングを利用するには、動画から動作部分のフレームを切り出し、ラベル付けされた行動の動画データセットが大量に必要となる。こうしたデータセットのうち、最大級のものがDeepMindの公開するKinetics※6である。Kineticsは、65万のビデオクリップから700の人間の動きを切り出し、ラベル付けしたデータセットである。これを使ってつくられたのがDeepMindのI3D (Inflated 3D ConvNet) である。画像のxy空間を時間方向に拡張 (inflated) した3次元畳込みを利用することを特徴とし、Kineticsで事前学習させることで、従来の時系列向けのニューラルネットワークであるLSTM※7 (Long Short-Term Memory) やTwo-Stream CNNと同等の精度を確保できることを示した。なお、I3Dで開発した3次元畳込みは行動認識以外、例えば立体の認識へも応用可能であり、DeepMindは、I3DをCT/MRなどの3次元の医用画像を基にした画像診断に応用し、「肺がん検診AI」として、人間の放射線科医に匹敵するか上回る精度で、早期がんを見つけ出すことができると発表した※8。

音声認識

音声認識とは、音素の時系列データから単語を同定する問題と考えることができる。古くからの研究分野であり、従来は、確率的に遷移する内部状態に応じて波形が生成されることをモデル化した隠れマルコフモデル (HMM；Hidden Markov model) とHMMの各状態の音響特徴量のパターンを連続分布でモデル化する混合正規分布 (GMM；Gaussian Mixture Model) との組み合わせなどが研究されてきた。

GMMの代わりに、ディープニューラルネットワークとHMMとを組み合わせた場合は、ディープニューラルネットワークが特定の音響特徴量分布を仮定しないため、GMMとHMMの組み合わせより高い精度が出るようになった。

また、音声認識は、音素の時間軸上の組み合わせを処理する必要があることから、再帰型ニュ

※5 CNN(Convolutional Neural Network)はディープラーニングのネットワーク構造の一つ、空間認識に優れているとされる(詳細は2.3参照)。
※6 https://deepmind.com/research/open-source/kinetics
※7 LSTMはディープラーニングのネットワーク構造の一つ、時系列データの処理に優れているとされる(詳細は2.3参照)。
※8 "A promising step forward for predicting lung cancer" <https://www.blog.google/technology/health/lung-cancer-prediction/>