Classificatie van afbeeldingen met geautomatiseerde machine learning platformen

Decorte Robbe, dr. Helsens Kenny, ir. Decorte Johan
Hogeschool Gent, Valentin Vaerwyckweg 1, 9000 Gent
robbe.decorte@student.hogent.be

Abstract

Deze technologie tracht *machine learning* toegankelijker te maken door een manier te bieden om vaak voorkomende problemen te automatiseren. Dit werd onderzocht door met AutoKeras en Google Cloud AutoML elk een prototype op te zetten dat voor een simpel maar realistisch classificatieprobleem de categorie van een afbeelding kan voorspellen. Er werden modellen getraind die katten van honden kunnen onderscheiden. Dit document beschrijft een studie naar de achterliggende gebruikte technieken, het verloop en de resultaten van beide prototypes. Er werd gevonden dat de alternatieven elk hun plaats hebben in verschillende fasen van een project. Google Cloud AutoML levert een productie waardig model terwijl AutoKeras kan dienen als hulpmiddel voor een *data scientist* of productie waardig kan zijn mits een extensieve voorbereiding van de data. De evolutie van de platformen zelf betekent enkel goed nieuws voor de toekomst. Mogelijks kan er nog onderzocht worden hoe het opschonen van de data geautomatiseerd kan worden. Dit is een grote stap binnen geautomatiseerde *machine learning* aangezien het een belangrijke factor is om *edge cases* te herkennen.

Introductie

Het informatie tijdperk centraliseert zich momenteel rond data. Bedrijven zien ook de toegevoegde waarde die het kan hebben in hun bedrijfsprocessen, kijk maar naar de grote spelers in de informaticawereld waar het begrip *Big Data* is ontstaan. Eén van deze toepassingen die intensief gebruik maakt van data, is *machine learning*. Mensen die goed overweg kunnen met de data om zo'n model te maken (Machine Learning Engineers / Data Scientists ...) zijn vaak moeilijk te vinden. Een werkgever die zo'n probleem aan wilt pakken heeft enkele keuzes, AutoML is een mogelijke optie. Alhoewel het interesseveld ontstaan is in de jaren '50, is het nog maar sinds kort een hot topic, met dank aan de grote hoeveelheid rekenkracht in moderne systemen en doorbraken binnen het onderzoeksveld die de toegangsdrempel verlagen.

Geautomatiseerde *machine learning* platformen trachten een oplossing te bieden voor *development* teams zonder een gespecialiseerde *machine learning* expert. Het platform voert alle stappen van het proces uit en uiteindelijk moeten ze het enkel in hun product integreren. Door de technische afhankelijkheid te verlagen kan de technologie sneller / meer gebruikt worden in bestaande projecten. Omdat bedrijven tot nu toe weinig contact hebben met AI en alles wat er toe behoort, zijn de meeste cases vergelijkbaar met elkaar. Zo heb je bijvoorbeeld binaire classificatie problemen, tekst analyse en meer. Waarom zou het dan niet mogelijk zijn om dit te automatiseren?

Experimenten

Figure 1: Voorbeelden van afbeeldingen uit de dataset die gebruikt werd om de modellen te trainen.

Om de werking van AutoKeras en Google Cloud AutoML te onderzoeken werd voor beide enkele modellen getraind dat voorspeld als het object op een afbeelding een kat of een hond is. Figuur 1 bevat twee afbeeldingen uit de gebruikte dataset. Het experiment blijft realistisch door de afbeeldingen niet te normaliseren (i.e. ook afbeeldingen gebruiken met ruis, zoals het meisje op de eerste foto).

Beide werden beoordeeld op de implementatie van het proces model uit figuur 2 alsook op extra (niet-) functionele *requirements*.

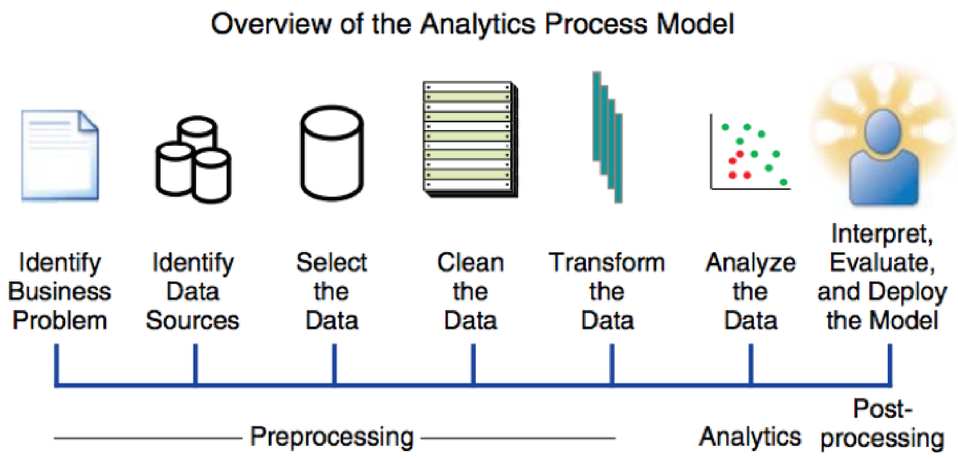

Figure 2: Afbeelding via Lemahieu, W., Broucke, S. V. & Baesens, B. (2018). Principles of Database Management: The Practical Guide to Storing, Managing and Analyzing Big and Small Data. USA, Cambridge University Press.

In het geval van AutoKeras kan er achteraf bekeken worden welke afbeeldingen verkeer voorspeld werden. Figuur 3 bevat negen van die gevallen. Door deze te bekijken is het mogelijk om een inkijk te krijgen in het model. Zo kom je mogelijks te weten waaraan het model gevoelig is, in dit geval zie je snel dat honden met gespitste oren soms als een kat voorspeld werden. Google Cloud AutoML heeft geen diepgaande evaluatiemogelijkheden zoals AutoKeras.

Figure 3: Voorbeelden van afbeeldingen die verkeerd geclassificeerd zijn. Bij elke afbeelding staat de zekerheid van de voorspelling (P, [0,0.5[= hond,]0.5,1] = kat) en de verwachte klasse (e).

Conclusies

Beide systemen komen de verwachtingen na maar moeten op de juiste plaats ingezet worden. Zo is Google Cloud AutoML een volwaardig *drop in replacement* in bestaande applicaties. Het proces kan niet eenvoudiger zijn en de verschillende manieren om het te integreren zorgen ervoor dat het in meeste situaties past. AutoKeras, in zijn huidige staat, is niet verfijnd genoeg om productie waardig te zijn. De extra moeite om de eerste stappen van het procesmodel te verbeteren kan evengoed verwisseld worden met een ML-ingenieur die het volledige proces uitvoert. Die niche kennis blijft noodzakelijk om te slagen. Anderzijds blijkt het wel een goede *tool* te zijn in de gereedschapskist van ML-ingenieurs.

AutoKeras is sterk gericht op de *core* van het probleem en zo blijven stappen vooraf en achteraf onbeantwoord terwijl die bij de werkwijze van Google Cloud AutoML een aanzienlijke rol hebben. Zo kan een model getraind en *deployed* zijn in een vijftal muisklikken, geen vooraf verwerkte afbeeldingen of andere zaken nodig. Dit terwijl AutoKeras pas gebruikt kan worden nadat de afbeeldingen omgezet zijn naar ruwe data, correct geschaald zijn en grijsfilters of andere optimalisaties toegepast worden. Achteraf is het de verantwoordelijkheid van de gebruiker om het model online te krijgen, wachtrij te optimaliseren en een interface te hebben die kan communiceren met het model.

Toekomstig onderzoek

De toekomst van geautomatiseerde *machine learning* ziet er alvast goed uit. De verbeteringen tussen versies van AutoKeras vallen op en ook steeds meer cloud platformen bieden een gelijkaardige service aan. Er is een echte *push* aan de gang, van de *community* en de bedrijven, om de toepassingen toegankelijker te maken. Verder onderzoek over dit onderwerp zou zich kunnen richten op individuele stappen van het procesmodel, bijvoorbeeld de *data preprocessing*. De automatisatie ervan is niet vanzelfsprekend omdat dit voor elke dataset anders is.