**Notulen meeting Maaike**

**Datum: 16 maart 2016**

**1. Hoever zijn we nu?**

**Vertellen:**

- Caffe heeft veel moeite gekost, maar het werkt. We kunnen nu voor 1 image de feature vector genereren.

- We kunnen dingen op de cluster laten draaien, via jobscripts.

**Vragen:**

- features laten zien 🡪 vragen of het er goed uit ziet. Kloppen ze?

- caffenet draaien op preprocessed images of op raw images?

*Misschien dat hij het automatisch cropt. Bij verschillende formaten, kun je gemakkelijker overfitten op de traindata. Even kijken hoe hij het cropt. Het zou niet moeten uitmaken hoe groot de plaatjes zijn, zolang ze maar even groot zijn allemaal. Moeten de plaatjes dan ook zo groot mogelijk? Nee niet per se, soms geven kleinere plaatjes een betere performance, want plaatjes moeten anders groter gemaakt worden 🡪 geeft ruis. Kleine plaatjes groot maken is niet handig.*

**2. Wat is ons plan?**

**Vertellen: ons plan:**

- per image een feature vector

- voor alle feature vectoren van de trainset clusters maken met K-means

- voor alle feature vectoren van de testset de images in de train-clusters indelen. Hiervoor Gaussian Mixture Model/EMD gebruiken, zodat ieder image in meerdere cluster kan zitten “soft assignment”.

- per business komt hier een vector uit: nr of clusters x 1.

- hierop SVM of Deep Learning toepassen voor classificatie.

**Prioriteiten:**

- features genereren voor alle images

- twee mensen aan SVM werken. Als dit gelukt is, Deep Learning.

- twee mensen aan K-means clustering, en genereren van cluster-features per business.

**Vragen:**

- feedback of tips over ons plan?

- We denken eraan de nr of clusters x 1 vectoren per business te normaliseren, zodat je alleen de verdeling over de clusters hebt. Is dit goed? Zijn er nadelen?

- andere ideeën om de feature vectors samen te voegen?

*SVMs kunnen ook meerdere classes categoriseren. Zoek naar het type. Vraag: hoe weet je dan hoe goed de clusters zijn? Is dat deterministisch? Het is moeilijk om te testen hoe de clusters de score beïnvloeden.*

*Verwacht je dat de clusters per bedrijf overeen gaan komen?*

*Er is een risico dat de clusters niet nuttig worden voor de classificatie. Bv. cluster met alcohol, cluster met voedsel. Stel dat iedere business een plaatje heeft in iedere cluster. En dan?*

*Probleem: niet zeker dat de clusters zo zijn dat je er het meest aan hebt voor je classificatie.*

*Wordnet voegt dat iets toe? Kan wel. De labels betekenen wel wat. Maar SVM leert zelf wat belangrijk is, dus dat hoef je niet te doen.*

*SVM wordt wel vaak gebruikt, werkt wel goed. Je kunt hiervoor sklearn gebruiken.*

*Geen neural netwerk, want te weinig data. 2000 businesses.*

*Meerdere machine learning technieken . Als je weinig weet van deep learning, vooral niet doen, want te veel parameters die getuned moeten worden.*

*Ja we moeten normaliseren!*

**3. Algemeen**

Vragen:

- Hoe is de vooruitgang? We twijfelen of het snel genoeg gaat, maar werken keihard. We verkrijgen weinig resultaten, doordat we steeds vastzitten bij Caffe, of de server/cluster aan de praat krijgen.

*Moeilijk te zeggen. Kan mailen als je ergens vast zit, zodat het in ieder geval bekend is. Een submissie, een idee, voor nu gaat het goed. Ook mailen: wie wat gedaan heeft. Wat zijn de taken.*

Volgende week:

donderdag is Maaike in Nijmegen. Om 14 uur weer in Mercator.

Doen:

- Maaike toevoegen aan git.