מכיוון ששם פרוייקט דל-לא מספיק תאורי יש לפחות
להוסיף תת כותרת המתארת את התחום והפעילות

30th November 2018

חסר כותרת המחלקה
והפקולטה

FACE2VEC

Preliminary Report

חסר מספר הפרוייקט
כותר הפרוייקט צריכה להיות
במרכז השורה
חסר לוגו המחלקה

ראה סיכום
הערות
בסוף הדוח

Authors:

xxxxxxxxxxxx xxxxxxxxxxx

Advisors:

xxxxxxxxxxxx xxxxxxxxxxx

In Cooperation with:

xxxxxxxxxxxx xxxxxxxxxxx

xxxxxxxxxxxx xxxxxxxxxxx

נבדק ללא ציון
נא לתקן הערות במסגרת
הגשת דוח התקדמות

TABLE OF CONTENTS

אותיות בתוך טבלה בפונטים גדולים
מידי
מעיד על דלות החומר
חסר מיספור לכותרות ולפרקים

TABLE OF FIGURES

כול ציור בטבלה צריך
להכנס בשורה אחת כולל
מספר העמוד

30th November 2018

PROJECT DEFINITION

BACKGROUND & MOTIVATION

Facial Expressions are a crucial aspect of a human’s body language, and thus interpersonal communication. By (consciously or subconsciously) manipulating facial muscles into various positions (thus creating an expression), a human can convey information that is central to our perception of his emotional state, intentions, mental state/condition, etc. Naturally, due to this degree of importance, it is a subject of great interest and extensive research, particularly in the discipline of psychology.

As technology develops, so too do our possibilities to use analytical tools to explore the world of facial expressions, by training computers to analyze massive amounts of facial expression data, and creating models based on that data. Creating a ‘good’ model would give us, as researchers, an opportunity to reach new conclusions regarding facial expressions, such as the relation and transition between them, and could contribute in the long term to develop technologies that could ‘understand’ the emotions in a person’s face, in a far more reliable and profound way than currently exists.

מספרי עמוד באנגלית לא
יכולים להיות בצד שמאל -
או במרכז או מצד ימין

סגנון הכתיבה שלכם הוא
סגנון סיפור ולא סגנון דוח
המדווח.
נא לא להשתמש בגוף
ראשון או שני יחיד או
רבים!

RESEARCH SUBJECT

What if we told you your computer could recognize your emotional state? How would it know how to do that?
That is the problem we are trying to solve.
The computer would need a ‘language’ that it could comprehend in order to correctly map what it ‘sees’ (your face), to what it tells you it sees (your emotion/expression). Since this is a computer we are talking about, it would need to be a mathematical model. How could we create such a model? According to our research hypothesis, we believe the answer may lie in the world of natural language processing (NLP), and specifically the word2vec method of creating a model that translates words from a corpus built of sentences into vectors, where ‘similar’ words (contextually similar) would have ‘similar’ vectors (similar direction/size in the vector space).

שם בגוף הדוח כמו למשל שם תוכנה חייב להתחיל באותיות גדולות

As you may have noticed, word2vec works on a corpus of words and by using the context of words in sentences, yet our research subject is of facial expressions, not words nor sentences. Our proposition is to take a large number of videos containing facial expressions (such a database of videos has already been created by the Dept. of Cognitive Science), and analyzing them frame-by-frame. Thus, a frame is our ‘word’ and a series of frames (section of video) is a ‘sentence’. We intend to use an existing and established tool called ‘FaceReader’, which is a facial analysis tool, to deconstruct the faces in the frames into ‘words’ that we can work with in our word2vec model. From the aforementioned proposed method we intend use, we derived the name for the project - ‘Face2Vec’.

PROJECT GOAL

Our ultimate goal is to arrive at a model that can reasonably detect human emotions from input video of a face.

More reasonably, we believe that we can arrive at a model that at the very least contains demonstrably valuable information regarding the distribution of emotions and expressions, and the transitions between them. As mentioned previously as motivation, such information could greatly benefit any future research and development of facial expression analysis tools for a multitude of applications (psychology, security etc.).

אין פה עניין של אמונה. זה
דוח מדעי. אפשר לדבר על
ניסוי ותהיה

אין התייחסות תאורתית
למה זה רגשות, כיצד
מכמתים (מלשון כמות),
כיצד מודדים רגשות, מה
הבעיה במידדת רגשות
ובכלל איך הופכים רגש
למשהו מטמטי?
מציע להתייחס לנושא מול
המנחה

לא מצאתי התייחסות,
למבנה הנתונים, תכולה
גודל, יחידות, פורמט

לא רואה תאור של
המעבדה שלכם, כלי
הניסוי, אופן המדידה
המעשי וכ"ו.

THEORETICAL BACKGROUND

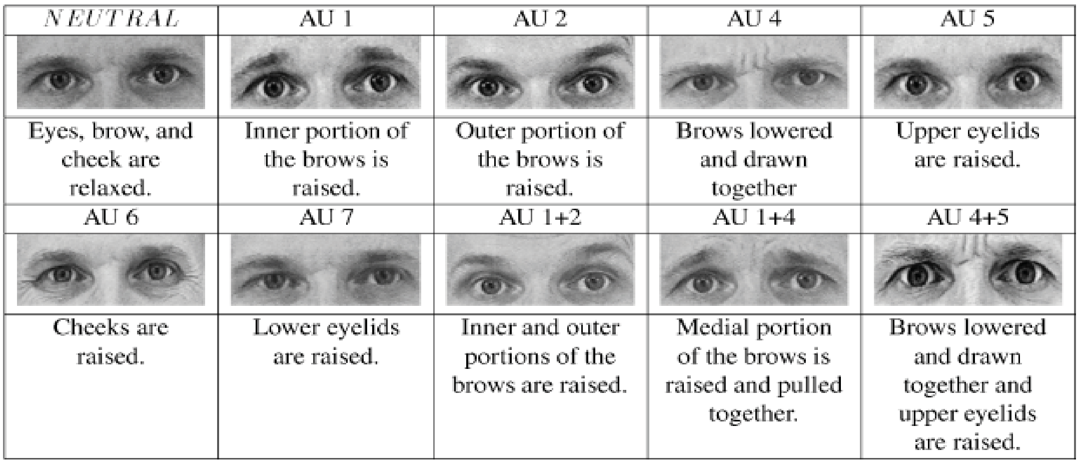
LITERATURE OVERVIEW

The research to which our project belongs is of a multidisciplinary nature, and has literary background in both the subjects of emotion recognition & classification, and computer science. While our focus, as CSE students, leans towards the computer science aspect of the research, it is of high importance to the success of our project that we have a good understanding of the existing concepts and conventions of emotion recognition & classification from the discipline of cognitive science. Thus, literature from both cognitive science & computer science shall be included in this overview, starting from the former.

Emotion Recognition & Classification Literature Overview

From our study of the existing literature on the subject, we have learned that there is a widely referenced convention that there are 6 'basic' emotions that are expressed in a similar fashion by all humans, and they are: anger, fear, happiness, sadness, disgust and surprise (Ekman & Friesen, 1971). Methods to detect these emotions, or to describe an emotion on a face as a linear combination of these emotions already exist, such as FaceReader (which we are going to use), who report an accuracy over these emotions of 89% (Uyl & Kuilenburg). For true emotion recognition, this resolution of emotions is considered by many to be inadequate and incomplete (Corardo & Keltner, 2015). Our aim is to try to reliably describe emotions with a much increased resolution compared to the 6 'basic' emotions convention.

A widely used method to analyze a facial expression is knowns as FACS (Facial Action Coding System), which breaks down the face into 30 ‘Action Units’. An Action Unit is a group of muscles on the face, and the facial expression can be described as a combination of which Action Units are active (Tian, Kanade, & Cohn, 2001). See *Figure 1* below for several examples. This method is used by FaceReader, and it can output its analysis in the form of Action Units, as specified by FACS. We will use this output as the ‘words’ for our word2vec neural network.



NEUTRAL	AU 1	AU 2	AU 4	AU 5
Eyes, brow, and cheek are relaxed.	Inner portion of the brows is raised.	Outer portion of the brows is raised.	Brows lowered and drawn together	Upper eyelids are raised.
AU 6	AU 7	AU 1+2	AU 1+4	AU 4+5
Cheeks are raised.	Lower eyelids are raised.	Inner and outer portions of the brows are raised.	Medial portion of the brows is raised and pulled together.	Brows lowered and drawn together and upper eyelids are raised.

Figure 1: Examples of Action Unit activation. (Example localized to the eye region)

Word2Vec Literature Overview

Word2Vec is a type of neural network that, given a large corpus of text, creates a vector space with a pre-defined, usually very large dimension. Each word in the corpus is translated into a vector in the vector space, in such a way that two words that have been found to have similar contexts within the corpus (appear in similar contexts in sentences that appear in the corpus) they will have vectors that are in close proximity. (Mikolov, 2013) & (Coyler, 2016). Word2Vec has established itself as at the forefront of NLP research (Buissek, 2017), and has proven itself to be a good method to create such word representations (Goldberg & Levi, 2014).

Word2Vec has two possible architectures for its training method – ‘Continuous Bag-of-Words’ (CBOW) and ‘Skip-Gram’, which both train from the data the corpus in notably different methods. Using CBOW, the sentences are scanned over in a ‘sliding window’ method, with a specified number of words within the window and one ‘central’ word within the window. The words either side of the ‘central’ word are the ‘context’ for the ‘central’ word. The model will attempt to maximize the conditional probability of the output being the ‘central’ word, given the ‘context’. With Skip-Gram, the roles are reversed – the model attempts to maximize the conditional probability of the output being the ‘context’, given the ‘central’ word (Coyler, 2016). A visual representation of both methods is presented in Figure 2 below

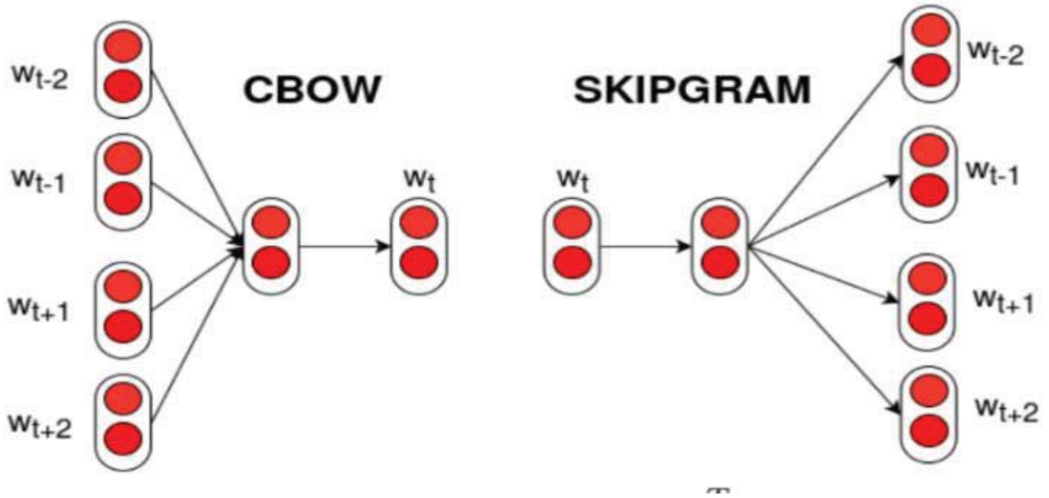


Figure 2 - CBOW and Skip-Gram layer representation with a window of size 5. Left layer is input, middle layer is the weights of the neural network, and right layer is the output.

Since it is usually required of Word2Vec to train on a very large corpus, as it will be in our case, optimizations are usually required in order to lower the training complexity. During our review of the literature we encountered three widely referenced methods:

- Hierarchial Softmax - Represent the vocabulary as a binary tree, where each leaf is a word from the vocabulary. A path to the word represents the probability of the word. Thus, instead of the model needing to evaluate V neural network output nodes, it needs to evaluate only around log(V) words in the tree. This method can be further optimized by making it a Huffman binary tree, so we can access more frequently appearing words quicker. *(Mikolov, 2013)*

- Negative Sampling – To lower the complexity involved in potentially updating every single one of the weights (of which there are typically up to 1000 of *(Mikolov, 2013)*) in the neural network for any given training sample, we can instead use the negative sampling method in order to have each training sample only update a small number of the weights. We do this by creating a sub-group of words to update the weights for – the target word an a few 'negative samples', where a 'negative sample' means a word that is not contextually related to the target word. The method of collecting the samples should be probabilistic, or can even be arbitrary *(Coyler, 2016)*.

- Subsampling – words with a very high frequency, such at 'the', 'a', 'in', can be considered to lack useful context in sentences and thus omitted from the training set. By doing this runtime during training can be saved. Mikolov's proposed method *(Mikolov, 2013)* to select the words to omit is a probabilistic one, where a word has a chance of being omitted as defined by the following formula:

$$P(w_i) = 1 - \sqrt{\frac{t}{f(w_i)}}$$

יש נוסחה ללא מיספור

Where $f(w_i)$ is the frequency of the word and t is a chosen threshold, typically around 10^{-5}. (Coyler, 2016)

Part of our challenge to implement Face2Vec is to decide on the features (such as the ones discussed above) most appropriate for our neural network to implement, especially taking into consideration that our training corpus is not precisely ‘words’ as intended in the literature.

THE METHOD OF IMPLEMENTATION

Firstly, we implement word2vec, using the PYTHON library. We have some ready results that were made by other students from the Industrial Engineering Dept., so we have something to compare our results with. We deduce our conclusions from this level of program training and transfer them into the world of the visual media.

How?
As a start, we are going to use a short-films database, made by students from the Cognition and Brain Science Dep. In those films, we are going to see people in their daily lives, using a variety of facial expressions.

While focusing on one film at a time, each one of us in the team, separately, will scour the video database for clips containing noticeable facial-changes between frames. After doing that separately, we combine our conclusions and decide where the changes really took place, by majority decision. In addition, we are going to define a threshold, which determines how many AU’s changed, in order for the human eye to notice. Different levels of change would be named with a different group, that contains all of those frames with minor change, or no change at all.

Next, we are going to treat each AU as a letter, each group of AUs, as a word, and a full sequence of frames, as a sentence. Now, we can take those pictures, and use the word2vec algorithm on the visual data we created and train our system on short films rather than words.

In order to analyze each frame, we will use both the "Meta-Face" program, and the "Face-Reader" program. Then we will use the R-studio program, for producing our results in a multi-dimensional graph to get some concrete conclusions.

STAGES OF THE PROJECT

תאור השלבים המתוכננים לביצוע הפרוייקט

לוח זמנים לביצוע כל שלב ושלב

יש להחליט האם הדוח
בעברית או באנגלית

STAGE	TASK	Time it takes	DEADLINE
1	Implementing word2vec	1 week	9.12.18
2	Learning how to work with the relevant programs	3 days	13.12.18
3	Filtering Short-films data base	2-3 weeks	31.12.18
4	Running the Face-Reader on the chosen films & extracting groups of changed AU's	2 weeks	15.1.19
5	Finding the best method of performing received results in a way we can get empirical conclusions & show them graphically using R-studio	1 week	22.1.19
6	Writing a code that would Embed frames into words – creating face2vec	3 weeks	12.2.19
7	Train the neural net on the algorithm, run it on the same short films we filtered, And compare the results	2 weeks	28.2.19
8	Fixes & improvements	2 weeks	15.3.19

זו טבלה לכל דבר לכן היא
צריכה לקבל כותר ומספר
וייצוג ברשימת טבלאות

REFERENCES

לא מקובל! אין מספור
למקורות ולכן קשה למצוא
התייחסות אלהם

Buissek, J. (2017). Demistifying Word2Vec.

Corardo, D., & Keltner, D. (2015). Understanding Multimodal Emotional Expressions: Recent Advances in Basic Emotion Theory. *EmotionResearcher*.

Coyler, A. (2016). The Amazing Power of Word Vectors.

Ekman, P., & Friesen, W. (1971). Constants Across Culture in the Face and Emotion. *Journal of Personality and Social Psychology*, 124.

Goldberg, Y., & Levi, O. (2014). Word2Vec Explained: Deriving Mikolov et al.'s Word Embedding Method.

Mikolov, T. (2013). Distributed Representations of Words and Phrases and their Compositionality.

Mikolov, T. (2013). Efficient Estimation of Word Representations in Vector Space. *Google*.

Tian, Kanade, & Cohn. (2001). Recognizing Action Units for Facial Expression Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Uyl, M. d., & Kuilenburg, H. v. (n.d.). The FaceReader: Online facial expression recognition.

אין אבני דרך כולל תאריכים ותוצרים
אין סכימת בלוקים של המערכת
אין תאור בדוח של מה יהיה התוצר הסופי של הפרוייקט
חסר טבלת קיצורים ומושגים
חבל שלא הצלחתם למצוא נוסחאה או חישוב מטמטי שאפשר להוסיף לפרוייקט
all pages should include headers and footer
אין התייחסות לסיכונים, ומה תעשו אם הם יתרחשו
אין רשימת מילות מפתח
נא לצרף שרטוט גאנט (שרטוט המתאר משימות ולוחות זמנים בצורה גראפית, חובה לציין נקודות קריטיות- משימה שתלויה במשימה אחרת)