Problem Statement
McMaster Text to Motion Database
CS 4ZP6

Brendan Duke
Andrew Kohnen
Udip Patel
Dave Pitkanen
Jordan Viveiros

April 8, 2017

1 Project Overview

The McMaster Text to Motion Database is a component of a larger project that is being managed by Dr.Graham Taylor from the University of Guelph. The larger project is a collaboration between the University of Guelph, SRI (a non-profit research organization) and other institutions, and was established with the goal of producing a "Computational Storytelling system".

This large-scale Computational Storytelling system is intended to work by taking text descriptions of a scene or dialogue and producing an animated video 'story' with that given content and characters.

Under Dr. Wenbo He from McMaster University, this smaller project will contribute a Python HTTP server that can be used to process images and videos with a deep learning algorithm, along with a website and database that can be used to store and view the processed media. Having access to the processed media will be a resource that can be of use to the larger project.

2 Problem Statement and Our Contribution

To implement the 'Computational Storytelling' functionality, the larger system should be able to produce an animation with moving characters that look reasonably realistic.

As of now, there is no simple, accessible database that links images and videos with text annotations or descriptions of the motions/positions of people observed in the image or video (ex. mapping the locations of some of the observed skeletal joints of a person in the image or video)

We propose to harvest the motion data (joint positions) of moving people by using a deep learning model and seeding our database with existing large image/video datasets of people doing common movements, and categorizing the processed media with tags. The developers of the larger system can use our website to search for specific actions, and take the motion data in order to generate the animated story with moving characters. The developers can also upload their own media to be processed as well at any time.

3 Motivation

The ability to move between natural language and motion data is at the core of what the larger project is trying to acheive. To enable research along the lines of converting text descriptions into observable motion, we propose to create a large databank of images and videos that are annotated both with human actions (via tags) and the motions observed (the positions of joints relative to each other).

4 Background Information and Challenges

Over the last ten years machine learning algorithms have significantly improved as well as the limits of our computer processing power, computer memory and available sources of many types of data. The growth in all these areas has made it possible for computer applications that use machine learning to improve their performance. The key advancement in the algorithms of machine learning has been the ability to train neural networks that are more than a single layer deep. With this ability neural networks are able to learn structures from long sequences of sequential data(long sequences of inputs such as natural language and video in our case).

Many different researchers have already been able to create a mapping between text and images (Kiros et al., 2014a, b; Socher et al., 2010). However these examples simply map from image to text but the problem of performing the mapping using videos is that there is now a new dimension of time and the volume of data grows very quickly as the length of the video increases.

The biggest challenge of this project is to produce a deep learning model that is trained to analyze an image or video (a set of image frames) and estimate the joint positions of a human, if there is one in the media. Luckily, there are some existing open-source code bases that work with varying platforms and frameworks (like C/C++ or Python). Implementing these models will take up the bulk of the project research.

Once the website can take in large amounts of media input from users and process the media, the databank can be populated by existing collections of images/videos that are pre-labeled with the actions observed in the media

The datasets with heavy labelling that we intend to use to seed the database of our website are: MovieQA, Charades, MSR-VTT

5 Objectives, Constraints and Deliverables

The goal of the entire collaboration is to create a basic text-to-motion tool which can link language concepts to a data structure representing motion (e.g. a series of joint angles). The first step in analyzing motion from a machine learning approach is typically to create a feature space – which is a lower dimensional representation of some of the important aspects of the data which the machine learning algorithms can then analyze.

The specific feature space we wish to consider is called a pose structure (Wei, 2016). This structure is used to model the motion of people. The feature space essentially consists of a labelling of the regions in an image where a specific limb is located (arm, head eye etc). Modelling the motion of the entire body is made much simpler when the problem can be decomposed into the independent motion of these features (applying the constraints that the features are connected in the conventional human form).

Therefore in order to achieve this first step we hierarchically break this set of deliverables into 2 components

1. Web-Interface and Database
    - The website allows users to upload media to be processed by the deep learning algorithm. The users are prompted to associate the processed media with some tags and save it to the database.
    - The website allows users to search for media with certain actions by having a full-text search on the tags that people that associate their uploads with in the database.
2. Tensorflow Deep Learning Model and HTTP Server
    - Will involve exploring pose estimation and tracking methods in Tensorflow that can extract motion data out of images and videos
    - Will use the Python Request library to handle incoming requests containing images/videos to process
    - References we will use as starting points for our model:
        i (Pfister, Charles, and Zisserman 2015)
        ii (Wei, Ramakrishna, and Kanade 2016)
        iii (Belagiannis and Zisserman 2016)
        iv (Tekin et al. 2016)
        v (Zhou et al. 2015)

These sub-goals reflect the requirement that we wish to create an interactive database that contains thousands of examples of our data structure (video motion that is labelled with the poses of the people in the video). These examples will then be used in the larger collaboration for a supervised learning algorithm that will perform the natural language to video mapping

6 Bibliography

Belagiannis, V., and A. Zisserman. 2016. Recurrent Human Pose Estimation. *arXiv Preprint arXiv:1605.02914*. arxiv.org. http://arxiv.org/abs/1605.02914.

Jain, Ashesh, Amir R. Zamir, Silvio Savarese, and Ashutosh Saxena. 2016. Structural-RNN:Deep Learning on Spatio-Temporal Graphs. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Kiros, R., Salakhutdinov, R., and Zemel, R. (2014a). Multimodal neural language models.In *ICML2014*. 102

Kiros, R., Salakhutdinov, R., and Zemel, R. (2014b). Unifying visual-semantic embeddings with multimodal neural language models. *a rXiv:1411.2539* [cs.LG]. 102 , 410

Pfister, Tomas, James Charles, and Andrew Zisserman. 2015. Flowing ConvNets for Human Pose Estimation in Videos. *arXiv [cs.CV]*. arXiv.

http://arxiv.org/abs/1506.02897 .

Socher Richard, Karpathy Andrej, Le Ve Quoc, Manning D. Christopher, Ng Y.
Connecting modalities: Semi-supervised segnmentation and annotation of images using un-aligned text corpora. CVPR (2010).

Taylor, Graham W., Geoffrey E. Hinton, and Sam T. Roweis. 2011.
Two Distributed-StateModels For Generating High-Dimensional Time Series. *Journal of Machine Learning Research: JMLR* 12 (Mar): 102568.

Tekin, Bugra, Isinsu Katircioglu, Mathieu Salzmann, Vincent Lepetit, and Pascal Fua. 2016.
Structured Prediction of 3D Human Pose with Deep Neural Networks. *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/1605.05180 .

Wei, S. E., V. Ramakrishna, and T. Kanade. 2016.
Convolutional Pose Machines. *arXiv Preprint arXiv*: arxiv.org. http://arxiv.org/abs/1602.00134 .

Zhou, Xiaowei, Menglong Zhu, Spyridon Leonardos, Kosta Derpanis, and Kostas Daniilidis. 2015.
Sparseness Meets Deepness: 3D Human Pose Estimation from Monocular Video. *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/1511.09439 .