sEMG Gesture Recognition With a Simple Model of Attention

David Josephs[1], Carson Drake[1], Che' Cobb[1], and John Santerre[1]

Master of Science in Data Science, Southern Methodist University, Dallas TX 75275
USA {`josephsd, drakec, cobbc`}`@smu.edu`

Abstract. This paper presents a novel method for fast classification of sEMG signals, using a simple model of attention. On a difficult, industry benchmark sEMG dataset, the proposed method yields excellent results, classifying 36 more gestures with about 20% higher accuracy than the current standards in the field. These results have direct and immediate application in the fields of robotics, myoelctric control, and prosthetics.

1 Introduction

Electrophysiological studies of the nervous system are the core area of research in clinical neurophysiology, where scientists attempt to link electrical signals from the body to real world effects. These studies include measuring brain waves (electroencephalography), comparison of sensory stimuli to electrical signals in the central nervous system (evoked potential), and the measure of electrical signals in skeletal muscles (electromyography).

Electromyography is of particular interest to this paper. The nervous system uses electrical signals to communicate with the rest of the body. When a signal from the nervous system reaches a skeletal muscle, the myocytes (muscle cells) contract, causing a physical motion. By measuring these electrical signals in a supervised manner, we can develop a link between signal and physical action. This connection yields many powerful uses, ranging from quantifying physical veracity to diagnosing neurodegenerative diseases. An example of the latter can be found in Akhmadeev et al. [2], where electromyographic (EMG) signals were used to classify Multiple Sclerosis patients from healthy control subjects with 82% accuracy.

Deep learning can be used to further improve the power and utility of the EMG analysis. A deep neural network is, in essence, a composition of neurons (regressors) that learns a functional mapping between two sets of data. By learning a mapping between EMG signal and physical effect, we can develop more sensitive and accurate models of what connects the two. This also allows us to use less intrusive measurement devices in studies and in the real world. The applications of this range from clinical trials and prognostication of neuromuscular diseases to gesture prediction in "brain-controlled" prosthetic limbs. This study focuses on the latter application. Current state of the art myoelectric prosthetic limbs are capable of detecting and performing between 7 and 18 motions [8] [1].

While this represents a significant increase in the quality of life of an amputee, there remains a large amount of room for improvement. Most of the potential for improvement does not lie in the robotics themselves, which are fairly robust, but within the device which maps brain signal to motion of the hand. The aim of this paper is therefore to build a highly accurate gesture classifier using EMG signal, capable of classifying a broad range of gestures.

Extending from this initial goal, there are certain metrics which must be met for the classifier to be deemed "useful". First, it must be fast. If there is a large degree of latency from thought to hand motion, the user will simply not use the device. Thus, it is important to determine the time window in which a prediction must be made. The absolute largest prediction latency for a model to be considered useful for myoelectric control lies between 250 and 300 milliseconds [10], [22]. For this study, the precedent set by [6] was followed, with a 260 millisecond prediction window. Another key consideration for the classifier is generalizability. Within the field of myoelectric control and gesture recognition, there are two important types of generalizability, intra-subject and inter-subject [20]. Intra subject generalizability indicates that the model is robust to personal elements, such as muscle fatigue. In contrast, inter-subject generalizability refers to the ability of the model to generalize to new people. This study focuses on intra-subject generalizability, as it is important that the device keep working for extended periods of time, while models can be fit to and personalized for amputees and non amputees alike [4].

In this paper, we propose a novel attentional architecture for sEMG recognition and myoelectric control. We demonstrate the model's validity on the 53-class NinaPro DB5 [17]. We also compare different techniques for dealing with the inherent class imbalance in sEMG, a synthetic data based approach (augmentation), an undersampling based approach, and a loss-based approach. All methods are evaluated on an intra-subject basis using the same holdout samples, yielding promising results.

2 Surface Electromyography

Before discussing the mechanics of the proposed model, it is important to first discuss electromyography more closely. As mentioned in the Introduction, electromyography is the measure of electrical signals from the brain in the rest of the body. To be more specific, the body controls the forces in the muscles using electrical signals, via units called *motor control units*, as seen in Figure 1. By measuring the amplitude and frequency of these signals, the strength or veracity of the motion and the type of motion can be determined. These signals are recorded in units of *Motor Unit Action Potential* (MUAP), either through internal electrodes or electrodes on the skin. This study is particularly focused on the measure of signals across the skin, or surface electromyography (sEMG). While sEMG is not quite as powerful as true EMG (e.g., the signals are weaker), it is far more practical and scalable, and far less disruptive, invasive, or sensitive to change. In this study, data are collected using two MYO armbands [25], on the

Fig. 1. A motor unit, consisting of: *1* an Axon, *2* a junction site, *3* a Myocyte, and *4* Myofibrils. The Axon sends electrical signal through the junction cite, causing the all of the myofibrils (and thus the entire myocyte) to contract.
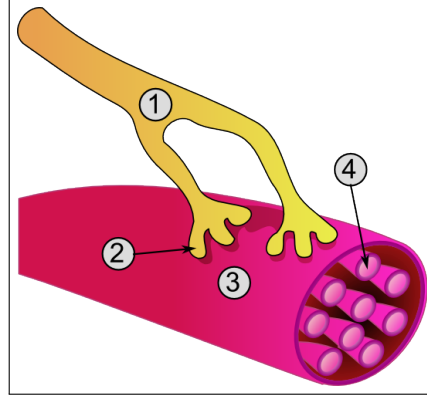
upper arm of the subject (just above the elbow), as seen in Figure 2. Intuitively, the idea is to collect the signal where the sensors are, and classify them with enough speed that it feels as if the hand were being moved on its own.

2.1 The NinaPro Database

The data in this study comes from the NinaPro Database, specifically database number 5 [17]. Within this database, there are 52 unique motions measured, as well as rest, collected over 10 subjects. Each subject does each exercise 6 times, with 3 seconds of rest between each repetition. The signal from these are collected at the frequency of the MYO armband, 200 Hertz.

This is one of the big challenges and benefits of the MYO armband. Low frequency sEMG is very sparse in information, and thus difficult to classify (especially relative to expensive HDsEMG sensors), however this also means that the MYO armband is cheap and accessible to the average consumer. The other benefit of MYOs cheapness and sparsity can be inferred from Figure 2. The cheap sensor constricts the arm and stays in place. Because the signals are weaker, slight shifts in the electrode do not have as much of an impact on the underlying signal (as the sensor is not incredibly sensitive). This means that the MYO armband is simpler and more robust for use by the general public.

The goal of this study is to develop an accurate mapping between MUAP in the upper arm and the motion of the hands.

Fig. 2. The MYO armbands are placed just above the elbow.

3   sEMG Classification

In order to justify the need for deep sEMG classifiers, it is important to briefly discuss why shallow classifiers are inappropriate. Shallow classifiers (and really all of classical machine learning) relies on *feature engineering*, or the creation of robust features which represent the data in a precise manner [16]. This is very powerful, but it also comes with the assumption that the same fundamental process is generating all data points, that is the same feature set that works at time a for person x works at time b for person y. However, in the case of surface electromyography, all brains and bodies are completely unique, which means that a classical, engineered feature set is neither robust between subjects or between repetitions (over time) [19]. Therefore, we must turn to deep learning, to *learn features from the data*, which are more robust over time and over subjects.

3.1   Deep Learning

Much of the previous work in sEMG classification (and time series classification in general [7]) with deep learning revolves around convolutional neural networks. In this subsection, notable and influential recent work will be highlighted. [3] very clearly demonstrated the power of convolutional neural networks with respect to sEMG classification, namely their flexibility and ease of use (due to feature learning vs feature engineering). This work is expanded in [6], in which the authors use convolutional neural networks, wavelet transformations, and deep transfer learning to slightly improve on the results highlighted in [3].

This study instead explores a new direction for both time series classification in general and for sEMG classification in particular: attention. In the last

few years of language modeling, attentional models have proven more and more utilitarian, managing to solve problems of long term memory (which are commonplace in time series and language) while also managing to avoid the training and gradient issues of recurrent neural networks. Attention is a flexible, intuitive mechanism with lots of untapped application. In this study, the simple, feed-forward model of attention described in [18] is expanded upon. The attention mechanism is shown in Figure 3.

Fig. 3. A simple model of temporal attention.
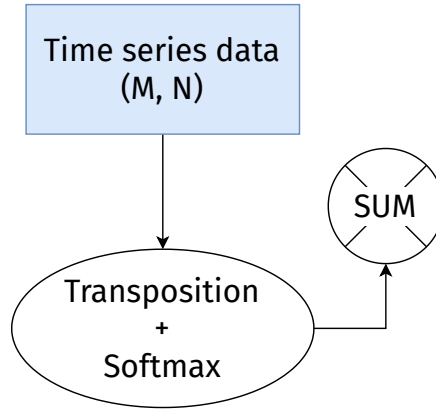
The mechanism works as follows: First time series data, an (M, N) matrix (where M represents timesteps and N represents features or channels). This matrix is transposed into an N by M matrix. In this matrix, an observation at a single row represents all timesteps of that feature or channel. For example, row x represents the time series produced by the feature $N = x$. This matrix is then fed in row by row to a standard, feedforward layer, using the softmax activation function. Thus, for each timestep of each feature, an importance (referred to as an "attention score") is calculated. Next, these are summed across time, producing a single number per feature, which represents the overall importance of the feature within that sample. As per [18], this simple mechanism, akin to a parametric time weighted average, actually successfully solves long term memory problems where order is not very important (or in this case, where memorization is to be avoided).

4 Data Processing

Before any gestures can be predicted, the data must be prepared. There are two considerations for this. First, the preprocessing method must be fast and simple, as it needs to run with near zero latency on edge hardware (for example in a prosthetic hand). Second, the goal of using a deep neural network is to *learn* robust features. Therefore, preprocessing methods must not profoundly alter the data, they must be simple methods which contain the maximum amount of available information. To start, the data (collected at 200 Hz), is sampled in time windows of 260 milliseconds (or 52 time steps), following the precedent set in [6]. One window of raw data can be seen in Figure 4.

Fig. 4. One window of sEMG data. The amplitude represents unitless motor activity potential, while the x axis represents the passage of time
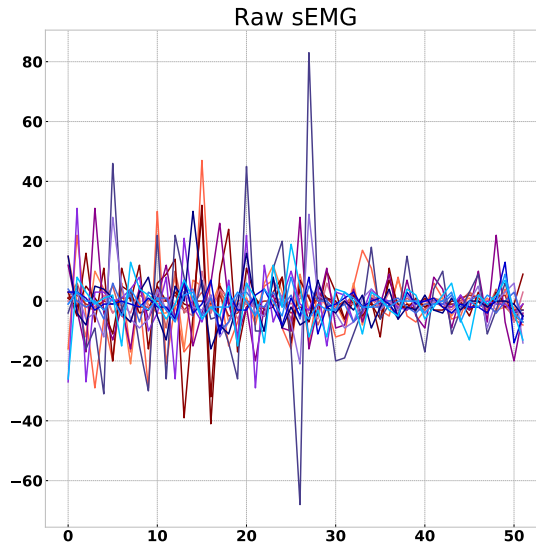

This raw data represents what is collected by the two MYO armbands. Each armband collects 8-channel signal across the arm at 200 Hz. To remove any powerline interference, the armband collects the data with a notch filter which automatically removes the effects of large nearby electronics (such as power lines). The next step taken to make the signal more rich in information was to rectify (in layman's terms, take the absolute value of) the signal. In [15], it was shown that rectification significantly increased the availability of information pertaining to the firing rate (temporal activity and pattern) of the motor units

producing the signal. This means that rectification of the sEMG makes the information which allows for the learning of a pattern or feature which maps to a specific gesture more readily available.

After making the information denser, the next important preprocessing step is to make the information clearer. This is done in two ways. First, the packet of rectified signal is passed through a high pass Butterworth filter at 20 Hz. This removes any low frequency artifacts (aliases), which obscure the relevant signal. Finally, the data is passed through a moving average filter, which makes it more evident when the muscles are going through different periods of activity [23]. The results of this preprocessing can be seen in Figure 5.

Fig. 5. One window of sEMG data. The amplitude represents unitless motor activity potential, while the x axis represents the passage of time. This data has been first rectified, then passed through a high pass filter, then passed through a 15-timestep moving average filter.
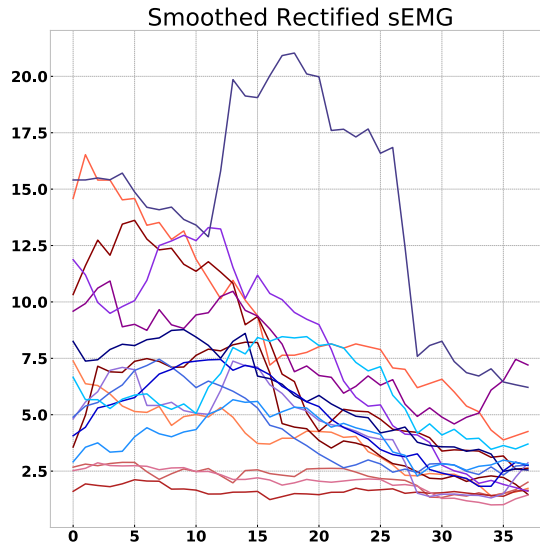
4.1 Data Augmentation

One of the great challenges of deep learning, even with barely processed data, is its incredible capacity to overfit. In order to learn more robust mappings, without collecting a significantly larger amount of data, synthetic data must be generated, in order to prevent the deep learning algorithm from simply memorizing the training data. In this study, a novel data augmentation technique was employed, as seen in Figure 6. The methodology for the data augmentation is a

simple random sample of a spectrum of different signal to noise ratios (SNR). A processed series is fed into the augmenter, and then random noise, at a given signal to noise ratio in magnitude is added. The signal to noise ratio for a realization is calculated at random as well, with linearly increasing likelihood as SNR increases. This allows the model to be trained for an extended period, without ever seeing the same observation twice.

Fig. 6. Data augmentation scheme, highlighted on a single channel of a single sample. Brighter realizations are more likely than dim ones.
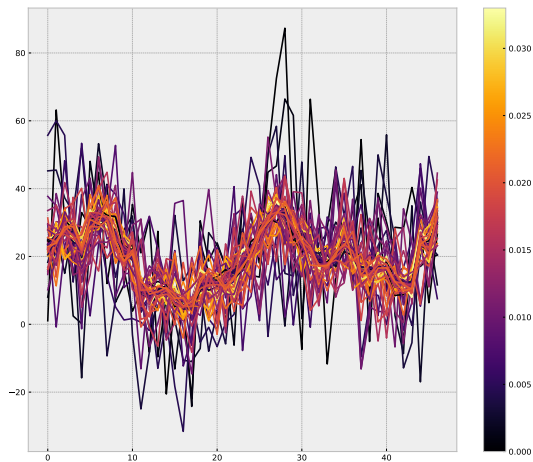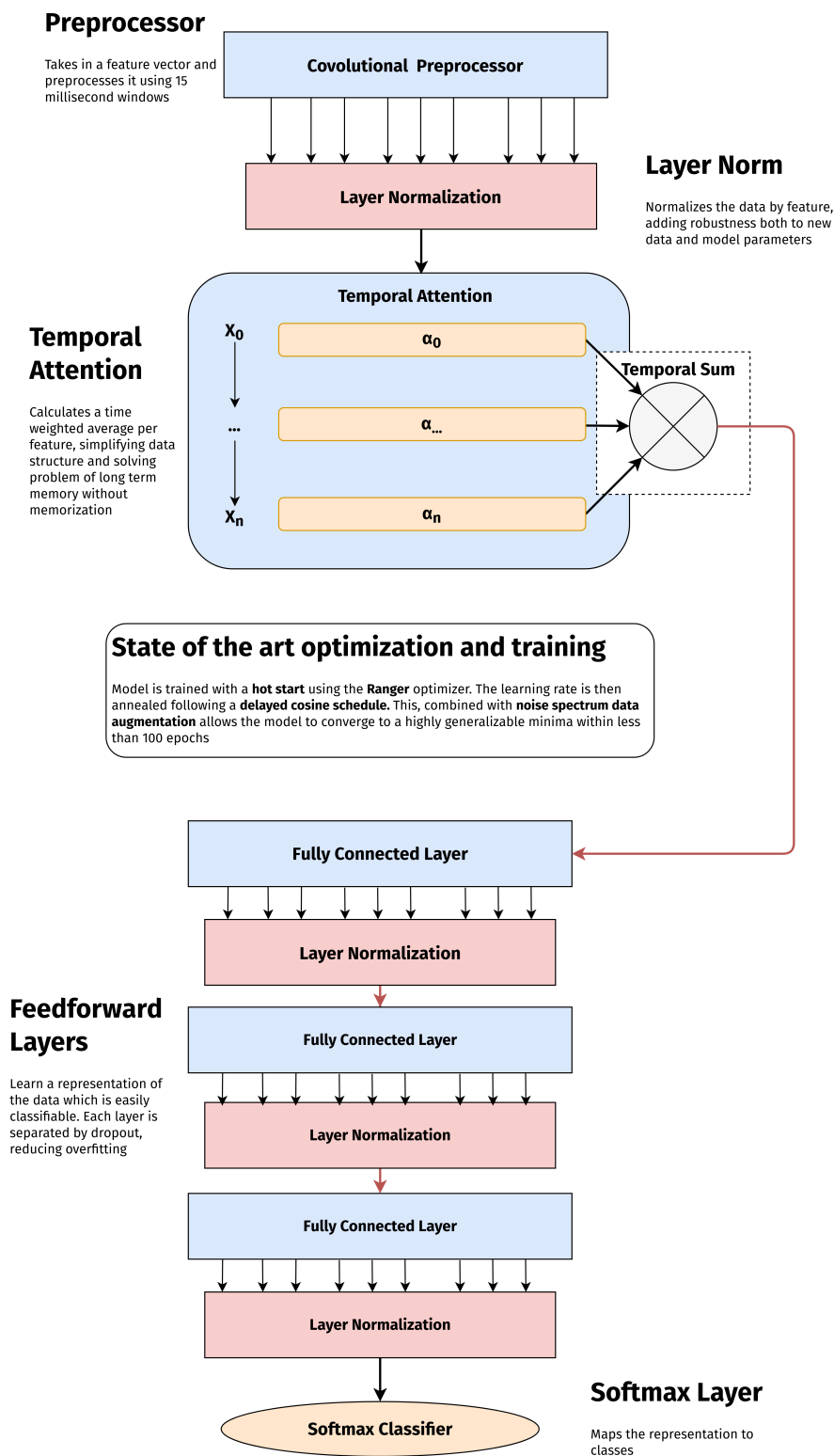

5 Modeling

The proposed model consists of three distinct phases: a convolutional preprocessor, an attention layer (as discussed in Section 3), and a classifier.

Fig. 7. The proposed architecture. Red arrows represent dropout connections.

Preprocessor

Takes in a feature vector and preprocesses it using 15 millisecond windows

Covolutional Preprocessor

Layer Normalization

Layer Norm

Normalizes the data by feature, adding robustness both to new data and model parameters

Temporal Attention

Calculates a time weighted average per feature, simplifying data structure and solving problem of long term memory without memorization

Temporal Attention

X_0 α_0

... $\alpha_{...}$

X_n α_n

Temporal Sum

State of the art optimization and training

Model is trained with a **hot start** using the **Ranger** optimizer. The learning rate is then annealed following a **delayed cosine schedule.** This, combined with **noise spectrum data augmentation** allows the model to converge to a highly generalizable minima within less than 100 epochs

Fully Connected Layer

Layer Normalization

Feedforward Layers

Learn a representation of the data which is easily classifiable. Each layer is separated by dropout, reducing overfitting

Fully Connected Layer

Layer Normalization

Fully Connected Layer

Layer Normalization

Softmax Classifier

Softmax Layer

Maps the representation to classes

As seen in Figure 7, the first stage of the model consists of a single, 1-dimensional convolutional preprocessing layer. This layer takes in the processed time series data, and filters them with a kernel size of three. That is, every three timesteps is filtered down into a single value (for each channel). This is done for each timestep, with 128 filters, resulting in a $(M, 128)$ matrix (where M is the length of the time series). The output of this layer, and the output of all subsequent layers, are normalized across features (layer normalization) [5]. Layer normalization differs from batch normalization in that it normalizes across features, instead of across batches. Mathematically, consider a batch with dimensions **(samples, timesteps, features)** (or, after attention, **(samples, features)**). When batch normalization is applied, the weights are normalized using summary statistics across the *samples*, while with layer normalization, the summary statistics used to normalize the weights are calculated across the *features*.

This matrix is then fed into the feed-forward attention mechanism described in Figure 3, resulting in a feature vector of 128 values. These are fed into the classification network, which consists of a set of standard, feed-forward layers. The arhitecture of the classifier is based off of the architecture proposed in [9], and was chosen due to its ability to effectively represent complex data. The output of this network is softmaxed in order to produce the final classifications. Each layer of the classification network is regularized using dropout.

5.1   Training

Several state of the art training techniques were employed in this model. First, instead of using the ReLU activation function, all layers use the "Mish" activation function [14]. The Mish activation function is slightly less regularized then ReLU, thus yielding results slightly faster (and avoiding the "dead neuron" problem). Instead of optimizing the network with Adam or SGD, the model was optimized using the Ranger optimizer. The Ranger optimizer consists of two components: Rectified Adam (RAdam) and Lookahead. The RAdam algorithm represents an improvement over the Adam optimizer in that it does not require a "warm up period", which Adam is notorious for needing. This allows the model to be trained to an optima much faster [13]. The Lookahead algorithm works in conjunction with a primary optimizer. The primary optimizer calculates weights as it normally does, and then the lookahead optimizer explores the loss landscape near the calculated weights. This allows for even faster convergence to an optima [26]. The combination of Lookahead and RAdam is the Ranger optimizer used in this paper [24]. Due to the very hot start of this combination (Mish and Ranger), the model learns very quickly. To further speed up training, the learning rate followed a delayed cosine annealing schedule. For the first 5 epochs, the model trains at a very high learning rate, and is subsequently annealed over the course of 50 epochs to a very low learning rate. This allows the model to quickly propel itself to a flat optima, and then work towards the bottom. This "hot start" methodology allows the model to quickly converge to a robust local minima [11].

A significant issue in sEMG classification is the large class imbalance present in the data. Most of the time, the hand is at rest. In the Nina 5 dataset, this means that there was over 30 times rest as any of the other 53 classes (and this also means that a non movable prosthetic hand would be about 60% accurate). Three methods were tested to combat this imbalance. First, undersampling the majority class (rest), so that it would have the same likelihood of occurring in the training dataset as the other classes yielded decent results. A similar approach, augmenting all the minority classes to match the majority class, was tested, however the slight benefit over undersampling was far outweighed by the prohibitive training cost of this method. Finally, a loss-based method was tested, using focal loss [12]. Focal loss is calculated in almost exactly the same manner as categorical cross entropy, except it gives a lower importance to well aligned (easy to classify) samples, and raises the importance of difficult samples. This yielded by far the best results.

6 Results

All models were tested on all 53 classes of the NinaPro 5 database, with the fourth and fifth repetitions as a validation set and the final repetition as a test set.

Table 1. Results compared to other recent papers.

Model	Window Size	Accuracy	Number of Gestures
Attentional Network, Focal Loss	260 ms	87%	53
Attentional Network, Undersampling	260 ms	83%	53
Allard et al, 2019	289 ms	68%	17
Simao, 2019	1.5 s	90%	7
Atzori, 2016	150 ms	66%	17

Table 1 makes the outcome of this study in relation to other papers in the field very clear. The addition of attention and focal loss resulted in extremely high accuracies, over a wide range of classes (from finger motions to functional grabbing motions and fine wrist motions). The only recent paper which performed better on the same dataset is mentioned in [21], however these results are over only 7 classes using 1.5 second time windows, and therefore are only relevant as a reference to the difficulty of 53-class classification. To further diagnose the successes and failures of this study, a confusion matrix is provided in Figure 8

Fig. 8. Confusion matrix of the classifications by the best performing model.
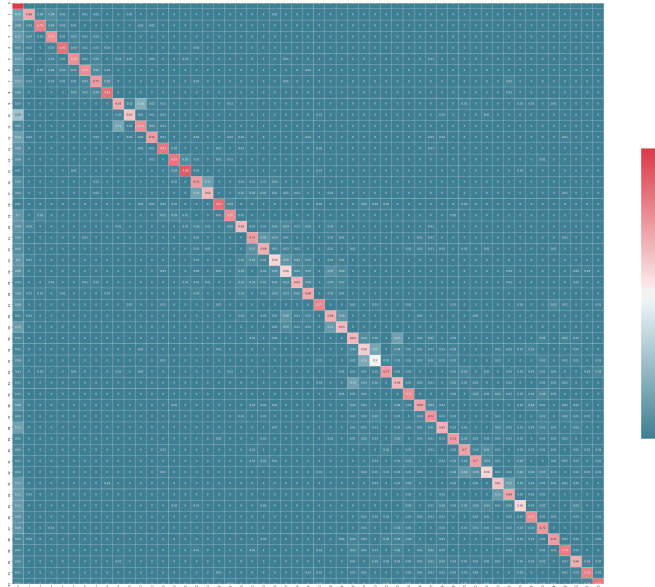
We can see clearly that the model performed excellently with rest, as well as with most of the fine finger and wrist motions (the upper left quarter of the confusion matrix). However, there were a few functional movements (bottom right quarter) where the model struggled slightly. Overall, the presence of red along the diagonal indicates that the model was able to classify the correct gesture the majority of the time, and with the broadness of the range of gesture, the model can be deployed in prosthetic limbs in order to significantly increase the quality of life of amputees.

7 Conclusions

In this study, a novel method for classifying sEMG signals for myoelectric control was highlighted. It was shown that using a simple attention mechanism and state of the art training techniques, a model can be built to accurately classify a broad range of gestures, in a short period of time. This has immediate applications in

the field of prosthetics, and can be expanded to other body parts with relatively little difficulty.

One general flaw of this study in general is the use of supervised learning at all. The presence of labels in human motion is rather unintuitive. In the future, research should be expanded into reinforcement learning and imitation learning, potentially combined with computer vision, in order to develop more robust models with an infinite range of gestures.

References

1. Myoelectric prosthetics, `https://www.ottobockus.com/prosthetics/upper-limb-prosthetics/solution-overview/myoelectric-prosthetics/`
2. Akhmadeev, K., Houssein, A., Moussaoui, S., Høgestøl, E.A., Tutturen, I., Harbo, H.F., Bos-Haugen, S.D., Graves, J., Laplaud, D.A., Gourraud, P.A.: Svm-based tool to detect patients with multiple sclerosis using a commercial emg sensor. 2018 IEEE 10th Sensor Array and Multichannel Signal Processing Workshop (SAM) pp. 376–379 (2018)
3. Atzori, M., Cognolato, M., Müller, H.: Deep learning with convolutional neural networks applied to electromyography data: A resource for the classification of movements for prosthetic hands. Frontiers in Neurorobotics **10** (Jul 2016). https://doi.org/10.3389/fnbot.2016.00009
4. Atzori, M., Gijsberts, A., Castellini, C., Caputo, B., Hager, A.G.M., Elsig, S., Giatsidis, G., Bassetto, F., Müller, H.: Electromyography data for non-invasive naturally-controlled robotic hand prostheses. Scientific data **1**(1), 1–13 (2014)
5. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization (2016)
6. Côté-Allard, U., Fall, C.L., Drouin, A., Campeau-Lecours, A., Gosselin, C., Glette, K., Laviolette, F., Gosselin, B.: Deep learning for electromyographic hand gesture signal classification using transfer learning. IEEE Transactions on Neural Systems and Rehabilitation Engineering **27**(4), 760–771 (April 2019). https://doi.org/10.1109/TNSRE.2019.2896269
7. Fawaz, H.I., Forestier, G., Weber, J., Idoumghar, L., Muller, P.A.: Deep learning for time series classification: a review. Data Mining and Knowledge Discovery **33**(4), 917–963 (Feb 2019). https://doi.org/10.1007/s10618-019-00619-1
8. Geethanjali, P.: Myoelectric control of prosthetic hands: state-of-the-art review. Medical Devices: Evidence and Research **Volume 9**, 247–255 (2016). https://doi.org/10.2147/mder.s91102
9. Hinton, G.E., Osindero, S., Teh, Y.W.: A fast learning algorithm for deep belief nets. Neural Computation **18**(7), 1527–1554 (2006). https://doi.org/10.1162/neco.2006.18.7.1527
10. Hudgins, B., Parker, P., Scott, R.N.: A new strategy for multifunction myoelectric control. IEEE Transactions on Biomedical Engineering **40**(1), 82–94 (Jan 1993). https://doi.org/10.1109/10.204774
11. Jastrzebski, S., Szymczak, M., Fort, S., Arpit, D., Tabor, J., Cho, K., Geras, K.: The break-even point on optimization trajectories of deep neural networks (2020)
12. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection (2017)
13. Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J., Han, J.: On the variance of the adaptive learning rate and beyond (2019)

14. Misra, D.: Mish: A self regularized non-monotonic neural activation function (2019)

15. Myers, L., Lowery, M., Omalley, M., Vaughan, C., Heneghan, C., Gibson, A.S.C., Harley, Y., Sreenivasan, R.: Rectification and non-linear pre-processing of emg signals for cortico-muscular analysis. Journal of Neuroscience Methods **124**(2), 157–165 (2003). https://doi.org/10.1016/s0165-0270(03)00004-9

16. Phinyomark, A., Scheme, E.: Emg pattern recognition in the era of big data and deep learning. Big Data and Cognitive Computing **2**(3) (2018). https://doi.org/10.3390/bdcc2030021, `https://www.mdpi.com/2504-2289/2/3/21`

17. Pizzolato, S., Tagliapietra, L., Cognolato, M., Reggiani, M., Müller, H., Atzori, M.: Comparison of six electromyography acquisition setups on hand movement classification tasks. PLOS ONE **12**, e0186132 (10 2017). https://doi.org/10.1371/journal.pone.0186132

18. Raffel, C., Ellis, D.P.W.: Feed-forward networks with attention can solve some long-term memory problems (2015)

19. Rehman, M.Z.U., Waris, A., Gilani, S., Jochumsen, M., Niazi, I., Jamil, M., Farina, D., Kamavuako, E.: Multiday emg-based classification of hand motions with deep learning techniques. Sensors **18**(8), 2497 (Jan 2018). https://doi.org/10.3390/s18082497

20. RING, F.: LEARNING APPROACHES IN SIGNAL PROCESSING. PAN STANFORD Publishing (2018)

21. Simão, M., Neto, P., Gibaru, O.: Emg-based online classification of gestures with recurrent neural networks. Pattern Recognition Letters **128**, 45–51 (2019). https://doi.org/10.1016/j.patrec.2019.07.021

22. Smith, L.H., Hargrove, L.J., Lock, B.A., Kuiken, T.A.: Determining the optimal window length for pattern recognition-based myoelectric control: Balancing the competing effects of classification error and controller delay. IEEE Transactions on Neural Systems and Rehabilitation Engineering **19**(2), 186–192 (Apr 2011). https://doi.org/10.1109/tnsre.2010.2100828, `http://dx.doi.org/10.1109/TNSRE.2010.2100828`

23. Tamura, H., Yan, M., Sakurai, K., Tanno, K.: Eog-semg human interface for communication. Computational Intelligence and Neuroscience **2016**, 1–10 (2016). https://doi.org/10.1155/2016/7354082

24. Tong, Q., Liang, G., Bi, J.: Calibrating the adaptive learning rate to improve convergence of adam (2019)

25. Visconti, P., Gaetani, F., Zappatore, G., Primiceri, P.: Technical features and functionalities of myo armband: An overview on related literature and advanced applications of myoelectric armbands mainly focused on arm prostheses. International Journal on Smart Sensing and Intelligent Systems **11**, 1–25 (06 2018). https://doi.org/10.21307/ijssis-2018-005

26. Zhang, M.R., Lucas, J., Hinton, G., Ba, J.: Lookahead optimizer: k steps forward, 1 step back (2019)