Multi-frame Weak Supervision to Label Wearable Sensor Data

Saelig Khattar [1] Hannah O'Day [2] Paroma Varma [1] Jason Fries [1] Jennifer Hicks [2] Scott Delp [2]
Helen Bronte-Stewart [2] Chris Re [1]

Abstract

Using modern deep learning models to make predictions on time series data from wearable sensors generally requires large amounts of labeled data. However, labeling these large datasets can be cumbersome since each sequence is comprised of many individual elements. In this paper, we present a weak supervision framework for programmatically labeling time series training data. We modify an existing weak supervision framework by (1) accepting supervision sources that operate over different temporal granularities and (2) using a multi-task model to capture the relation among elements that belong to the same sequence. We apply our algorithm to label clinically relevant freezing behavior (i.e., transient, interrupted walking) in time series data from sensors worn by patients with Parkinson's disease. Training an LSTM model with our weakly supervised method outperforms traditional weak supervision by 4.3 F1 points and comes within 4.8 F1 points of models using $4\times$ more hand-labeled data.

1. Introduction

Time series data generated by wearable sensors are an increasingly common source of biomedical data. With their ability to monitor events in non-laboratory conditions, sensors offer new insights into human health across a diverse range of applications, including continuous glucose monitoring (Cappon et al., 2017), atrial fibrillation detection (Tison et al., 2018), fall detection (Casilari et al., 2017), and general human movement monitoring (Kumari et al., 2017). Machine learning could help automate many of these monitoring tasks and enable medical professionals to make more informed decisions. However, building large labeled training sets is time consuming and expensive, especially for human movement data that has considerable inter-subject variability (Halilaj et al., 2018). Thus, there is a need to efficiently label the large amounts of data that supervised machine learning algorithms require for time series tasks.

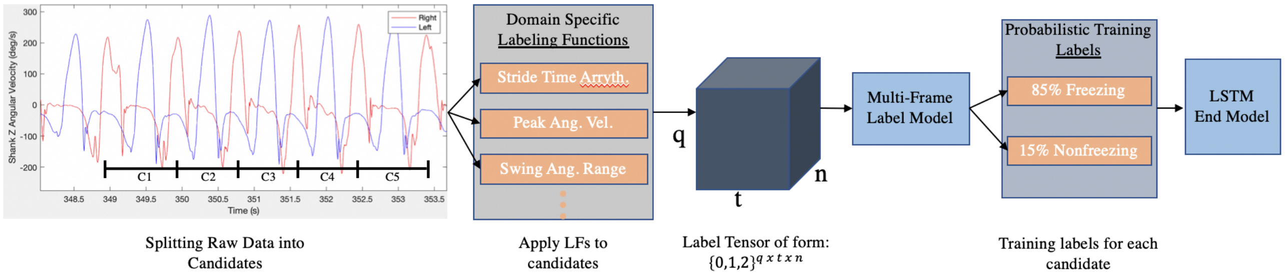
Weak supervision has proven effective at mitigating this problem in a variety of imaging and text classification applications (Xiao et al., 2015; Ratner et al., 2017; Fries et al., 2018). Instead of using manually labeled training data, weak supervision encodes domain insights into the form of noisy, heuristic *labeling functions*, which are combined by a graphical model to create probabilistically labeled training sets. However, current weak supervision paradigms assume classification targets are i.i.d. and thus do not model correlations between consecutive samples in sensor data. As a first step towards applying weak supervision to time series data, we adapt an existing multi-task weak supervision method (Ratner et al., 2018) to operate over temporally correlated data. We call this approach *multi-frame weak supervision*.

The two key contributions of this work are as follows. First, sensor data is decomposed into a series of *frames*, where each frame in a sequence is modeled as a separate task in a multi-task weak supervision model. This formulation captures the chain dependency structure between frames and enables defining multi-granular labeling functions, which can operate over individual frames or larger frame sequences. Second, we learn multiple accuracy parameters for each labeling function, conditioned on the frame index in a sequence. This is useful when frames at different indices follow different distributions. These additional modeling capabilities enable our algorithm to label correlated samples more accurately than traditional weak supervision approaches.

Once the weak supervision model learns labeling function parameters, we generate probabilistic training labels that can be used to train any downstream classification model such as a deep neural network. As a motivating use case, we use ankle sensor data to classify freezing behaviors in people with Parkinson's disease, relying on labeling functions to encode biomechanical knowledge about human movement and Parkinson's (Halilaj et al., 2018). Using our weak supervision framework to model temporal correlations, we

[1] Department of Computer Science, Stanford University, Stanford, California, USA [2] Department of Bioengineering, Stanford University, Stanford, California, USA. Correspondence to: Saelig Khattar <saelig@cs.stanford.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).



Shank Z Angular Velocity (deg/s)
Time (s)
Right
Left
C1
C2
C3
C4
C5
Domain Specific Labeling Functions
Stride Time Arryth.
Peak Ang. Vel.
Swing Ang. Range
q
t
n
Multi-Frame Label Model
Probabilistic Training Labels
85% Freezing
15% Nonfreezing
LSTM End Model
Splitting Raw Data into Candidates
Apply LFs to candidates
Label Tensor of form: {0,1,2}^{q x t x n}
Training labels for each candidate

Figure 1. Full multi-frame weak supervision pipeline: from raw data to training the end model.

reduce the amount of hand-labeled data required by a factor of $4\times$, and come within 4.8 F1 points of a fully supervised model that uses all the hand-labeled data.

2. Methodology

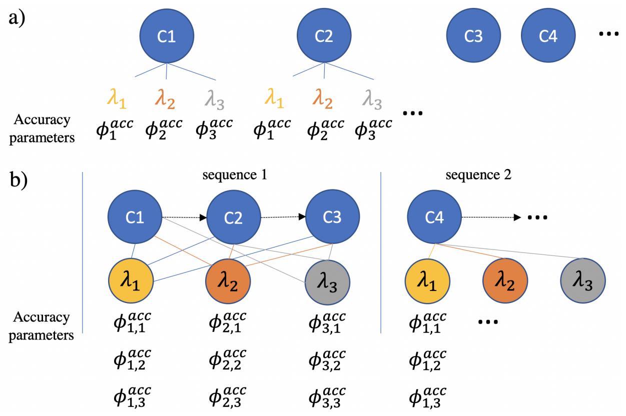


Figure 2. Example with 3 labeling functions ($\lambda_1, \lambda_2, \lambda_3$) voting on candidates $C_1, ..., C_m$ in a (a) classical weak supervision setup and (b) multi-frame weak supervision setup (sequence length 3).

2.1. Multi-frame Weak Supervision

In weak supervision, noisy training labels are programmatically generated for unlabeled data using several heuristic labeling functions which encode specific domain knowledge. These labeling functions are modeled using a generative process which allows us to denoise the labels by learning their correlation structure and accuracies (Ratner et al., 2016).

However, current weak supervision paradigms model each data point as being independent (See Figure 2a), making them inappropriate for the temporally correlated data present in time series problems. To model these correlations and dependencies, we adapt a multi-task weak supervision approach to time series data, where successive data points, or candidates, are treated as different tasks. We refer to this approach as *multi-frame weak supervision*. In this setting, we divide up our data into q small sequences which consist of t candidates or tasks. If we need to label m total candidates, and $m \mod t \neq 0$, we pad the last sequence with abstain labels so that every sequence is of length t. In each sequence, we treat each candidate as a separate task, where all the tasks are correlated in some manner.

To label the candidates in these sequences, we use labeling functions $\lambda_i : \mathcal{X} \rightarrow \mathcal{Y} \cup \emptyset$ for $i = 1...n$ which takes in one or more candidates $x_i..x_{i+K} \in \mathcal{X}$, and output a label $y \in \mathcal{Y}$ or $\emptyset$, if the function abstains, . Using n labeling functions on q sequences with t candidates each, we create a 3D label tensor $L = (\mathcal{Y} \cup \emptyset)^{q \times t \times n}$ (See Figure 1).

Then, within each of the q sequences, we define a dependency structure among the t tasks. This structure is a chain dependency between the tasks, where the second task depends on the first, the third on the second, and so on. In this new setting, each labeling function $\lambda_i, i = 1...n$ has an accuracy parameter $\phi_{i,j}^{Acc}$, where $j = 1...t$. Here, each labeling function has t accuracy parameters, one for each index in the sequence. This is in direct contrast to a traditional weak supervision setting, where each labeling function learns a single accuracy parameter ϕ_i^{Acc} across all the candidates (See Figure 2). By having a different accuracy parameter per labeling function per candidate in a sequence, we can capture information about the correlations among the candidates within a sequence.

Using our label matrix L, we learn a label model $P_{\phi^{acc}}(Y|L)$ that is parameterized by these accuracy parameters and uses our chain dependency structure (Ratner et al., 2018). With this label model, we generate probabilistic labels for our candidates which we use to train a discriminative model that we aim to generalize beyond the information encoded in the labeling functions. We do this by minimizing the expected loss with respect to $\tilde{Y}$:

$$\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{m} \mathbb{E}_{y \sim \tilde{Y}} l(h_\theta(x_i), y_i)$$

With our multi-frame weakly supervised label model, we can better model the temporal correlations between successive candidates and more accurately assign probabilistic labels compared to traditional weak supervision models.

2.2. End Model

We then train a discriminative model on the probabilistic labels generated from the label model. We use a single layer bi-directional LSTM and hidden state dimension 300 for our end model that takes in a multivariate sensor stream as input. In order to provide longer temporal context, we pass in a windowed version of each candidate that includes past and future frames. Window size was tuned empirically, with [-3,+1] performing best overall. Since the sequence length of each frame slightly varies, we then pad these sequences (with 0's) and truncate any sequences over a pre-defined maximum sequence length. To provide more contextual signal, we also add multiplicative attention to pool over the hidden states in the LSTM. Note, even with the windowed version of each candidate, the classification task is still to make a prediction on just the candidate itself. See Figure 1 for a depiction of the full process of training the end model.

3. Labeling Patient Data

3.1. Dataset

To test our approach, we use a dataset that contains series of wearable sensor measurements from 36 trials from 9 patients that have Parkinson's Disease (PD) and exhibit freezing behavior. PD is a neurodegenerative disease marked by tremor, loss of balance, and other motor impairments, that affects over 10 million people worldwide. Freezing of gait (FOG) — a sudden and brief episode where an individual is unable to produce effective forward stepping (Giladi et al., 1992; 1997) —is one of the disabling problems caused by PD, and often leads to falls (Bloem et al., 2004).

In this dataset, subjects walked in a laboratory setting that the investigators designed to elicit freezing events. The average trial length was about 2.5 minutes. Leg or shank angular velocity was measured during the forward walking task using wearable inertial measurement units (sampled at 128 Hz), which were positioned in a standardized manner for all subjects and tasks on both shanks (lower leg).

3.2. Preprocessing

From each trial, we extract left and right shank gyroscope data in the z-direction, along with the respective gold labels, which were manually recorded by a neurologist. We combine the data from all trials, and segment the sensor data by *gait cycle*, i.e., the time interval required for one foot to make successive contact with the ground. Gait cycle time is computed analytically from the right shank sensor data and is defined as the time period between two successive peaks on an angular velocity versus time plot.

Algorithm 1 Example labeling function

```
**Input:** Candidate $x_i$, size $p \times l$
**if** $arrhythmicity(x_i) > 0.55$ **then** $2$ (Non-freezing)
**else if** $arrhythmicity(x_i) < 0.15$ **then** $1$ (Freezing)
**else** $0$ (Abstain)
```

We then define a single candidate to be $x^{p \times l} \in \mathcal{X}$ where p is the number of sensor streams and l is the sequence length. For our task, $p = 2$ since we use the left and right shank sensor streams, and l is the sequence length for a single gait cycle ($\sim$1.2 seconds). With this definition, our dataset is composed of approximately 3500 candidates. Our binary classification task is to predict a label $y \in \mathcal{Y} = \{1, 2\}$, where we define $y = 1$ to indicate freezing behavior and $y = 2$ to indicate non-freezing behavior.

3.3. Labeling Functions

To programatically label data, we use five labeling functions with varying temporal granularity which draw on biomechanical domain knowledge and empirical observations. Specifically, these labeling functions target features which can distinguish freezing and non-freezing events.

For all labeling functions, we assign positive, negative, or abstain labels based on empirically measured threshold values from the validation set. For example, one heuristic we employ uses stride time arrhythmicity (Plotnik et al., 2005; 2007), which we calculate as average coefficient of variation for the past 3 stride times of the left and right leg. For this function, we label a candidate as *freezing* if the arrhythmicity of that candidate is greater than 0.55, and *not freezing* if the arrhythmicity is less than 0.15 (See Algorithm 1). If arrhythmicity for a particular candidate is in between these two values, we abstain. In addition to stride time arrhythmicity (LF1), other labeling functions we use involve the swing angular range of the shank (LF2, LF3), and the amplitude and variance in shank angular velocity (LF4, LF5). See Table 1 for the individual performance of each labeling function on the validation set (data split described in Section 4.1) .

4. Experimental Evaluation

4.1. Experimental Setup

We create training/validation/test sets by splitting on patient trials/sessions. In this setting, both the validation and test set have a single trial from each patient, and the training set has one or more trials from each patient. With this split, our training set consists of about 1800 candidates and our

Table 1. Labeling function (LF) evaluation results, where coverage is the percentage of the validation set that the LF does not abstain on, and empirical accuracy is the accuracy as compared to the hand-labels (ground truth).

LFs	Coverage (%)	Emp. Acc. (%)	F1
LF1	48.9	53.9	56.2
LF2	71.2	55.3	64.3
LF3	71.9	59.6	60.5
LF4	42.7	78.1	75.5
LF5	75.4	63.2	65.0

validation set has about 600 candidates. For both weak supervision settings, we treat the training set as an unlabeled dataset, and only use hand labels from the validation set. Using the respective label models, we then generate probabilistic training labels for the training set to use in the end model. In the fully supervised setting, we train with the hand-labeled training set and validate with the hand-labeled validation set.

4.2. Label Model Results

Using the labeling functions described in Section 3.3, we build factor graph-based label models in both the classical weak supervision and multi-frame weak supervision settings, and predict probabilistic labels $y \in \mathcal{Y}$ for each candidate $x^{p \times l} \in \mathcal{X}$ in the training set (Figure 1). In the multi-frame setting, we test our label models (on the validation set) with sequence lengths of 3 and 5 candidates, and experiment with different class balance priors. Our best multi-frame label model used a sequence length of 3 candidates with a class balance prior derived from the validation set.

We then compare performance to a simple ensembling of the labeling functions (Majority Vote) by directly using our label models on the test set. The performance of these label models on the test set are summarized in Table 2. We note that our multi-frame label model beats a classic weak supervision label model by 3.1 F1 points, and 14.7 accuracy points. Our multi-frame model does exhibit a precision/recall trade-off, however, as we get a higher precision but a lower recall than a classical weak supervision and majority vote label model. In practice, higher precision in the label model is helpful in weak supervision settings, since the end model can generalize better as models are scaled with more unlabeled data (Ratner et al., 2017).

4.3. End Model Results

We evaluate our end model in both the weakly supervised and multi-frame weakly supervised setting, and compare performance with that of a fully supervised model. We also compare performance with training an end model on probabilistic labels generated using a Majority Vote (MV) ensemble of the labeling functions (Table 3), as well as just using the trained label models directly (Table 2).

Table 2. Label Model (LM) test set performance

Label Model Type	P	R	F1	Acc
Majority Vote	39.7	**90.3**	55.2	52.4
Classical LM	41.5	84.2	55.6	56.3
Multi-frame LM	**54.6**	63.5	**58.7**	**71.0**

Table 3. End model (LSTM) test set performance

End Model Type	P	R	F1	Acc
Fully Supervised	*62.0*	*84.0*	*71.0*	*77.8*
MV (with End Model)	46.3	**92.1**	61.6	62.7
Classical WS	48.3	86.3	61.9	65.6
Multi-frame WS	**60.2**	73.6	**66.2**	**75.6**

From Table 3, we note that our weakly supervised model comes within 10 F1 points of a fully supervised model. This model also beats using both ensembles of labeling functions, the classical label model and majority vote, directly on the test set by more than 6 F1 points (See Table 2). Further, our multi-frame weakly supervised model improves upon this classical weakly supervised model by 4.3 F1 points and 10.0 points in accuracy. This multi-frame weakly supervised model comes within just 4.8 F1 points and 2.2 accuracy points of the fully supervised model. As with the label models, we note similar precision/recall trade-offs with our multi-frame weak supervision end model: comparing with the majority vote end model, we increase precision by 13.9 points at the expense of 18.5 points in recall, yielding an overall improvement of 4.6 F1 points.

5. Conclusion and Next Steps

Our work demonstrates the potential of multi-frame weak supervision on time series tasks. In our experiments, our multi-frame weakly supervised model performed close to fully supervised models, but used just a fourth (∼600 out of ∼2400) of the number of hand labels. This model also performed better than a classical weakly supervised model. Further, the amount of data available for our gait freezing task was fairly small — with more unlabeled data, we expect to see improved performance in both weak supervision settings.

In the future, we plan to add more and different types of sensor streams and modalities (e.g., video). We are also interested in generating more unlabeled data from more novel supervision sources, such as incorporating biomechanical simulation data (Seth et al., 2018).

References

Bloem, B. R., Hausdorff, J. M., Visser, J. E., and Giladi, N. Falls and freezing of gait in parkinson's disease: a review of two interconnected, episodic phenomena. *Movement disorders: official journal of the Movement Disorder Society*, 19(8):871–884, 2004.

Cappon, G., Acciaroli, G., Vettoretti, M., Facchinetti, A., and Sparacino, G. Wearable continuous glucose monitoring sensors: A revolution in diabetes treatment. *Electronics*, 6(3):65, 2017.

Casilari, E., Santoyo-Ramón, J.-A., and Cano-García, J.-M. Analysis of public datasets for wearable fall detection systems. *Sensors*, 17(7):1513, 2017.

Fries, J. A., Varma, P., Chen, V. S., Xiao, K., Tejeda, H., Saha, P., Dunnmon, J., Chubb, H., Maskatia, S., Fiterau, M., et al. Weakly supervised classification of rare aortic valve malformations using unlabeled cardiac mri sequences. *BioRxiv*, pp. 339630, 2018.

Giladi, N., McMahon, D., Przedborski, S., Flaster, E., Guillory, S., Kostic, V., and Fahn, S. Motor blocks in parkinson's disease. *Neurology*, 42(2):333–333, 1992.

Giladi, N., Kao, R., and Fahn, S. Freezing phenomenon in patients with parkinsonian syndromes. *Movement disorders: official journal of the Movement Disorder Society*, 12(3):302–305, 1997.

Halilaj, E., Rajagopal, A., Fiterau, M., Hicks, J. L., Hastie, T. J., and Delp, S. L. Machine learning in human movement biomechanics: best practices, common pitfalls, and new opportunities. *Journal of biomechanics*, 2018.

Kumari, P., Mathew, L., and Syal, P. Increasing trend of wearables and multimodal interface for human activity monitoring: A review. *Biosensors and Bioelectronics*, 90: 298–307, 2017.

Plotnik, M., Giladi, N., Balash, Y., Peretz, C., and Hausdorff, J. M. Is freezing of gait in parkinson's disease related to asymmetric motor function? *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, 57(5):656–663, 2005.

Plotnik, M., Giladi, N., and Hausdorff, J. M. A new measure for quantifying the bilateral coordination of human gait: effects of aging and parkinsons disease. *Experimental brain research*, 181(4):561–570, 2007.

Ratner, A., Bach, S. H., Ehrenberg, H., Fries, J., Wu, S., and Ré, C. Snorkel: Rapid training data creation with weak supervision. *Proceedings of the VLDB Endowment*, 11 (3):269–282, 2017.

Ratner, A., Hancock, B., Dunnmon, J., Sala, F., Pandey, S., and Ré, C. Training complex models with multi-task weak supervision. *arXiv preprint arXiv:1810.02840*, 2018.

Ratner, A. J., De Sa, C. M., Wu, S., Selsam, D., and Ré, C. Data programming: Creating large training sets, quickly. In *Advances in neural information processing systems*, pp. 3567–3575, 2016.

Seth, A., Hicks, J. L., Uchida, T. K., Habib, A., Dembia, C. L., Dunne, J. J., Ong, C. F., DeMers, M. S., Rajagopal, A., Millard, M., et al. Opensim: Simulating musculoskeletal dynamics and neuromuscular control to study human and animal movement. *PLoS computational biology*, 14 (7):e1006223, 2018.

Tison, G. H., Sanchez, J. M., Ballinger, B., Singh, A., Olgin, J. E., Pletcher, M. J., Vittinghoff, E., Lee, E. S., Fan, S. M., Gladstone, R. A., et al. Passive detection of atrial fibrillation using a commercially available smartwatch. *JAMA cardiology*, 3(5):409–416, 2018.

Xiao, T., Xia, T., Yang, Y., Huang, C., and Wang, X. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2691–2699, 2015.