Weakly-Supervised Temporal Localization via Occurrence Count Learning

Julien Schroeter
schroeterj1@cardiff.ac.uk

Dr Kirill Sidorov
Prof David Marshall
Weakly-Supervised Temporal Localization via Occurrence Count Learning | Julien Schroeter | Kirill Sidorov | David Marshall
CONTEXT
Temporal Localization

Input Data

Temporal Sequence

Target

DNN

Precise Localization

Impulse-like Events

Weakly-Supervised Temporal Localization via Occurrence Count Learning | Julien Schroeter | Kirill Sidorov | David Marshall
Weakly-Supervised Temporal Localization via Occurrence Count Learning | Julien Schroeter | Kirill Sidorov | David Marshall

**CONTEXT**

Temporal Localization

---

Fully-Supervised
Weakly-Supervised Temporal Localization via Occurrence Count Learning | Julien Schroeter | Kirill Sidorov | David Marshall
Temporal Localization

Training Labels

Input Data

DNN

Fully-Supervised
OBJECTIVE
Weakening the annotation requirement

Our approach
OBJECTIVE
Weakening the annotation requirement

Input Data

LoCo
Training

Our approach
OBJECTIVE
Weakening the annotation requirement

Our approach
OBJECTIVE
Weakening the annotation requirement

Input Data
Training Labels
Occurrence Counts

Our approach

LoCo
Training
OBJECTIVE
Weakening the annotation requirement

Training

Inference

LoCo

Occurrence Counts

3 4 6

LoCo
**OBJECTIVE**

Weakening the annotation requirement

---

**Training**

Input

---

**Inference**

LoCo

LoCo

Occurrence Counts

3 4 6

Weakly-Supervised Temporal Localization via Occurrence Count Learning | Julien Schroeter | Kirill Sidorov | David Marshall
OBJECTIVE
Weakening the annotation requirement

Input

Output

LoCo

LoCo

Occurrence Counts

3 4 6

Training

Inference
OBJECTIVE
Weakening the annotation requirement

Input

Training

LoCo

Output

Inference

LoCo

Occurrence Counts

Input

Occurrence Counts

Output

5 4 9

Weakly-Supervised Temporal Localization via Occurrence Count Learning | Julien Schroeter | Kirill Sidorov | David Marshall
OBJECTIVE
Weakening the annotation requirement

Input
LoCo
Occurrence Counts
Precise Localization

Output
LoCo
Occurrence Counts

Training

Inference

Weakly-Supervised Temporal Localization via Occurrence Count Learning | Julien Schroeter | Kirill Sidorov | David Marshall
**OBJECTIVE**

Weakening the annotation requirement

**Weakly-Supervised Temporal Localization via Occurrence Count Learning**

- **Training**
- **Inference**

**Input**

Occurrence Counts → LoCo

**Output**

LoCo → Precise Localization

Occurrence Counts

**Weakly-Supervised**
Is it useful?
OBJECTIVE
Weakening the annotation requirement

Label Piano Music
OBJECTIVE
Weakening the annotation requirement
OBJECTIVE
Weakening the annotation requirement

Precise hand-labeling is very tedious
OBJECTIVE
Weakening the annotation requirement

- Precise hand-labeling is very tedious
- Prone to labeling inaccuracy
OBJECTIVE
Weakening the annotation requirement
OBJECTIVE
Weakening the annotation requirement

How many notes per pitch?
OBJECTIVE
Weakening the annotation requirement

How many notes per pitch?
**OBJECTIVE**
Weakening the annotation requirement

How many notes per pitch?
OBJECTIVE
Weakening the annotation requirement

How many notes per pitch?
OBJECTIVE
Weakening the annotation requirement

How many notes per pitch?
Unlike existing methods, in which localization is explicitly achieved by design, our model learns localization *implicitly* as a byproduct of learning to count instances.
MODEL
Counting Occurrences

\[ p_i(t) = f \left( \left( x_i(n) \right)_{n=1}^{t} \right) \]

Probability of Event occurrence
\[ p_i(t) = f \left( \left( x_i(n) \right)_{n=1}^t \right) \]

\[ E_i(t) = \mathcal{B} \left( p_i(t) \right), \text{ ind. Bernoulli} \]
MODEL
Counting Occurrences

\[ p_i(t) = f \left( \left( x_i(n) \right)_{n=1}^t \right) \]

\[ E_i(t) = \mathcal{B} \left( p_i(t) \right), \text{ ind. Bernoulli} \]

\[ Y_i = \sum_t E_i(t) \]

Occurrence Count
\[ p_i(t) = f \left( \left( x_i(n) \right)_{n=1}^t \right) \]

\[ E_i(t) = \mathcal{B}(p_i(t)), \text{ ind. Bernoulli} \]

\[ Y_i = \sum_{t} E_i(t) \]
**MODEL**

**Counting Occurrences**

**Input Data**

\[ p_i(t) = f \left( \left( x_i(n) \right)_{n=1}^t \right) \]

\[ E_i(t) = \mathcal{B} (p_i(t)), \text{ ind. Bernoulli} \]

**Occurrence Count**

\[ Y_i = \sum_t E_i(t) \]
**MODEL**

**Counting Occurrences**

**Input Data**

\[ p_i(t) = f \left( \left( x_i(n) \right)_{n=1}^t \right) \]

\[ E_i(t) = \mathcal{B} \left( p_i(t) \right), \text{ ind. Bernoulli} \]

\[ Y_i = \sum_t E_i(t) \]

**Occurrence Count**
**Model**

Counting Occurrences

Estimated through RNN (e.g. LSTM)

Input Data

\[ p_i(t) = f\left(\left(\mathbf{x}_i(n)\right)_{n=1}^t\right) \]

\[ E_i(t) = \mathbb{B}(p_i(t)), \text{ ind. Bernoulli} \]

\[ Y_i = \sum_t E_i(t) \]
MODEL

Loss

\[ Y_i = \sum_t E_i(t) \]

Occurrence Count

Weakly-Supervised Temporal Localization via Occurrence Count Learning | Julien Schroeter | Kirill Sidorov | David Marshall
\[ Y_i = \sum_{t} E_i(t) \]

Compare them to true observed counts.
\[ Y_i = \sum_t E_i(t) \]

**Occurrence Count**

Compare them to true observed counts.

\[ L(\theta) = -\sum \log \left( \Pr \left( Y_i, \theta = y_i | X_i \right) \right) \]
\[ Y_i = \sum_t E_i(t) \]

**Occurrence Count**

Compare them to true observed counts.

\[ L(\theta) = - \sum \log \left( \Pr \left( Y_i, \theta = y_i \mid X_i \right) \right) \]

**Observed Count**
\[ Y_i = \sum_t E_i(t) \]

Comparison of occurrence counts to true observed counts.

\[ L(\theta) = -\sum \log \left( \Pr(Y_i, \theta = y_i | X_i) \right) \]

\[ L(\theta) = -\log \left( \frac{1}{2} \right) - \log \left( \frac{1}{3} \right) - \log \left( \frac{1}{4} \right). \]
\[ Y_i = \sum_t E_i(t) \]

**Occurrence Count**

Compare them to true observed counts.

\[ L(\theta) = - \sum \log \left( \Pr \left( Y_i, \theta = y_i \mid X_i \right) \right) \]

Optimized with standard backpropagation
MODEL
Full Pipeline

\[ L(\theta) = -\log(\cdot) - \log(\cdot) - \log(\cdot) \]
Why does it work?
Y follows a Poisson-binomial distribution
\[ \mathcal{Y}_i(k, t) := \Pr(Y_{i, \theta}(t) = k) \]

Bin $k$ of count
distribution at time $t$
\[ Y_i(k, t) := \Pr(Y_{i, \theta}(t) = k) \]

**Property 2** (Recursion on \( k, t \))

\[
Y_i(k, t) = \begin{cases} 
(1 - p_i(t))Y_i(k, t-1) & k = 0 \\
(1 - p_i(t))Y_i(k, t-1) + p_i(t)Y_i(k-1, t-1) & k > 0 
\end{cases}
\]

where \( Y_i(k, 0) = 1_{k=0} \).
Property 2 (Recursion on $k, t$)

\[
\Upsilon_i(k, t) = \begin{cases} 
(1-p_i(t))\Upsilon_i(k, t-1) & k=0 \\
(1-p_i(t))\Upsilon_i(k, t-1) + p_i(t)\Upsilon_i(k-1, t-1) & k>0 
\end{cases}
\]

where $\Upsilon_i(k, 0) = 1_{k=0}$.
Property 2 (Recursion on $k, t$)

$$\Upsilon_i(k, t) = \begin{cases} 
(1-p_i(t))\Upsilon_i(k, t-1) & k=0 \\
(1-p_i(t))\Upsilon_i(k, t-1) + p_i(t)\Upsilon_i(k-1, t-1) & k > 0 
\end{cases}$$

where $\Upsilon_i(k, 0) = 1_{k=0}$.

\[ p_i = 6\% \]
\( p_i = 6\% \)

**Property 2** (Recursion on \( k, t \))

\[
\gamma_i(k, t) = \begin{cases} 
(1-p_i(t))\gamma_i(k, t-1) & \text{if } k=0 \\
(1-p_i(t))\gamma_i(k, t-1) + p_i(t)\gamma_i(k-1, t-1) & \text{if } k>0 
\end{cases}
\]

where \( \gamma_i(k, 0) = \mathbb{1}_{k=0} \).

\[ (9) \]
Property 2 (Recursion on $k$, $t$)

$$
\Upsilon_i(k, t) = \begin{cases} 
(1-p_i(t))\Upsilon_i(k, t-1) & k=0 \\
(1-p_i(t))\Upsilon_i(k, t-1) + p_i(t)\Upsilon_i(k-1, t-1) & k > 0 
\end{cases}
$$

(9)

where $\Upsilon_i(k, 0) = 1_{k=0}$. 

$p_i = 30\%$
Property 2 (Recursion on $k$, $t$)

\[ Y_i(k, t) = \begin{cases} 
(1 - p_i(t)) Y_i(k, t - 1) & \text{if } k = 0 \\
(1 - p_i(t)) Y_i(k, t - 1) + p_i(t) Y_i(k - 1, t - 1) & \text{if } k > 0 
\end{cases} \]

where $Y_i(k, 0) = 1_{k=0}$. 

$p_i = 30\%$
Property 2 (Recursion on $k, t$)

\[
\gamma_i(k, t) = \begin{cases} 
(1-p_i(t))\gamma_i(k, t-1) & k=0 \\
(1-p_i(t))\gamma_i(k, t-1) + p_i(t)\gamma_i(k-1, t-1) & k > 0 
\end{cases}
\]

where $\gamma_i(k, 0) = \mathbb{1}_{k=0}$.
Weakly-Supervised Temporal Localization via Occurrence Count Learning | Julien Schroeter | Kirill Sidorov | David Marshall

**MODEL**

Recursion

\[ p_i = 30\% \]

\[ \gamma_i(k, t) = \begin{cases} (1 - p_i(t)) \gamma_i(k, t - 1) & k = 0 \\ (1 - p_i(t)) \gamma_i(k, t - 1) + p_i(t) \gamma_i(k - 1, t - 1) & k > 0 \end{cases} \]

\[ \gamma_i(k, 0) = 1_{k=0}. \]
\( p_i = 30\% \)

**Property 2** (Recursion on \( k, t \))

\[
\Upsilon_i(k, t) = \begin{cases} 
(1 - p_i(t)) \Upsilon_i(k, t-1) & k = 0 \\
(1 - p_i(t)) \Upsilon_i(k, t-1) + p_i(t) \Upsilon_i(k-1, t-1) & k > 0
\end{cases}
\]

where \( \Upsilon_i(k, 0) = 1_{k=0} \).
Property 1 (Mass shift irreversibility)

\[(X_{i},\theta(t))_{i=1}^{T_{i}} \text{ is monotonically increasing.}\]
**Property 1 (Mass shift irreversibility)**

\[(Y_i, \theta(t))_{t=1}^{T_i} \text{ is monotonically increasing.}\]

Mass moves to the right
Property 1 (Mass shift irreversibility)

\((Y_{i,t}(t))^{T}_{t=1}\) is monotonically increasing.

Mass moves to the right
**Property 1** (Mass shift irreversibility)

\[(Y_i, \theta(t))_{t=1}^{T_i} \text{ is monotonically increasing.}\]
Property 1 (Mass shift irreversibility)

\[ (Y_i, \theta(t))_{t=1}^{T_i} \] is monotonically increasing.

Mass moves to the right
**Property 1** (Mass shift irreversibility)

\[
(Y_{i,\theta(t)})_{t=1}^{T_i} \text{ is monotonically increasing.}
\]
Property 1 (Mass shift irreversibility)

\[(Y_{i,\theta(t)})_{t=1}^{T_i}\] is monotonically increasing.

Mass moves to the right
**Property 1** (Mass shift irreversibility)

\[
\left( Y_i, \theta(t) \right)_{t=1}^{T_i} \text{ is monotonically increasing.}
\]
.Property 1 (Mass shift irreversibility)

\[(Y_i, \theta(t))^{T_i}_{t=1}\] is monotonically increasing.

Mass moves to the right
**Model**

No early triggering

**Property 1 (Mass shift irreversibility)**

\[(Y_{i,\theta(t)})_{t=1}^{T_i} \text{ is monotonically increasing.}\]

Consequence: Mass shifts are irreversible
**Property 1** (Mass shift irreversibility)

\[(Y_{i,\theta}(t))_{t=1}^{T_i}\] is monotonically increasing.

**Consequence:** Mass shifts are irreversible

- prevents the model from triggering early
- prevents the model from false alarms
Lemma 2 (First upper bound)

\[
\max_k \gamma_i(k, t) \leq \frac{1}{2} + \min_{j \leq t} \| \frac{1}{2} - p_i(j) \|
\]
Lemma 2 (First upper bound)

\[
\max_k \gamma_i(k, t) \leq \frac{1}{2} + \min_{j \leq t} \left\| \frac{1}{2} - p_i(j) \right\|
\]

\[
L(\theta) = - \sum_i \log (\Pr (Y_{i,\theta} = y_i \mid X_i)) \quad \text{Counting Loss}
\]
Lemma 2 (First upper bound)

\[
\max_k \gamma_i(k, t) \leq \frac{1}{2} + \min_{j \leq t} \left\| \frac{1}{2} - p_i(j) \right\|
\]

\[
L(\theta) = - \sum_i \log \left( \Pr \left( Y_{i,\theta} = y_i \mid X_i \right) \right)
\]

\[
= - \sum_i \log \left( \gamma_i(y_i, T_i) \right)
\]

Counting Loss
**Lemma 2 (First upper bound)**

\[
\max_k \gamma_i(k, t) \leq \frac{1}{2} + \min_{j \leq t} \| \frac{1}{2} - p_i(j) \|
\]

\[
L(\theta) = - \sum_i \log \left( \Pr \left( Y_i, \theta = y_i \mid X_i \right) \right)
\]

\[
= - \sum_i \log \left( \gamma_i(y_i, T_i) \right)
\]

\[
\geq - \sum_i \log \left( \frac{1}{2} + \min_{j \leq t} \| \frac{1}{2} - p_i(j) \| \right)
\]
**Lemma 2** (First upper bound)

\[
\max_k \mathcal{Y}_i(k, t) \leq \frac{1}{2} + \min_{j \leq t} \left\| \frac{1}{2} - p_i(j) \right\|
\]

**Learns to count**

\[
L(\theta) = -\sum_i \log \left( \Pr \left( Y_i, \theta = y_i \mid X_i \right) \right)
= -\sum_i \log \left( \mathcal{Y}_i(y_i, T_i) \right)
\geq -\sum_i \log \left( \frac{1}{2} + \min_{j \leq t} \left\| \frac{1}{2} - p_i(j) \right\| \right)
\]

**Counting Loss**
**Model**

**Mass Convergence**

\[
\text{Lemma 2 (First upper bound)}
\]

\[
\max_k \mathcal{Y}_i(k, t) \leq \frac{1}{2} + \min_{j \leq t} \left\| \frac{1}{2} - p_i(j) \right\|
\]

*Learn to count*

\[
L(\theta) = - \sum_i \log \left( \Pr \left( Y_{i,\theta} = y_i \mid X_i \right) \right)
\]

\[
= - \sum_i \log \left( \mathcal{Y}_i(y_i, T_i) \right)
\]

\[
\geq - \sum_i \log \left( \frac{1}{2} + \min_{j \leq t} \left\| \frac{1}{2} - p_i(j) \right\| \right)
\]

*Counting Loss*

*Converge towards 0,1 extremes*
**Property 3** (Sparse mass concentration) The inequality derived below reveals that, as the loss decreases, small $p_i(\cdot)$ will quickly converge towards zero.

\[
\max_k \gamma_i(k, t) \overset{(8)}{\leq} \min_{l \leq t} \max_k \gamma_i(k, l) \overset{\text{ind}}{=} \min_{\sigma, l \leq t} \max_k \gamma_i,\sigma(k, l)
\]

\[
\text{Le Cam} \leq \min_{\sigma, l \leq t} \max_k \frac{\lambda_{i,\sigma, l}^k e^{-\lambda_{i,\sigma, l}}}{k!} + 2 \sum_{j=1}^l p_{i,\sigma}(j)^2
\]

\[
\overset{\text{def}}{=} \min_{\sigma, l \leq t} \max_k \left[ \sum_{j=1}^l p_{i,\sigma}(j) \right]^k e^{-\left[ \sum_{j=1}^l p_{i,\sigma}(j) \right]} \frac{1}{k!} + 2 \sum_{j=1}^l p_{i,\sigma}(j)^2,
\]
Property 3 (Sparse mass concentration) The inequality derived below reveals that, as the loss decreases, small $p_i(\cdot)$ will quickly converge towards zero.

\[
\max_k \gamma_i(k, t) \leq \min_{l \leq t} \max_k \gamma_i(k, l) = \min_{\sigma, l \leq t} \max_k \gamma_i,\sigma(k, l)
\]

\[
\text{Le Cam } \leq \min_{\sigma, l \leq t} \max_k \frac{\lambda_{i,\sigma,l}^k e^{-\lambda_{i,\sigma,l}}}{k!} + 2 \sum_{j=1}^l p_{i,\sigma}(j)^2
\]

\[
def \min_{\sigma, l \leq t} \max_k \left[ \frac{\left( \sum_{j=1}^l p_{i,\sigma}(j) \right)^k e^{-\left( \sum_{j=1}^l p_{i,\sigma}(j) \right)}}{k!} \right] + 2 \sum_{j=1}^l p_{i,\sigma}(j)^2,
\]

A detection cannot be split into numerous small $p_i(\cdot)$ contributions
As the model learns to count event occurrences:
MODEL
Mass Convergence

As the model learns to count event occurrences:

• $p_i(\cdot)$ converge towards 0,1 extremes
As the model learns to count event occurrences:

- $p_i(\cdot)$ converge towards 0,1 extremes
- A detection cannot be split into numerous small $p_i(\cdot)$ contributions
As the model learns to count event occurrences:

- \( p_i(\cdot) \) converge towards 0, 1 extremes
- A detection cannot be split into numerous small \( p_i(\cdot) \) contributions

A single \( p_i(\cdot) \) will contain almost all of them mass for an event.
1. Almost **binary** predictions
1. Almost **binary** predictions

2. No **early** triggering
MODEL Properties

1. Almost binary predictions
2. No early triggering
3. No systematic late bias \(\longrightarrow\) Not a theoretical property
MODEL Properties

1. Almost **binary** predictions

2. No **early** triggering

3. No systematic **late** bias  
   \(\text{Not a theoretical property}\)

   Achieved through an implementation trick:
1. Almost **binary** predictions

2. No **early** triggering

3. No systematic **late** bias \(\text{Not a theoretical property}\)

Achieved through an implementation trick: **Feeding sequences of variable length**
1. Almost **binary** predictions
2. No **early** triggering
3. No systematic **late** bias
MODEL
Properties

1. Almost binary predictions

2. No early triggering

3. No systematic late bias

If the model accurately learns to count occurrences and if the events are detectable, then a coherent localization will emerge naturally.
Experiments
DRUM DETECTION
Experiment Specifications

Detection of three different drum types in drum audio extracts
Detection of three different drum types in drum audio extracts

- Tight tolerance of 50ms for a prediction to be correct
Detection of three different **drum** types in drum audio extracts

- **Tight tolerance** of 50ms for a prediction to be correct
- Comparison with **fully-supervised benchmark** models
**DRUM DETECTION**

Our approach

![Signal](image)
DRUM DETECTION

Our approach

Signal

Mel-spectrogram

1st order derivative

Fourier
DRUM DETECTION

Our approach

- Signal
- Mel-spectrogram
- 1st order derivative
- Fourier
- CNNs

Convolutional Representations

Our approach

- DRUM DETECTION

Convolutional Representations

- Fourier

Weakly-Supervised Temporal Localization via Occurrence Count Learning | Julien Schroeter | Kirill Sidorov | David Marshall
DRUM DETECTION

Our approach

Signal

Mel-spectrogram

1st order derivative

Fourier

CNNs

Convolutional Representations

LSTM
Our approach

Signal
Mel-spectrogram
1\textsuperscript{st} order derivative

Convolutional Representations

CNNs
LSTM
FCs
DRUM DETECTION

Our approach

- Signal
- Mel-spectrogram
- 1st order derivative
- Fourier
- CNNs
- Convolutional Representations
- LSTM
- FCs
- Localization
DRUM DETECTION

Our approach

- Signal
- Mel-spectrogram
- 1st order derivative

Convolutional Representations
- Trained with our loss (using only occurrence counts)

Localization

Fourier

CNNs

LSTM

FCs
# DRUM Detection

## Results

<table>
<thead>
<tr>
<th>Method</th>
<th>KD</th>
<th>SD</th>
<th>HH</th>
<th>PRE</th>
<th>REC</th>
<th>F₁</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Random</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RNN</td>
<td>94.7</td>
<td>79.5</td>
<td>88.3</td>
<td>84.1</td>
<td><strong>93.3</strong></td>
<td>87.5</td>
</tr>
<tr>
<td>TANHB</td>
<td>92.4</td>
<td>84.6</td>
<td>87.1</td>
<td>86.3</td>
<td>92.1</td>
<td>88.0</td>
</tr>
<tr>
<td>RELU₂</td>
<td>91.3</td>
<td>83.8</td>
<td>85.2</td>
<td>83.7</td>
<td>92.3</td>
<td>86.8</td>
</tr>
<tr>
<td>LSTM₉</td>
<td>94.4</td>
<td>84.1</td>
<td>91.4</td>
<td>90.8</td>
<td>90.8</td>
<td><strong>90.0</strong></td>
</tr>
<tr>
<td>GRU₂</td>
<td>94.2</td>
<td><strong>87.1</strong></td>
<td>87.7</td>
<td>88.6</td>
<td>92.7</td>
<td>89.7</td>
</tr>
<tr>
<td>ours (LoCo)</td>
<td>92.3</td>
<td>81.2</td>
<td><strong>93.0</strong></td>
<td><strong>90.9</strong></td>
<td>87.1</td>
<td>88.9</td>
</tr>
<tr>
<td><strong>Subset</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RNN</td>
<td><strong>91.0</strong></td>
<td>57.8</td>
<td>82.2</td>
<td>72.8</td>
<td><strong>88.3</strong></td>
<td>77.0</td>
</tr>
<tr>
<td>TANHB</td>
<td>82.7</td>
<td>61.6</td>
<td>84.8</td>
<td>74.1</td>
<td>83.8</td>
<td>76.4</td>
</tr>
<tr>
<td>RELU₂</td>
<td>79.4</td>
<td>62.1</td>
<td>80.8</td>
<td>69.6</td>
<td>84.2</td>
<td>74.1</td>
</tr>
<tr>
<td>LSTM₉</td>
<td>85.8</td>
<td><strong>68.8</strong></td>
<td>83.7</td>
<td>78.3</td>
<td>84.7</td>
<td><strong>79.4</strong></td>
</tr>
<tr>
<td>GRU₂</td>
<td>87.7</td>
<td>62.3</td>
<td>79.4</td>
<td>73.0</td>
<td>85.2</td>
<td>76.5</td>
</tr>
<tr>
<td>ours (LoCo)</td>
<td>84.9</td>
<td>59.4</td>
<td><strong>90.0</strong></td>
<td><strong>84.8</strong></td>
<td>73.5</td>
<td>78.1</td>
</tr>
</tbody>
</table>
## Drum Detection Results

<table>
<thead>
<tr>
<th>Method</th>
<th>KD</th>
<th>SD</th>
<th>HH</th>
<th>PRE</th>
<th>REC</th>
<th>F₁</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>RANDOM</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RNN</td>
<td>94.7</td>
<td>79.5</td>
<td>88.3</td>
<td>84.1</td>
<td><strong>93.3</strong></td>
<td>87.5</td>
</tr>
<tr>
<td>TANHB</td>
<td>92.4</td>
<td>84.6</td>
<td>87.1</td>
<td>86.3</td>
<td>92.1</td>
<td>88.0</td>
</tr>
<tr>
<td>RELU5</td>
<td>91.3</td>
<td>83.8</td>
<td>85.2</td>
<td>83.7</td>
<td>92.3</td>
<td>86.8</td>
</tr>
<tr>
<td>LSTMPB</td>
<td>94.4</td>
<td>84.1</td>
<td>91.4</td>
<td>90.8</td>
<td>90.8</td>
<td><strong>90.0</strong></td>
</tr>
<tr>
<td>GRU5</td>
<td>94.2</td>
<td><strong>87.1</strong></td>
<td>87.7</td>
<td>88.6</td>
<td>92.7</td>
<td>89.7</td>
</tr>
<tr>
<td>ours (LoCo)</td>
<td>92.3</td>
<td>81.2</td>
<td><strong>93.0</strong></td>
<td><strong>90.9</strong></td>
<td>87.1</td>
<td>88.9</td>
</tr>
<tr>
<td><strong>SUBSET</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RNN</td>
<td>91.0</td>
<td>57.8</td>
<td>82.2</td>
<td>72.8</td>
<td><strong>88.3</strong></td>
<td>77.0</td>
</tr>
<tr>
<td>TANHB</td>
<td>82.7</td>
<td>61.6</td>
<td>84.8</td>
<td>74.1</td>
<td>83.8</td>
<td>76.4</td>
</tr>
<tr>
<td>RELU5</td>
<td>79.4</td>
<td>62.1</td>
<td>80.8</td>
<td>69.6</td>
<td>84.2</td>
<td>74.1</td>
</tr>
<tr>
<td>LSTMPB</td>
<td>85.8</td>
<td><strong>68.8</strong></td>
<td>83.7</td>
<td>78.3</td>
<td>84.7</td>
<td><strong>79.4</strong></td>
</tr>
<tr>
<td>GRU5</td>
<td>87.7</td>
<td>62.3</td>
<td>79.4</td>
<td>73.0</td>
<td>85.2</td>
<td>76.5</td>
</tr>
<tr>
<td>ours (LoCo)</td>
<td>84.9</td>
<td>59.4</td>
<td><strong>90.0</strong></td>
<td><strong>84.8</strong></td>
<td>73.5</td>
<td>78.1</td>
</tr>
</tbody>
</table>

State-of-the-art
## DRUM Detection

### Results

<table>
<thead>
<tr>
<th>Method</th>
<th>KD</th>
<th>SD</th>
<th>HH</th>
<th>PRE</th>
<th>REC</th>
<th>F₁</th>
</tr>
</thead>
<tbody>
<tr>
<td>RANDOM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RNN</td>
<td>94.7</td>
<td>79.5</td>
<td>88.3</td>
<td>84.1</td>
<td>93.3</td>
<td>87.5</td>
</tr>
<tr>
<td>TANHB</td>
<td>92.4</td>
<td>84.6</td>
<td>87.1</td>
<td>86.3</td>
<td>92.1</td>
<td>88.0</td>
</tr>
<tr>
<td>RELUts</td>
<td>91.3</td>
<td>83.8</td>
<td>85.2</td>
<td>83.7</td>
<td>92.3</td>
<td>86.8</td>
</tr>
<tr>
<td>LSTMkB</td>
<td>94.4</td>
<td>84.1</td>
<td>91.4</td>
<td>90.8</td>
<td>90.8</td>
<td>90.0</td>
</tr>
<tr>
<td>GRUts</td>
<td>94.2</td>
<td>87.1</td>
<td>87.7</td>
<td>88.6</td>
<td>92.7</td>
<td>89.7</td>
</tr>
<tr>
<td>ours (LoCo)</td>
<td>92.3</td>
<td>81.2</td>
<td>93.0</td>
<td>90.9</td>
<td>87.1</td>
<td>88.9</td>
</tr>
<tr>
<td>SUBSET</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RNN</td>
<td>91.0</td>
<td>57.8</td>
<td>82.2</td>
<td>72.8</td>
<td>88.3</td>
<td>77.0</td>
</tr>
<tr>
<td>TANHB</td>
<td>82.7</td>
<td>61.6</td>
<td>84.8</td>
<td>74.1</td>
<td>83.8</td>
<td>76.4</td>
</tr>
<tr>
<td>RELUts</td>
<td>79.4</td>
<td>62.1</td>
<td>80.8</td>
<td>69.6</td>
<td>84.2</td>
<td>74.1</td>
</tr>
<tr>
<td>LSTMkB</td>
<td>85.8</td>
<td>68.8</td>
<td>83.7</td>
<td>78.3</td>
<td>84.7</td>
<td>79.4</td>
</tr>
<tr>
<td>GRUts</td>
<td>87.7</td>
<td>62.3</td>
<td>79.4</td>
<td>73.0</td>
<td>85.2</td>
<td>76.5</td>
</tr>
<tr>
<td>ours (LoCo)</td>
<td>84.9</td>
<td>59.4</td>
<td>90.0</td>
<td>84.8</td>
<td>73.5</td>
<td>78.1</td>
</tr>
</tbody>
</table>

Great Overall F1-Score

State-of-the-art
Detection of three different drum types in drum audio extracts

Further tests on HH reveal that:
Further tests on HH reveal that:

- In that setting, the standard deviation is only of $4.35\text{ms}$ for the distance between true and predicted hits.
PIANO ONSET DETECTION

Results

Detection of piano notes in audio extracts 🎹
Detection of **piano** notes in audio extracts

- Complex task with 88 channels
Detection of piano notes in audio extracts

- Complex task with 88 channels
- Tight tolerance of 50ms for a prediction to be considered correct
Detection of piano notes in audio extracts

- Complex task with 88 channels
- Tight tolerance of 50ms for a prediction to be considered correct
- Comparison with fully-supervised benchmark models
## PIANO ONSET DETECTION

### Results

<table>
<thead>
<tr>
<th>Method</th>
<th>PRE</th>
<th>REC</th>
<th>F₁</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sigtilia et al. (2016)</td>
<td>44.97</td>
<td>49.55</td>
<td>46.58</td>
</tr>
<tr>
<td>Kelz et al. (2016)</td>
<td>44.27</td>
<td>61.29</td>
<td>50.94</td>
</tr>
<tr>
<td>Hawthorne et al. (2017)</td>
<td>84.24</td>
<td>80.67</td>
<td>82.29</td>
</tr>
<tr>
<td>ours (LoCo)</td>
<td>76.22</td>
<td>68.61</td>
<td>71.99</td>
</tr>
</tbody>
</table>
PIANO ONSET DETECTION

Results
Digit Detection Experiment
DIGIT DETECTION EXPERIMENT

Main Idea

Initial Image

Not a sequence
DIGIT DETECTION EXPERIMENT

Main Idea

Not a sequence
DIGIT DETECTION EXPERIMENT

Main Idea

Not a sequence

A sequence
DIGIT DETECTION EXPERIMENT

Main Idea

Input
DIGIT DETECTION EXPERIMENT

Main Idea

Input → Model
**DIGIT DETECTION EXPERIMENT**

**Main Idea**

- **Input**: Images of digits.
- **Model**: Processing the images.
- **Predictions**: Results of the model, indicating presence or absence of digits.

The diagram illustrates the process of weakly-supervised temporal localization via occurrence count learning, showcasing how the model processes input images to predict digit occurrences.
DIGIT DETECTION EXPERIMENT

Main Idea

Predictions
DIGIT DETECTION EXPERIMENT

Main Idea

Predictions

Hilbert Space-filling curve
**DIGIT DETECTION EXPERIMENT**

**Main Idea**

**Predictions**

```
[0] [1] [0] [0] [0] [0]
[0] [0] [0] [0] [1] [0]
```

**Hilbert Space-filling curve**

**Object Detection**
Figure 4. Digit Representations. Comparison of t-SNE digit feature representations resulting from the fully-supervised VGG-13 architecture (left) and from our weakly-supervised approach (right).
Weakly-Supervised Temporal Localization via Occurrence Count Learning

Julien Schroeter | Kirill Sidorov | David Marshall

**DIGIT DETECTION EXPERIMENT**

Detection Performance

Mean absolute distance between true and estimated bounding box centers: 9:04 pixels (approx. step size of the space filling curve)
The model learnt:
The model learnt:

1. Feature representation
The model learnt:

1. Feature representation
2. Space-mapping
The model learnt:

1. Feature representation
2. Space-mapping
3. Object detection
The model learnt:

1. Feature representation
2. Space-mapping
3. Object detection

Using only occurrence counts as training labels
**CONCLUSION**

This work shows that implicit model constraints can be used to ensure that **accurate localization emerges as a byproduct of learning to count occurrences.**
CONCLUSION

This work shows that implicit model constraints can be used to ensure that accurate localization emerges as a byproduct of learning to count occurrences.

Competitive results against fully-supervised state-of-the-art models.
Questions?