Improving weakly supervised sound event detection with self-supervised auxiliary tasks

Soham Deshmukh\(^1\)  
Bhiksha Raj\(^2\)  
Rita Singh\(^2\)

\(^1\) Microsoft  
\(^2\) Carnegie Mellon University  
The work was done at Carnegie Mellon University
Sound event detection

What kind of sounds do you imagine listening in each scene?
Training SED models

Strong supervision: Audio events and their start and end time

Weak supervision: Audio tags
Sound event detection in present

Has progressed in past years due to larger datasets

However, sound event detection rarely explored in “in the wild” and noisy settings

Noise in pipeline

SED used for predictive maintenance
Sound event detection in present

Has progressed in past years due to larger datasets

However, sound event detection rarely explored in “in the wild” and noisy settings

Noise in pipeline

Inference in real-life noisy environments
Sound event detection in present

Has progressed in past years due to larger datasets

However, sound event detection rarely explored in “in the wild” and noisy settings

Noise in pipeline

Inference in real-life noisy environments

New applications have limited data

SED used for unobstructive healthcare
How to improve SED in noisy settings?

Learning better representations/feature detectors for each audio event from such noisy training data

- Self-supervised auxiliary tasks

Improving pooling method used in these networks

- Two step attention pooling
Proposed architecture
Proposed architecture

(A) Self-supervised learning architecture

\[ g_1 : \hat{X} \mapsto Z \quad g_2 : Z \mapsto P \]
\[ g_1(.) = g_3(.) = g(.) \]
\[ g_4^{-1}(g(\cdot)) = g^{-1}(g_4(\cdot)) = I \]

\[ \min_W \mathcal{L}_1(P, y|w, w_2) + \alpha \mathcal{L}_2(\{\hat{x}_i\}_{i=1}^T, \{\hat{x}_i\}_{i=1}^T|w, w_2) \]

B) Primary decoder

\[ Z_{a_1} = \frac{e^{\sigma(ZW_{a_1}^T + b_{a_1})}}{\sum_{i=1}^F e^{\sigma(ZW_{a_1}^T + b_{a_1})}} \quad Z_{c_1} = (ZW_{c_1}^T + b_{c_1}) \]
\[ Z_{p_1} = \sum_{i=0}^F Z_{c_1} \cdot Z_{a_1} \]
\[ Z_{a_2} = \frac{e^{\sigma(Z_{p_1} W_{a_2}^T + b_{a_2})}}{\sum_{t=1}^T e^{\sigma(Z_{p_1} W_{a_2}^T + b_{a_2})}} \quad Z_{p_1} = (ZW_{c_2}^T + b_{c_2}) \]
\[ Z_{p_2} = \sum_{t=0}^T Z_{c_2} \cdot Z_{a_2} \]
Experiments

We form a noisy dataset by mixing:
- DCASE 2019 Task 1 of Acoustic Scene Classification (ASC)
- DCASE 2018 Task 2 of General purpose Audio tagging

The DCASE 2019 Task 1 provides background sounds (noise) recorded from a variety of real world scenes in which the sounds from DCASE 2019 Task 2 are randomly embedded.
Experiments

We form a noisy dataset by mixing:
- DCASE 2019 Task 1 of Acoustic Scene Classification (ASC)
- DCASE 2018 Task 2 of General purpose Audio tagging

The DCASE 2019 Task 1 provides background sounds (noise) recorded from a variety of real world scenes in which the sounds from DCASE 2019 Task 2 are randomly embedded.

Results in 32000 audio clips with 8000 audio clips for each 20,10,0 dB SNR.
Results

Performance across different SNR

<table>
<thead>
<tr>
<th>Network</th>
<th>encoder</th>
<th>pooling</th>
<th>aux.</th>
<th>SNR 20 dB</th>
<th>micro-p</th>
<th>macro-p</th>
<th>AUC</th>
<th>SNR 10 dB</th>
<th>micro-p</th>
<th>macro-p</th>
<th>AUC</th>
<th>SNR 0 dB</th>
<th>micro-P</th>
<th>macro-p</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>VGGish</td>
<td>GAP</td>
<td>X</td>
<td></td>
<td>0.5067</td>
<td>0.6127</td>
<td></td>
<td>0.9338</td>
<td>0.4291</td>
<td>0.5390</td>
<td></td>
<td>0.9144</td>
<td>0.3295</td>
<td>0.4093</td>
<td></td>
<td>0.8694</td>
</tr>
<tr>
<td>VGGish</td>
<td>GMP</td>
<td>X</td>
<td></td>
<td>0.5390</td>
<td>0.5186</td>
<td></td>
<td>0.8497</td>
<td>0.5263</td>
<td>0.5023</td>
<td></td>
<td>0.8422</td>
<td>0.4640</td>
<td>0.4441</td>
<td></td>
<td>0.8189</td>
</tr>
<tr>
<td>VGGish</td>
<td>GWRP</td>
<td>X</td>
<td></td>
<td>0.7018</td>
<td>0.7522</td>
<td></td>
<td>0.9362</td>
<td>0.6538</td>
<td>0.7129</td>
<td></td>
<td>0.9265</td>
<td>0.5285</td>
<td>0.6084</td>
<td></td>
<td>0.8985</td>
</tr>
<tr>
<td>VGGish (dil.)</td>
<td>AP</td>
<td>X</td>
<td></td>
<td>0.7391</td>
<td>0.7586</td>
<td></td>
<td>0.9279</td>
<td>0.6740</td>
<td>0.7404</td>
<td></td>
<td>0.9211</td>
<td>0.5714</td>
<td>0.6341</td>
<td></td>
<td>0.9014</td>
</tr>
<tr>
<td>VGGish</td>
<td>2AP</td>
<td>✓</td>
<td></td>
<td>0.7829</td>
<td>0.7645</td>
<td></td>
<td>0.9390</td>
<td>0.7603</td>
<td>0.7486</td>
<td></td>
<td>0.9343</td>
<td>0.6986</td>
<td>0.6892</td>
<td></td>
<td>0.9177</td>
</tr>
</tbody>
</table>

The proposed architecture beats existing benchmark by
- SNR 20 dB: 5.9%,
- SNR 10 dB: 12.8%
- SNR 0 dB: 22.3%
Results

Ablation study of components

\[
\min_W \mathcal{L}_1(P, y|w, w_4) + \alpha \mathcal{L}_2(\{\bar{x}_i\}^T_{i=1}, \{\hat{x}_i\}^T_{i=1}|w, w_2)
\]

<table>
<thead>
<tr>
<th>auxiliary task</th>
<th>SNR 20 dB</th>
<th>SNR 10 dB</th>
<th>SNR 0 dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\alpha = 0.0)</td>
<td>0.7772</td>
<td>0.7430</td>
<td>0.6937</td>
</tr>
<tr>
<td>(\alpha = 0.001)</td>
<td><strong>0.7829</strong></td>
<td><strong>0.7603</strong></td>
<td><strong>0.6986</strong></td>
</tr>
<tr>
<td>(\alpha = 0.1)</td>
<td>0.7637</td>
<td>0.7428</td>
<td>0.6792</td>
</tr>
</tbody>
</table>

Varying alpha:
- \(\alpha = 0 \rightarrow\) two step attention pooling: 5.2%, 10.2%, 21.4% on 20, 10, 0 dB SNR
- \(\alpha = 1e-3 \rightarrow\) two step attention pooling and aux task: 0.7%, 2.3%, 0.7 % on 20, 10, 0 dB SNR
- \(\alpha = 1e-2 \rightarrow\) two step attention pooling: decreased
## Results

Performance on different type of sound event

### WEAKLY LABELLED SED AUDIO EVENT SPECIFIC RESULTS FOR SNR = 0

<table>
<thead>
<tr>
<th>Model</th>
<th>Guitar</th>
<th>Appl add</th>
<th>Ba rk</th>
<th>Bass drum</th>
<th>Bur ping</th>
<th>Bus</th>
<th>Cel lo</th>
<th>Chi me</th>
<th>Clar incr</th>
<th>Comp. keyb.</th>
<th>Cou gh</th>
<th>Cou bell</th>
<th>Double bass</th>
<th>Dra wer</th>
<th>Elec. piano</th>
<th>Fa rt</th>
<th>Finger snap.</th>
<th>Fire work</th>
<th>Flu te</th>
<th>Glock ensp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>GAP</td>
<td>0.548</td>
<td>0.818</td>
<td>0.477</td>
<td>0.191</td>
<td>0.508</td>
<td>0.168</td>
<td>0.361</td>
<td>0.626</td>
<td>0.389</td>
<td>0.502</td>
<td>0.334</td>
<td>0.197</td>
<td>0.212</td>
<td>0.212</td>
<td>0.386</td>
<td>0.409</td>
<td>0.36</td>
<td>0.286</td>
<td>0.359</td>
<td></td>
</tr>
<tr>
<td>GMP</td>
<td>0.517</td>
<td>0.539</td>
<td>0.53</td>
<td>0.535</td>
<td>0.426</td>
<td>0.145</td>
<td>0.378</td>
<td>0.406</td>
<td>0.356</td>
<td>0.208</td>
<td>0.275</td>
<td>0.077</td>
<td>0.31</td>
<td>0.393</td>
<td>0.623</td>
<td>0.322</td>
<td>0.384</td>
<td>0.889</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GWRR</td>
<td>0.728</td>
<td>0.933</td>
<td>0.742</td>
<td>0.242</td>
<td>0.741</td>
<td>0.254</td>
<td>0.511</td>
<td>0.766</td>
<td>0.449</td>
<td>0.587</td>
<td>0.629</td>
<td>0.262</td>
<td>0.296</td>
<td>0.349</td>
<td>0.652</td>
<td>0.514</td>
<td>0.517</td>
<td>0.418</td>
<td>0.893</td>
<td></td>
</tr>
<tr>
<td>AtrousAP</td>
<td>0.72</td>
<td>0.956</td>
<td>0.782</td>
<td>0.169</td>
<td>0.804</td>
<td>0.2</td>
<td>0.562</td>
<td>0.767</td>
<td>0.502</td>
<td>0.685</td>
<td>0.756</td>
<td>0.781</td>
<td>0.214</td>
<td>0.187</td>
<td>0.691</td>
<td>0.734</td>
<td>0.566</td>
<td>0.318</td>
<td>0.902</td>
<td></td>
</tr>
<tr>
<td>2APAE</td>
<td>0.869</td>
<td>0.942</td>
<td>0.865</td>
<td>0.62</td>
<td>0.849</td>
<td>0.572</td>
<td>0.71</td>
<td>0.633</td>
<td>0.542</td>
<td>0.59</td>
<td>0.628</td>
<td>0.921</td>
<td>0.386</td>
<td>0.552</td>
<td>0.569</td>
<td>0.907</td>
<td>0.579</td>
<td>0.97</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2APAE e-3</td>
<td>0.792</td>
<td>0.951</td>
<td>0.839</td>
<td>0.812</td>
<td>0.974</td>
<td>0.627</td>
<td>0.669</td>
<td>0.606</td>
<td>0.503</td>
<td>0.699</td>
<td>0.631</td>
<td>0.94</td>
<td>0.59</td>
<td>0.403</td>
<td>0.562</td>
<td>0.941</td>
<td>0.565</td>
<td>0.532</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2APAE e-2</td>
<td>0.759</td>
<td>0.943</td>
<td>0.787</td>
<td>0.789</td>
<td>0.891</td>
<td>0.605</td>
<td>0.677</td>
<td>0.485</td>
<td>0.68</td>
<td>0.632</td>
<td>0.816</td>
<td>0.563</td>
<td>0.377</td>
<td>0.522</td>
<td>0.867</td>
<td>0.61</td>
<td>0.522</td>
<td>0.853</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### WEAKLY LABELLED SED AUDIO EVENT SPECIFIC RESULTS FOR SNR = 10

<table>
<thead>
<tr>
<th>Model</th>
<th>Guitar</th>
<th>Appl add</th>
<th>Ba rk</th>
<th>Bass drum</th>
<th>Bur ping</th>
<th>Bus</th>
<th>Cel lo</th>
<th>Chi me</th>
<th>Clar incr</th>
<th>Comp. keyb.</th>
<th>Cou gh</th>
<th>Cou bell</th>
<th>Double bass</th>
<th>Dra wer</th>
<th>Elec. piano</th>
<th>Fa rt</th>
<th>Finger snap.</th>
<th>Fire work</th>
<th>Flu te</th>
<th>Glock ensp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>GAP</td>
<td>0.69</td>
<td>0.974</td>
<td>0.691</td>
<td>0.238</td>
<td>0.642</td>
<td>0.373</td>
<td>0.57</td>
<td>0.763</td>
<td>0.372</td>
<td>0.648</td>
<td>0.529</td>
<td>0.507</td>
<td>0.394</td>
<td>0.438</td>
<td>0.873</td>
<td>0.481</td>
<td>0.364</td>
<td>0.644</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GMP</td>
<td>0.004</td>
<td>0.691</td>
<td>0.626</td>
<td>0.732</td>
<td>0.732</td>
<td>0.63</td>
<td>0.163</td>
<td>0.494</td>
<td>0.508</td>
<td>0.391</td>
<td>0.399</td>
<td>0.284</td>
<td>0.421</td>
<td>0.083</td>
<td>0.414</td>
<td>0.267</td>
<td>0.076</td>
<td>0.812</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GWRR</td>
<td>0.777</td>
<td>0.960</td>
<td>0.808</td>
<td>0.452</td>
<td>0.452</td>
<td>0.608</td>
<td>0.685</td>
<td>0.608</td>
<td>0.907</td>
<td>0.608</td>
<td>0.766</td>
<td>0.829</td>
<td>0.517</td>
<td>0.597</td>
<td>0.668</td>
<td>0.567</td>
<td>0.621</td>
<td>0.471</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### WEAKLY LABELLED SED AUDIO EVENT SPECIFIC RESULTS FOR SNR = 20

<table>
<thead>
<tr>
<th>Model</th>
<th>Guitar</th>
<th>Appl add</th>
<th>Ba rk</th>
<th>Bass drum</th>
<th>Bur ping</th>
<th>Bus</th>
<th>Cel lo</th>
<th>Chi me</th>
<th>Clar incr</th>
<th>Comp. keyb.</th>
<th>Cou gh</th>
<th>Cou bell</th>
<th>Double bass</th>
<th>Dra wer</th>
<th>Elec. piano</th>
<th>Fa rt</th>
<th>Finger snap.</th>
<th>Fire work</th>
<th>Flu te</th>
<th>Glock ensp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>GAP</td>
<td>0.72</td>
<td>0.986</td>
<td>0.747</td>
<td>0.399</td>
<td>0.699</td>
<td>0.56</td>
<td>0.64</td>
<td>0.803</td>
<td>0.485</td>
<td>0.707</td>
<td>0.571</td>
<td>0.584</td>
<td>0.501</td>
<td>0.332</td>
<td>0.652</td>
<td>0.481</td>
<td>0.593</td>
<td>0.988</td>
<td>0.766</td>
<td></td>
</tr>
<tr>
<td>GMP</td>
<td>0.907</td>
<td>0.843</td>
<td>0.654</td>
<td>0.838</td>
<td>0.631</td>
<td>0.336</td>
<td>0.565</td>
<td>0.499</td>
<td>0.657</td>
<td>0.534</td>
<td>0.44</td>
<td>0.889</td>
<td>0.137</td>
<td>0.579</td>
<td>0.328</td>
<td>0.653</td>
<td>0.226</td>
<td>0.54</td>
<td>0.931</td>
<td></td>
</tr>
<tr>
<td>GWRR</td>
<td>0.83</td>
<td>0.986</td>
<td>0.922</td>
<td>0.529</td>
<td>0.869</td>
<td>0.649</td>
<td>0.727</td>
<td>0.813</td>
<td>0.67</td>
<td>0.728</td>
<td>0.742</td>
<td>0.696</td>
<td>0.626</td>
<td>0.627</td>
<td>0.672</td>
<td>0.636</td>
<td>0.729</td>
<td>0.897</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AtrousAP</td>
<td>0.877</td>
<td>0.991</td>
<td>0.922</td>
<td>0.562</td>
<td>0.924</td>
<td>0.622</td>
<td>0.773</td>
<td>0.819</td>
<td>0.746</td>
<td>0.77</td>
<td>0.89</td>
<td>0.716</td>
<td>0.573</td>
<td>0.708</td>
<td>0.806</td>
<td>0.746</td>
<td>0.755</td>
<td>0.745</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2APAE</td>
<td>0.903</td>
<td>0.969</td>
<td>0.911</td>
<td>0.936</td>
<td>0.959</td>
<td>0.761</td>
<td>0.787</td>
<td>0.642</td>
<td>0.666</td>
<td>0.736</td>
<td>0.605</td>
<td>0.936</td>
<td>0.825</td>
<td>0.592</td>
<td>0.665</td>
<td>0.589</td>
<td>0.956</td>
<td>0.834</td>
<td>0.913</td>
<td></td>
</tr>
</tbody>
</table>
Results

Performance on different type of sound event

<table>
<thead>
<tr>
<th>model</th>
<th>aux.</th>
<th>bus</th>
<th>cowbell</th>
<th>gong</th>
<th>meow</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atrous + AP</td>
<td>X</td>
<td>0.2</td>
<td>0.781</td>
<td><strong>0.692</strong></td>
<td>0.583</td>
</tr>
<tr>
<td>VGGish + 2AP</td>
<td>X</td>
<td>0.572</td>
<td>0.921</td>
<td>0.643</td>
<td>0.483</td>
</tr>
<tr>
<td>VGGish + 2AP</td>
<td>✓</td>
<td><strong>0.627</strong></td>
<td>0.94</td>
<td>0.663</td>
<td>0.532</td>
</tr>
</tbody>
</table>

Some key insights:

- Proposed model outperforms other models on almost all audio events across different SNR
- Most improvement observed on events like `Bass drum', `bus', `double bass', `cowbell'
- Atrous model outperforms proposed on `gong', `chime', `meow'. Indicates atrous models is better at detecting audio events whose energy is spread wide in the temporal domain
Results

Two step attention pooling visualisation

- Input audio mel spectrogram
- Aux. decoder output
- Attention weights-f1
- Attention weights-f2
- Attention weights-f3
- Output of 1st step attention pooling
- Attention weights-t
- Output of 2nd step attention pooling

Two step attention pooling visualisation
Conclusion

Two step attention pooling helps in learning features to better discriminate between sound events

- Both in clean and noisy settings
- Makes training stable
- Improves localisation of the audio event in T-F

Self-supervised auxiliary tasks can improve network performance in noisy settings

- Appropriate auxiliary task: reconstruction of input T-F representation
- Right contribution of auxiliary task
- Most benefit in SNR 10 dB
Thank you for listening

Soham Deshmukh\textsuperscript{1}

Bhiksha Raj\textsuperscript{2}

Rita Singh\textsuperscript{2}

\textsuperscript{1} Microsoft
\textsuperscript{2} Carnegie Mellon University

The work was done at Carnegie Mellon University
References


References


Sound event detection in wild

Applications with lot of background noise

Training data does not represent inference distribution

Current sound event detection models lose performance in noisy setting
Sound event detection in wild

Applications with lot of background noise

Training data does not represent inference distribution and event detection models lose performance in noisy setting