ASR Performance Prediction on Unseen Broadcast Programs using Convolutional Neural Networks
Zied Elloumi, Laurent Besacier, Olivier Galibert, Benjamin Lecouteux

To cite this version:
Zied Elloumi, Laurent Besacier, Olivier Galibert, Benjamin Lecouteux. ASR Performance Prediction on Unseen Broadcast Programs using Convolutional Neural Networks. Blackbox NLP Workshop and EMLP 2018, 2018, Bruxelles, Belgium. 2018. hal-02102831

HAL Id: hal-02102831
https://hal.archives-ouvertes.fr/hal-02102831
Submitted on 17 Apr 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Analyzing Learned Representations of a Deep ASR Performance Prediction Model
Zied Elloumi, Laurent Besacier, Olivier Gallibert, Benjamin Lecouteux

In short
- **Task**: prediction of ASR performance on unseen broadcast programs at utterance level
- **Goal**: understand which information is captured by our deep model (Elloumi et al., 2018) and its relation with different conditioning factors
- **Main results**: a clear signal is captured about speech style, accent, and broadcast type

Our ASR performance prediction system
- In (Elloumi et al., 2018), we proposed a new approach using convolution neural networks (CNNs) to predict ASR performance from a collection of heterogeneous broadcast programs (both radio and TV)
- We particularly focused on the combination of text (ASR transcription) and signal (raw speech) inputs which both proved useful for CNN prediction

The network input can be either a pure text input, a pure signal input (raw signal) or a dual (text+speech) input at utterance level
- Our best approach gave 19.24% in terms of MAE (Mean Absolute Error)

Methodology
- Generate utterance level features (colored in yellow) from our deep model
- Follow (Belinkov and Glass, 2017) approach to better understand which information is captured by our deep model and its relation with different conditioning factors: speech style, accent, and broadcast program origin
  - **Classification task**: build three shallow feed-forward neural network classifiers (SHOW, STYLE, ACCENT) with a similar architecture: one hidden layer of 128 units followed by dropout (rate of 0.5), a ReLU non-linearity and a softmax layer for mapping onto the label set size
  - **Visualization task**: t-SNE algorithm to plot hidden representations

Data
- Data set from (Elloumi et al., 2018) divided into 3 subsets: TRAIN (67.5K), DEV (7.5K) and TEST (6.7K) → The TEST set contains unseen broadcast programs that are different from those present in TRAIN and DEV

Extract a balanced version of our TRAIN/DEV/TEST sets by filtering among over-represented labels

<table>
<thead>
<tr>
<th>#Catg</th>
<th>Turns of speech per category</th>
<th>TRAIN</th>
<th>DEV</th>
<th>TEST</th>
</tr>
</thead>
<tbody>
<tr>
<td>SHOW</td>
<td>5</td>
<td>4487±3</td>
<td>493±3</td>
<td>-</td>
</tr>
<tr>
<td>STYLE</td>
<td>2</td>
<td>1327±2</td>
<td>1403±2</td>
<td>3109±2</td>
</tr>
<tr>
<td>ACCENT</td>
<td>2</td>
<td>2304±2</td>
<td>2559±2</td>
<td>1539±2</td>
</tr>
</tbody>
</table>

Results
- **Classification task**
  - | Layer | Dim. | SHOW | STYLE | ACCENT |
  - | TXT  |      |      |       |        |
  - A1  | 1280 | 57.12 | 80.72 | 68.99 | 70.75 | 66.54 |
  - A2  | 256  | 54.89 | 80.11 | 69.56 | 70.30 | 69.43 |
  - A3  | 128  | 51.04 | 79.23 | 68.27 | 68.25 | 70.89 |
  - RAW-SIG
  - B1  | 512  | 42.35 | 72.92 | 58.64 | 64.60 | 55.85 |
  - B2  | 512  | 41.22 | 72.20 | 58.41 | 64.44 | 54.84 |
  - B3  | 256  | 41.22 | 72.38 | 58.44 | 64.50 | 54.65 |
  - B4  | 128  | 40.77 | 72.38 | 58.52 | 64.74 | 54.87 |
  - TXT + RAW-SIG
  - C1  | (A3+B4)| 256 | 57.04 | 81.29 | 70.36 | 71.41 | 65.98 |
  - C2  | 128  | 53.06 | 79.62 | 70.55 | 70.01 | 65.20 |

Visualisation task
- Duration between 4 and 5s
- Duration between 5 and 6s

Multi-task learning
- We perform multi-task learning providing the additional information about broadcast type, speech style and speaker’s accent during training
- The architecture of the multi-task model is similar to the single-task WER prediction model but we add additional outputs: a Softmax function is added for each new classification task after the last fully connected layer (C2)

We propose an analysis of learned representations of our deep ASR performance prediction system
- Experiments show that hidden layers convey a clear signal about speech style, accent, and broadcast type
- We proposed a multi-task learning approach to simultaneously predict WER and classify utterances according to style, accent and broadcast program origin
- A slight improvements on the test set are observed for MAE and Kendall metrics using multi-task systems

Conclusion

Reference:

Contact: Zied.elloumi@lne.fr