Multitask Learning with Capsule Networks for Speech-to-Intent Applications

– Presented at ICASSP 2020

Jakob Poncelet and Hugo Van hamme

KU Leuven – Department Electrical Engineering ESAT/PSI, Belgium

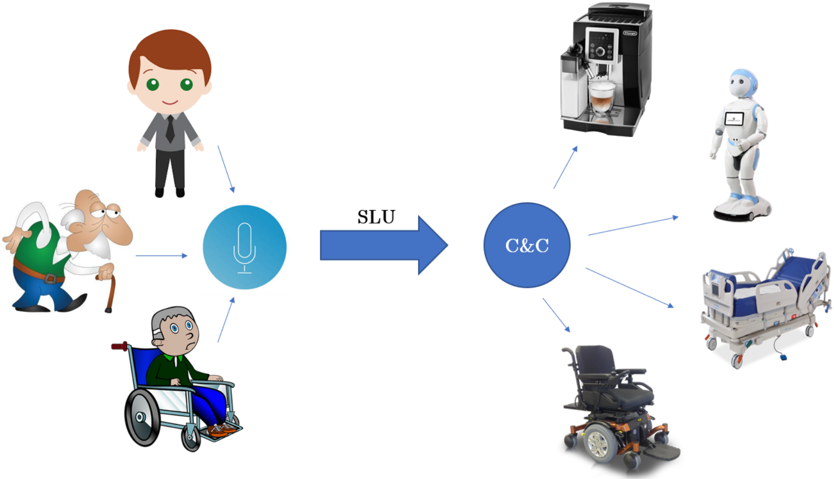
0 Overview

KU LEUVEN

1 Overview

1. Introduction
2. Model
3. Data
4. Results
5. Conclusion

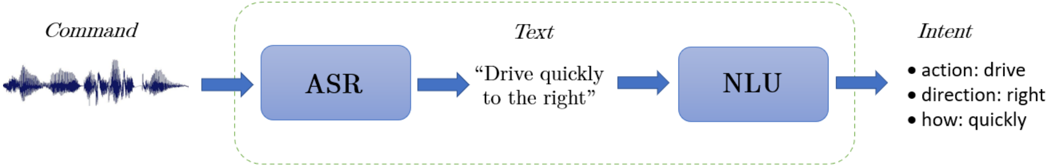
1 Introduction



Spoken Language Understanding system for Command-and-Control applications

1 Introduction

Conventional SLU

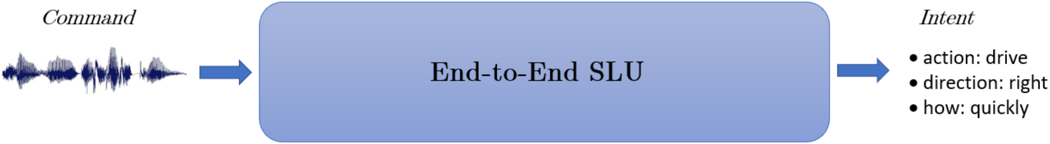


Problems:

- Dysarthric speech
- Strong dialects
- Domain specific

1 Introduction

Direct *speech-to-intent*



Idea: Build model from scratch with demonstrations from the user

→ All kinds of speech
→ Language and domain independent
→ User can choose his/her phrases

1 Introduction

E2ESLU Approaches:

- Non-Negative Matrix Factorization (NMF) [1]
- Encoder-Decoder neural networks [2]
- **Capsule networks** [3]

Key points:

→ Fast learning models (due to explicit user dependency)

→ High asymptotic accuracy

[1] B. Ons, J.F. Gemmeke, H. Van hamme, "Fast vocabulary acquisition in an nmf-based self-learning vocal user interface," Computer Speech and Language, vol. 28, no. 4, pp. 997 – 1017, 2014

[2] D. Serdyuk, Y. Wang, C. Fuegen, A. Kumar, B. Liu, Y. Bengio, "Towards end-to-end spoken language understanding," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 5754–5758

[3] V. Renkens, H. Van hamme, "Capsule networks for low resource spoken language understanding," Interspeech, Sep 2018

1 This work

- Analysis of the capsules in the proposed architecture for E2ESLU [1], more specifically how the different intents are represented

- Introducing multitask learning in the capsule network by applying task-specific regularisations to the output capsules
  $\rightarrow$ Speaker recognition (generalisable: extra designer possibilities!)

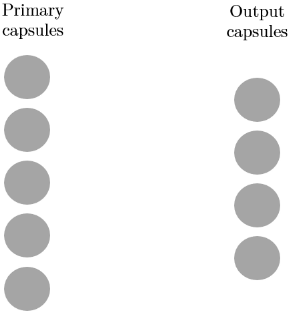
- Performance comparison between the baseline and the multitask model on small and large datasets, when used by multiple speakers

[1] V. Renkens, H. Van hamme, "Capsule networks for low resource spoken language understanding," Interspeech, Sep 2018

2 Overview

KU LEUVEN

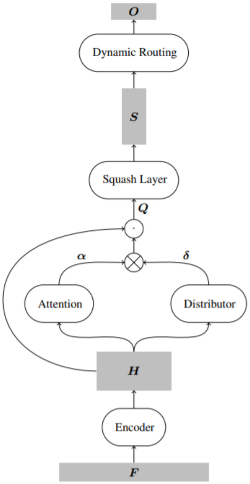
2 Capsule Networks: main idea



S. Sabour, N. Frosst, G. E Hinton, "Dynamic routing between capsules," in Proceedings NIPS, 2017.

- Capsule: activation vector
  - Length = probability object/pattern is present
  - Orientation = instantiation parameters of object/pattern
- Lower layer capsules predict output of next layer capsule
  - $\hat{\boldsymbol{u}}_{j|i} = \boldsymbol{W}_{ij}\boldsymbol{u}_i$
  - Dynamic routing: multiple lower layer capsules should agree on the higher level property
  - Parts $\rightarrow$ Whole

2 Baseline Capsule Network Model



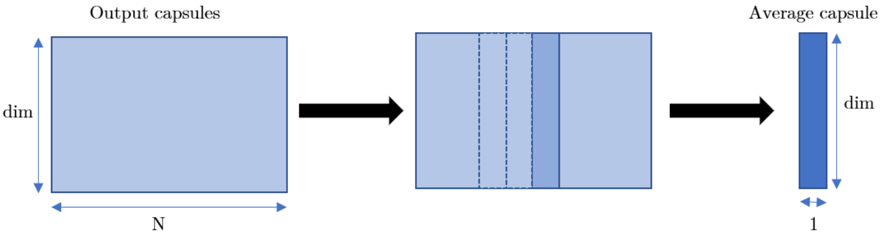
- Every primary capsule corresponds to a word/subword pattern in the input features
- Every output capsule corresponds to a specific label
  - Semantic frame is filled from all output capsules that are "active" (= length vector close to 1)
- Training: max-margin loss on output vectors $\boldsymbol{v}$
  - $L_l = \sum_{k=1}^{K} T_k \, max(0, 0.9 - \|\boldsymbol{v}_k\|) + (1 - T_k) \, max(0, \|\boldsymbol{v}_k\| - 0.1)$

Image: V. Renkens, H. Van hamme, "Capsule networks for low resource spoken language understanding," Interspeech, Sep 2018

2 Multitask Model – 1

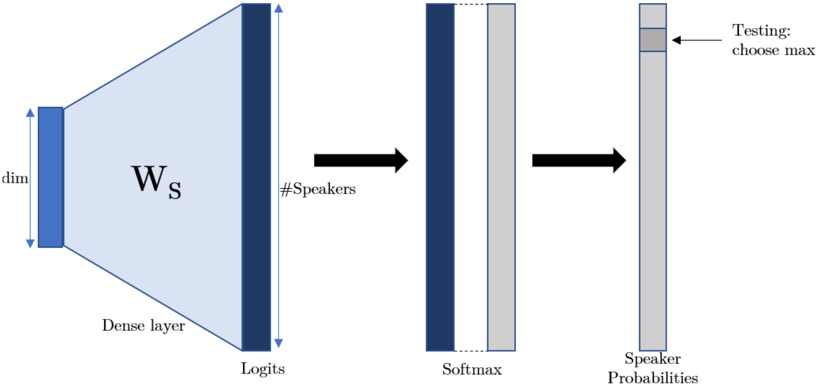
- Average capsule: combine information from every output capsule, for every output dimension separately

$$\boldsymbol{z} = \frac{\sum_{i=1}^{N} \boldsymbol{v}_i}{\sum_{i=1}^{N} \|\boldsymbol{v}_i\|}$$



2 Multitask Model – 2

- Speaker recognition: map the average capsule to speaker probabilities with a single softmax layer

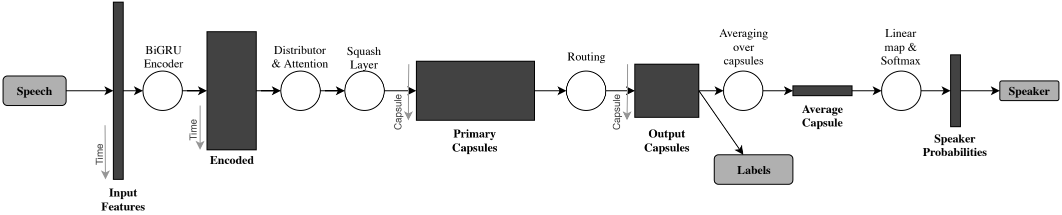


KU LEUVEN

2 Multitask Model – 3

- Define speaker loss with cross-entropy on predicted speaker probabilities
  $L_s = -\sum_{i=1}^{M} t_i log(P_i)$

- Add with regularisation parameter to total loss
  $L_{tot} = L_l + \lambda_s L_s$



3 Overview

3 Datasets

- `GRABO`
  - Commands to robot, e.g. “drive quickly to the right”, “drive slowly a little bit forward”, “pointer on”, “grab object”
  - 33 output labels
  - 10 Dutch, 1 English speaker
  - ca 6000 utterances in total

- `Fluent Speech Commands`
  - Smart home, virtual assistant, e.g. “turn on the lights in the bedroom”, “turn up the volume”, “I need to practice my German, change the language”
  - 31 output labels
  - 97 English speakers
  - ca 30000 utterances in total

3 Figures

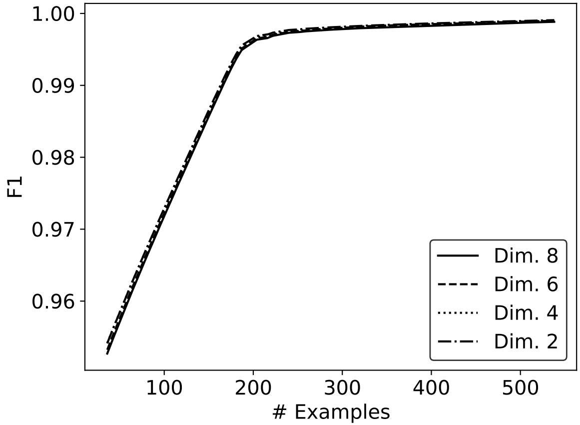
- Learning curves created by cross-validation experiments
  $\rightarrow$ performance in function of amount of training data
  - F1-measure for label classification
  - Percentage of correctly decoded speakers

- Distinction between
  - Speaker dependent experiments: perform experiment on data of one speaker only, average over results of all speakers
  - Speaker independent experiments: perform experiment on data of all speakers mixed together

4 Overview

KU LEUVEN

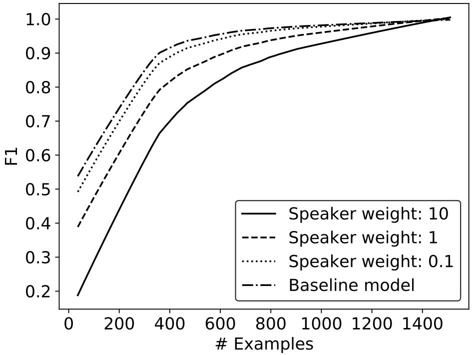
4 Dimension Analysis of Baseline Model



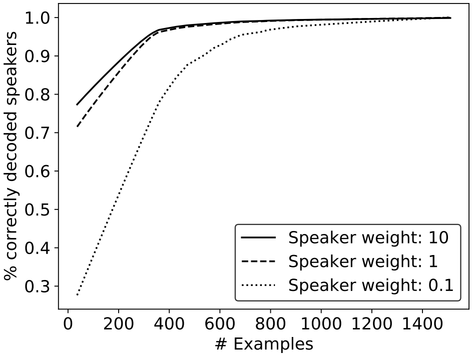
- Comparison for different output dimension (speaker dependent experiment on GRABO)
- Vector of dimension 2 suffices to represent the different intents

KU LEUVEN

4 Performance of Multitask Model on GRABO (speaker independent experiments)

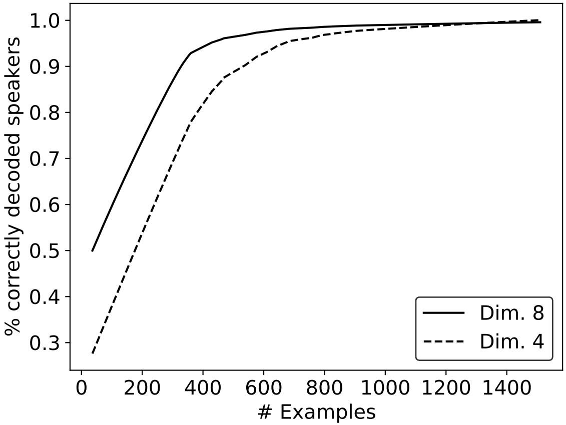


Intent Recognition



Speaker Recognition

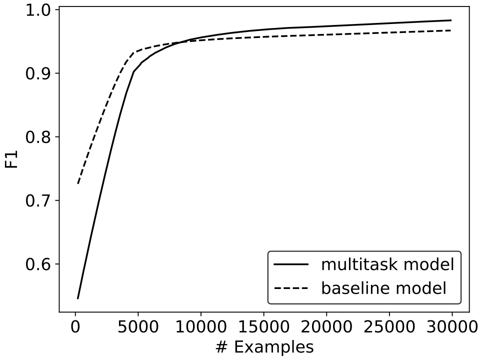
4 Dimension Analysis of Multitask Model



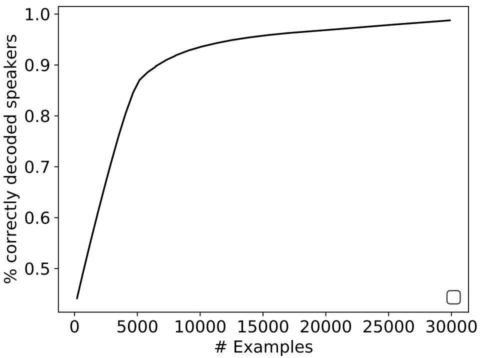
- The applied regularisation has given meaning to the orientation of the output vectors (= speaker identity)

KU LEUVEN

4 Performance of Multitask Model on Fluent Speech Commands (speaker independent experiments)



Intent Recognition



Speaker Recognition

KU LEUVEN

5 Overview

5 Conclusion

- The regularisation in the multitask model has given an interpretable meaning to the orientation of the activation vectors of the output capsules $\rightarrow$ Speaker-ID

- Multitask learning of speaker-ID improved the performance of the capsule network on the larger, challenging, Fluent Speech Commands dataset

THE END

Thank you for your attention.