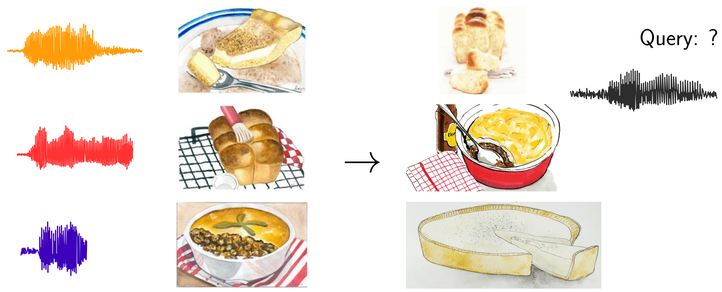
Multimodal One-Shot Learning of Speech and Images

Ryan Eloff Herman A. Engelbrecht Herman Kamper
E&E Engineering, Stellenbosch University, South Africa

Read me at arXiv.org

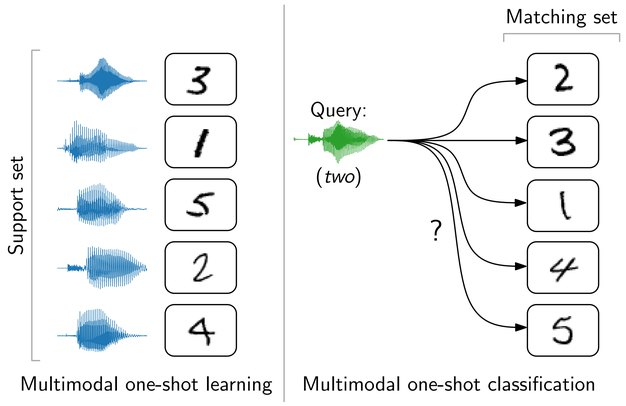
Motivation

Imagine that you are a robot chef (in a kitchen) ...



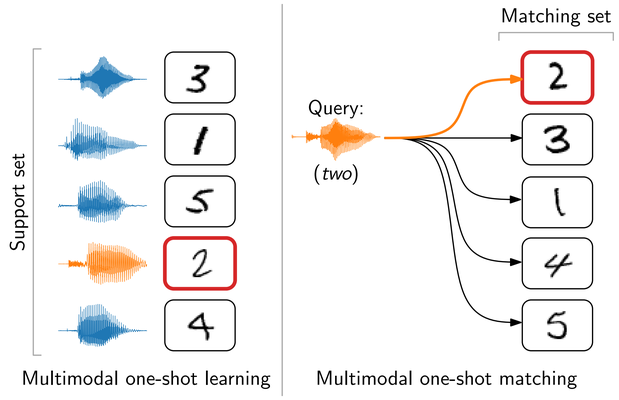
- Humans quickly learn new words and object categories from one or a few examples.
- Artificial agents should do the same, yet current speech and vision processing algorithms require thousands of labelled examples to complete a similar task.
- **One-shot learning**: acquisition of novel concepts from a single labelled example.
- Different to the above example, since you directly associate visual signals to spoken words without class labels, and generalise to new visual/spoken instances!
- **Multimodal one-shot learning**: a new task we formalise, where agents learn novel concepts from a single example of co-occurring multimodal sensory inputs.

Multimodal One-Shot Learning and Matching



- Multimodal one-shot learning on a dataset of spoken digits paired with images.
- At test time, a model must match a test query in one modality to the matching item in a test (or *matching*) set in the other modality.
- This is done using information from the *support set*, where neither the query nor the matching set instances occur in the support set.

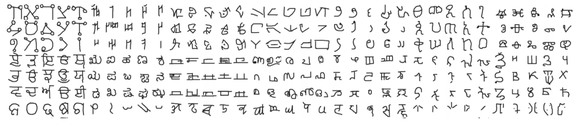
Our approach



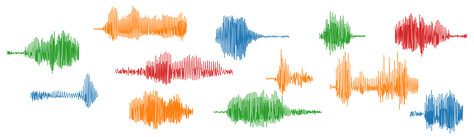
- Cross-modal test-time matching via unimodal comparisons with the support set.
- Assumes we can measure within modality similarity → unimodal one-shot learning!

Metric learning from background data

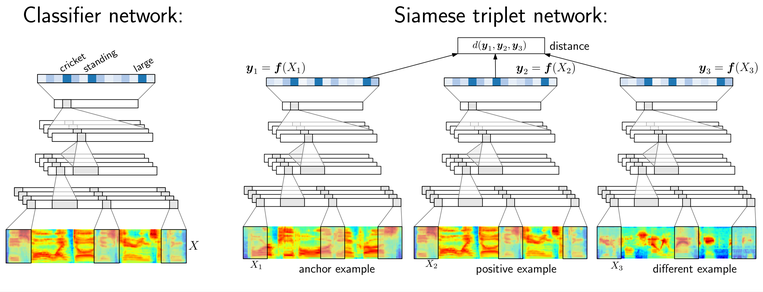
Omniglot labelled characters (no digits):



Isolated labelled words (no digits):



Neural network models for metric learning



Experimental details

- Simple benchmark dataset: one-shot learning from spoken digits paired with handwritten digit images.
  Speech: TIDigits corpus of spoken digit sequences split into isolated digits.
  Images: MNIST handwritten digits dataset.
- Treat utterances labelled "oh" and "zero" as separate classes → 11 class labels.
- Models evaluated on one-shot task accuracy averaged over 400 test episodes.

One-shot speech classification

11-way one-shot and five-shot speech classification results on isolated spoken digits.

Model	Train time	11-way Accuracy one-shot	11-way Accuracy five-shot
DTW	–	67.99% ± 0.29	91.30% ± 0.20
FFNN Classifier	13.1m	71.39% ± 0.81	89.49% ± 0.45
CNN Classifier	60.6m	82.07% ± 0.92	93.58% ± 0.98
Siamese CNN (offline)	70.5m	89.40% ± 0.54	95.12% ± 0.37
Siamese CNN (online)	15.0m	**92.85% ± 0.38**	**97.65% ± 0.22**

One-shot matching of speech to images

11-way one- and five-shot cross-modal matching of spoken and visual digits. Speaker invariance tests are 11-way one-shot, where all support set items are from the same speaker as the query, except for the item actually matching the query.

Model	11-way Accuracy one-shot	five-shot	speaker invariance
DTW + Pixels	34.92% ± 0.42	44.46% ± 0.69	28.00% ± 1.86
FFNN Classifier	36.49% ± 0.41	44.29% ± 0.56	34.95% ± 2.28
CNN Classifier	56.47% ± 0.76	63.97% ± 0.91	53.71% ± 2.2
Siamese CNN (offline)	67.41% ± 0.56	70.92% ± 0.36	66.70% ± 0.92
Siamese CNN (online)	**70.12% ± 0.68**	**73.53% ± 0.52**	**69.73% ± 1.04**

Conclusions

- Introduced and formalised multimodal one-shot learning, specifically for learning from speech and images.
- Developed a one-shot cross-modal matching dataset that may be used to benchmark other approaches.
- Unimodal one-shot learning approaches may be used for this task, but result in compounding errors through successive unimodal comparisons.
- **Future:** explore methods that can directly match one modality to another, particularly looking into recent meta-learning approaches.
- Full code recipe available at:
  `https://github.com/rpeloff/multimodal_one_shot_learning`

IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 2019

ryan.peter.eloff@gmail.com | ryaneloff.com | rpeloff