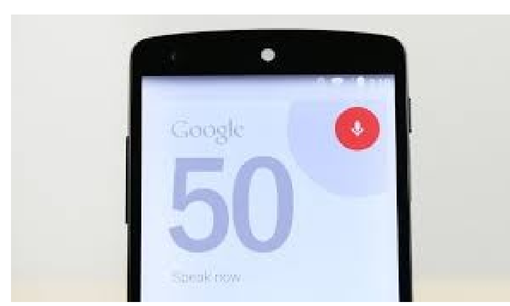
Symbolic inductive bias for visually grounded learning of spoken language

Grzegorz Chrupała

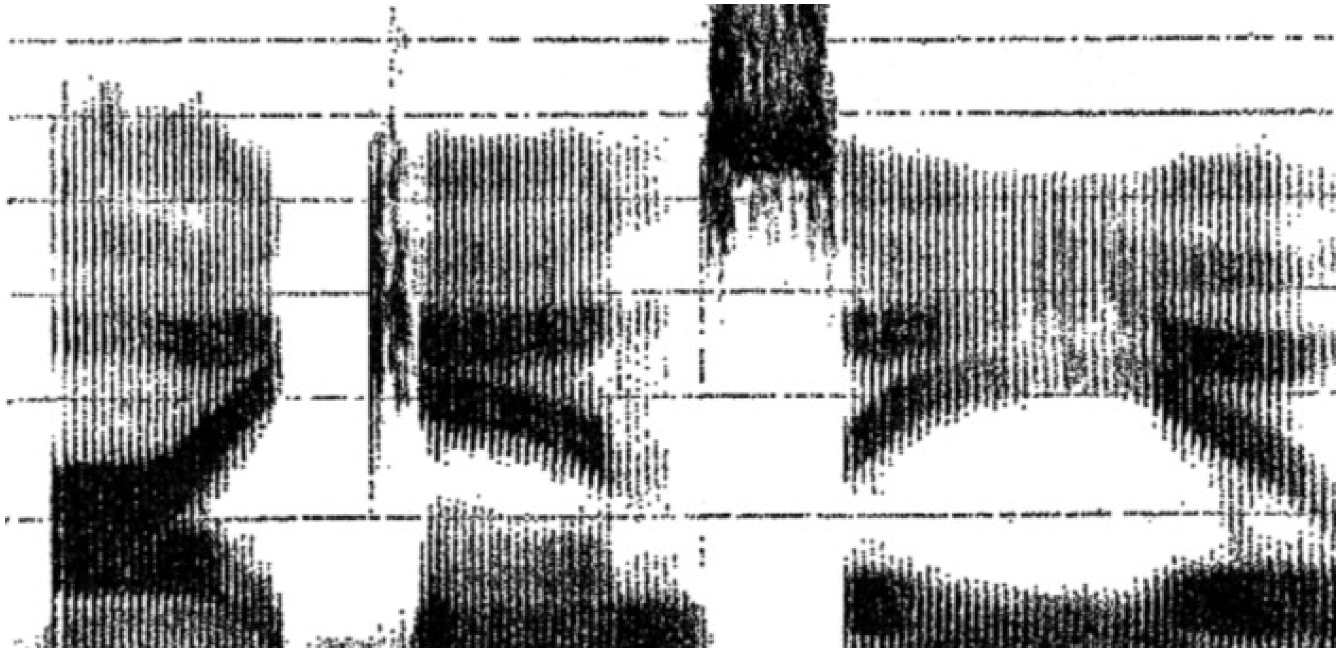
Tilburg University

Automatic Speech Recognition

A major commercial success story in Language Technology

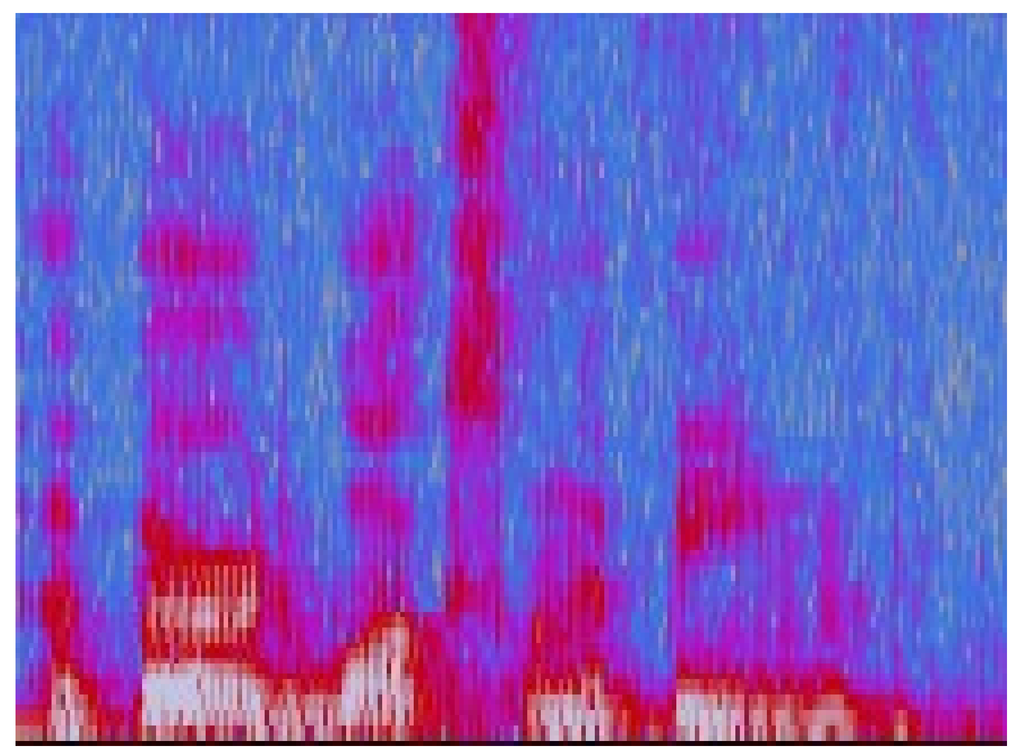


Very strong supervision



I can see you

Weaker supervision: Visually grounded spoken language



Data

- Flickr8K Audio Caption Corpus

  (Harwath and Glass 2016)
  - Written captions read by crowd workers
  - 8K images, five audio captions each

Existing models

- Convolutional neural network applied to a spectrogram
  - Harwath and Glass 2016 (NIPS)
- Multi-layer Highway recurrent network applied to Mel-frequency Cepstral Coefficient features
  - Chrupała et al 2017 (ACL)

Learning language via visual grounding

- Closer to human language learning
- May be easier to obtain data
  - Low-resource languages
  - Languages with no standard writing system
    - Cantonese, Hokkien
- **BUT: difficult, less constrained task**

Inductive bias

The **inductive bias** of a learning algorithm is the **set of assumptions** that the learner uses to predict outputs given inputs that it has **not encountered**.

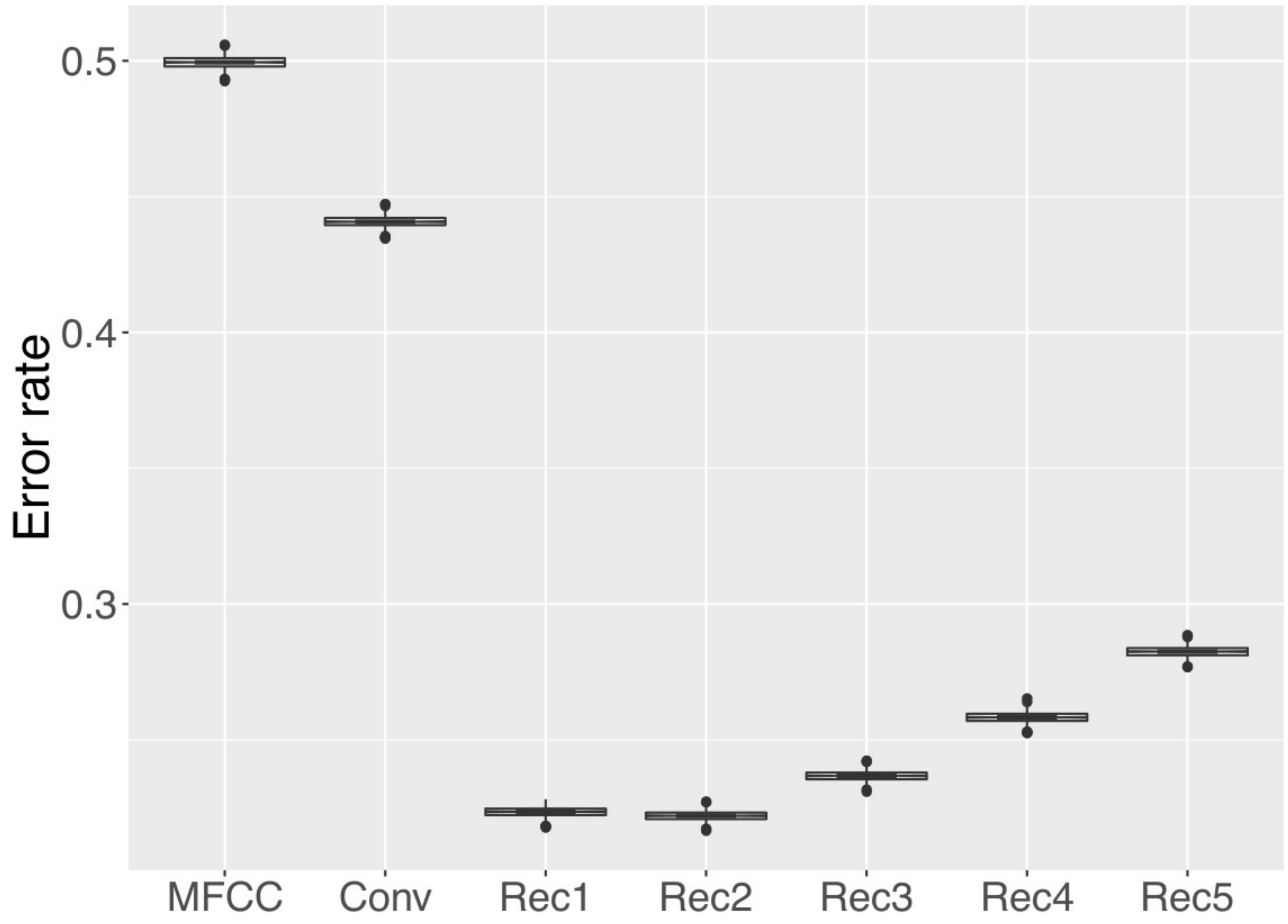
https://en.wikipedia.org/wiki/Inductive_bias

(Recurrent) Neural Networks

- RNN: autoregressive neural nets.
- Do **not** assume any **linguistically-motivated** structure.
- They may **discover** the existence of discrete phonemes in speech, **despite** this lack of bias.

Learned representations encode phonemes

Alishahi et al 2017 (CoNLL)

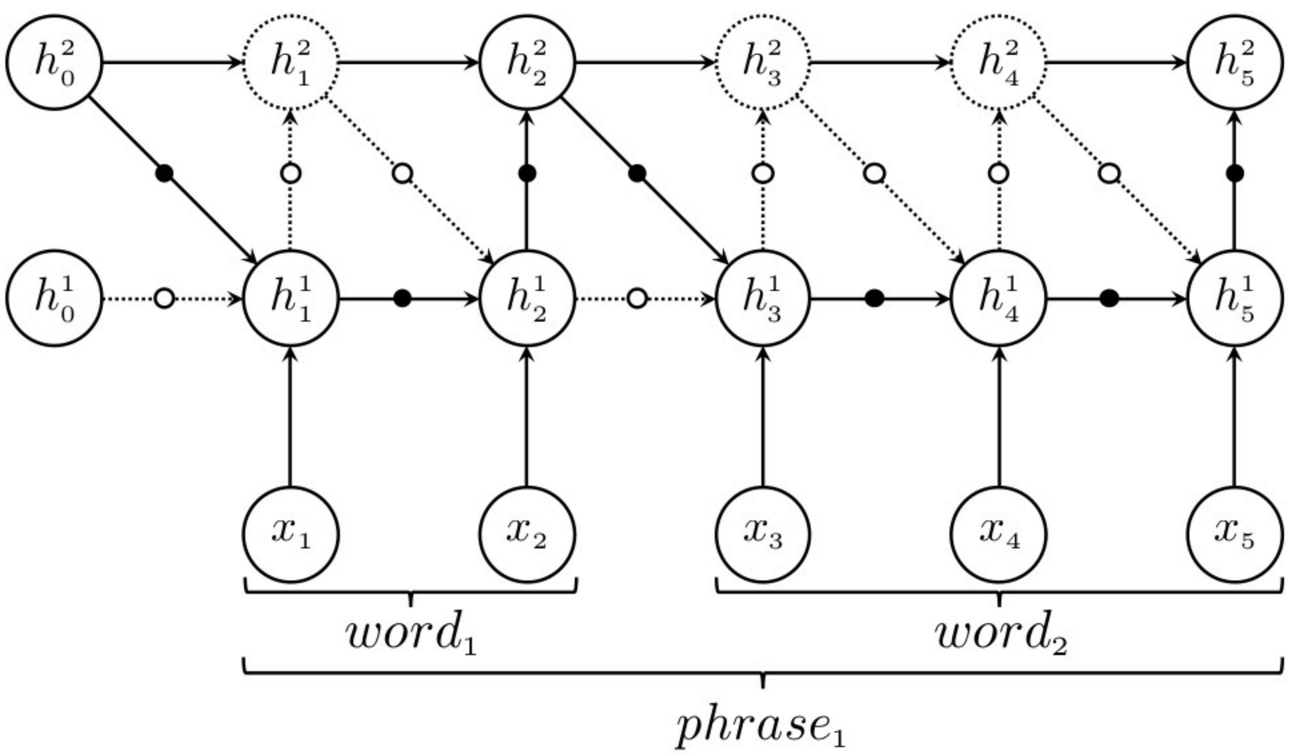


Inject inductive bias via Multi-task learning

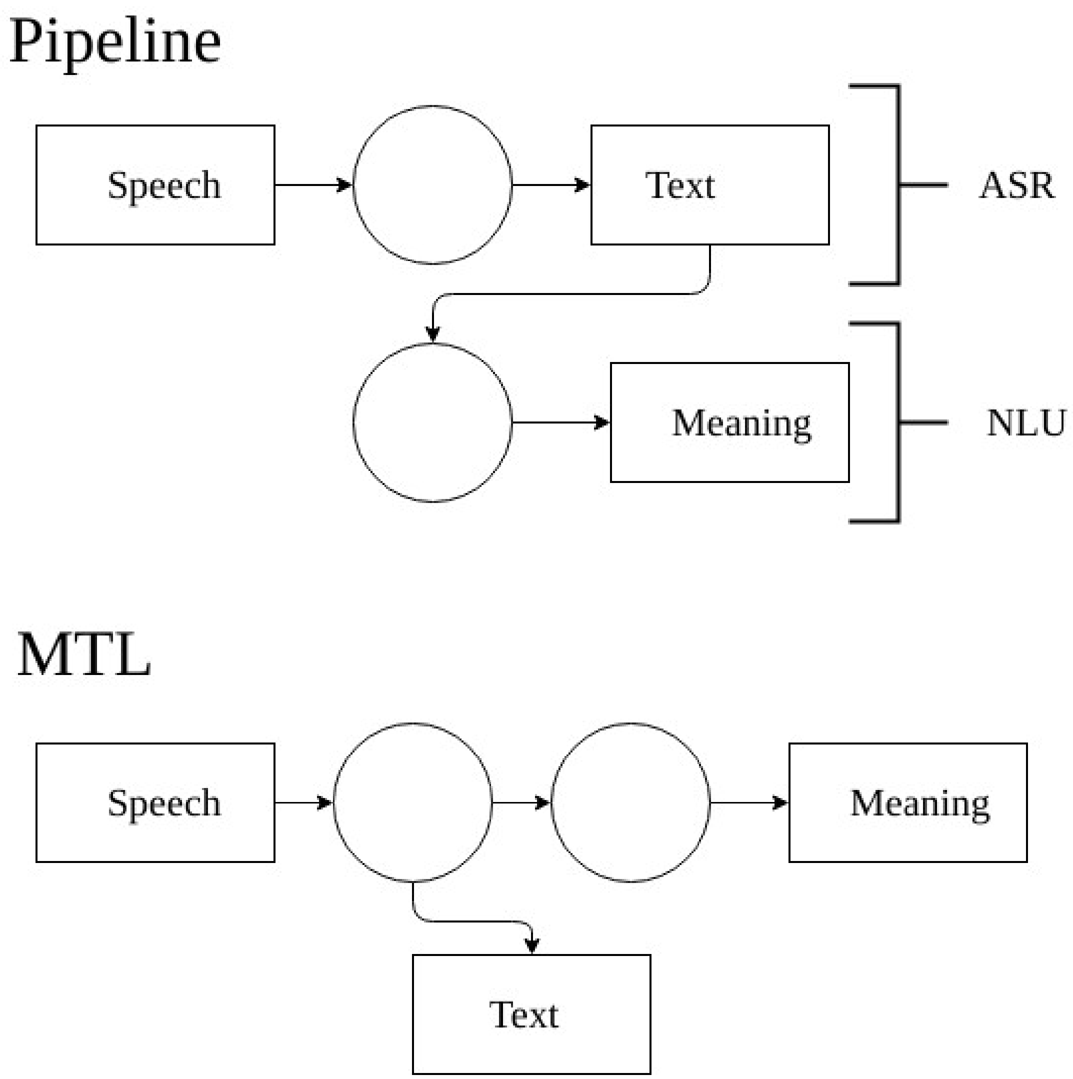
- Human learners – biases encoded in the genome via evolution
- ML – biases encoded via
  - Architectural design
  - Multi-task learning

Inductive bias via architecture

- Bias encoded as hard constraint on architecture.
- Example: Chung et al 2017 (ICLR)
- **But hard to get to work**
  - Kádár et al 2018 (COLING)



Pipeline vs MTL



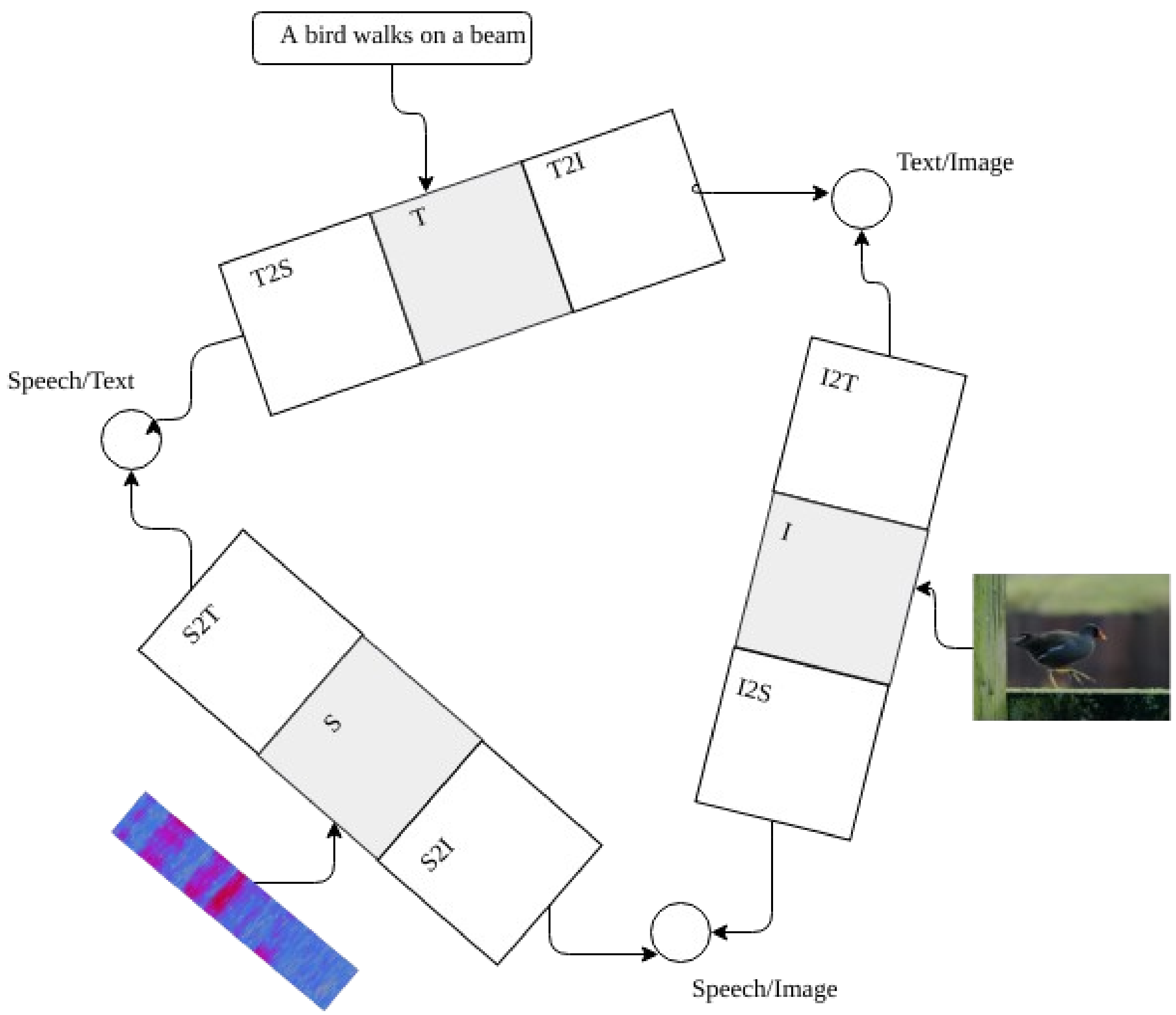
MTL

- Text only used for training (not as input)
- Representations able to encode text, but can encode other info
- No hard constraint on representations, just a nudge.

Questions

- Does MTL help?
  - Because of inductive bias or extra data?
- Which parameters should be shared?
- Which should be which task-specific?

Three-task model



- Tasks
  - **Speech/Image**
  - Speech/Text
  - Text/Image
- Tasks share some parameters

Project two modalities to joint space

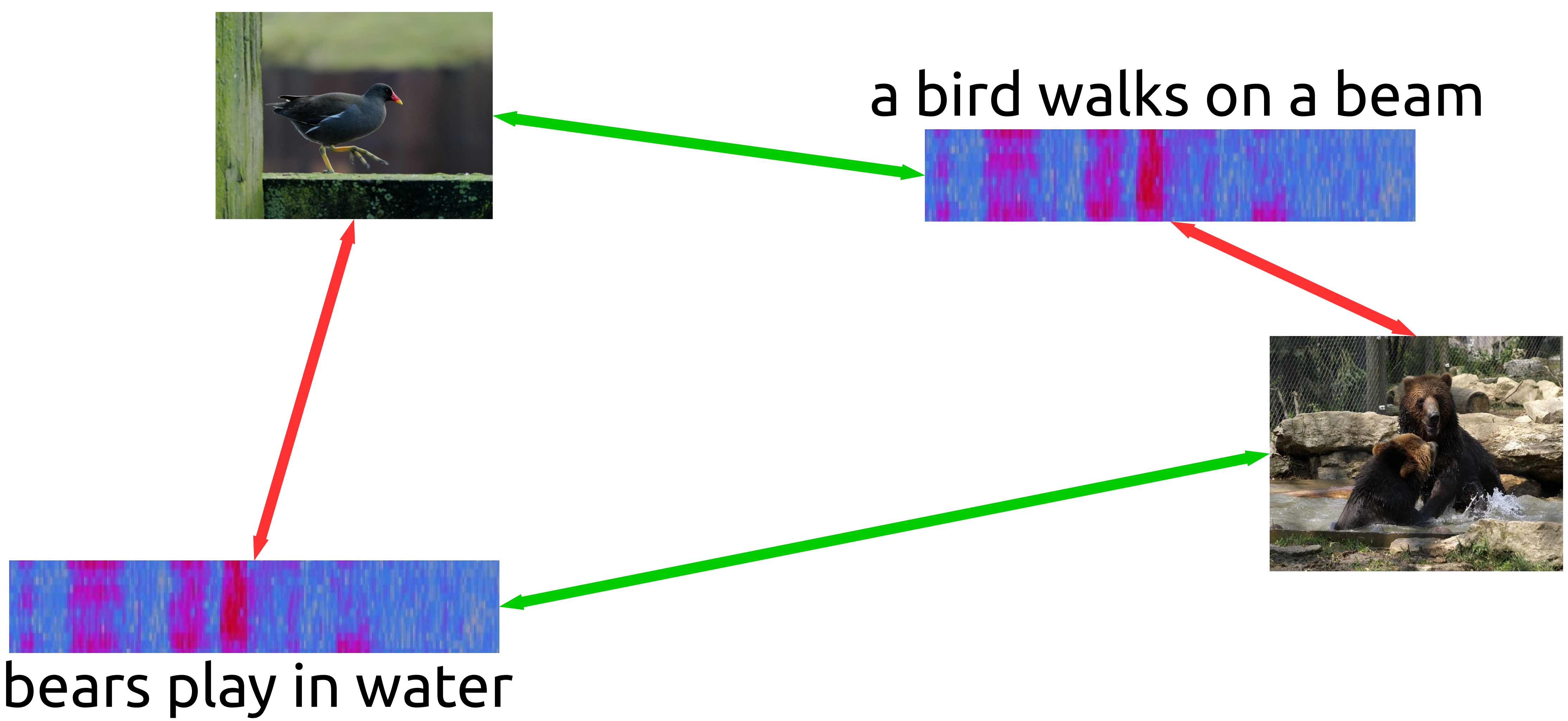
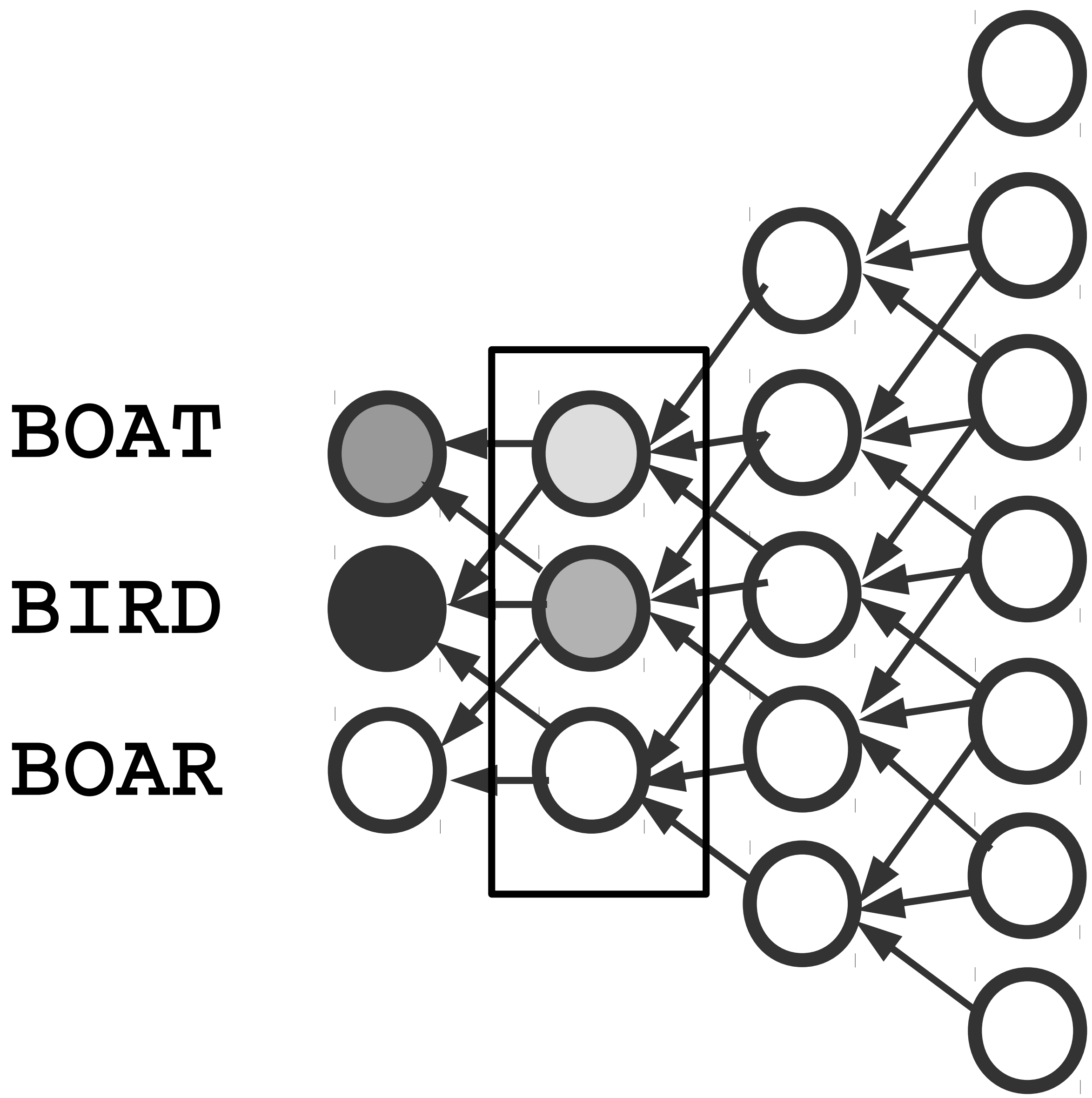
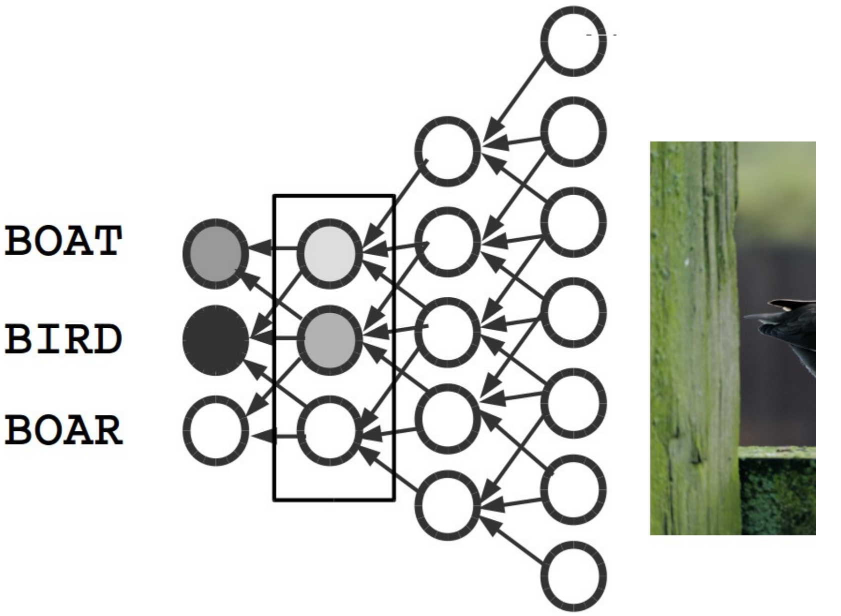


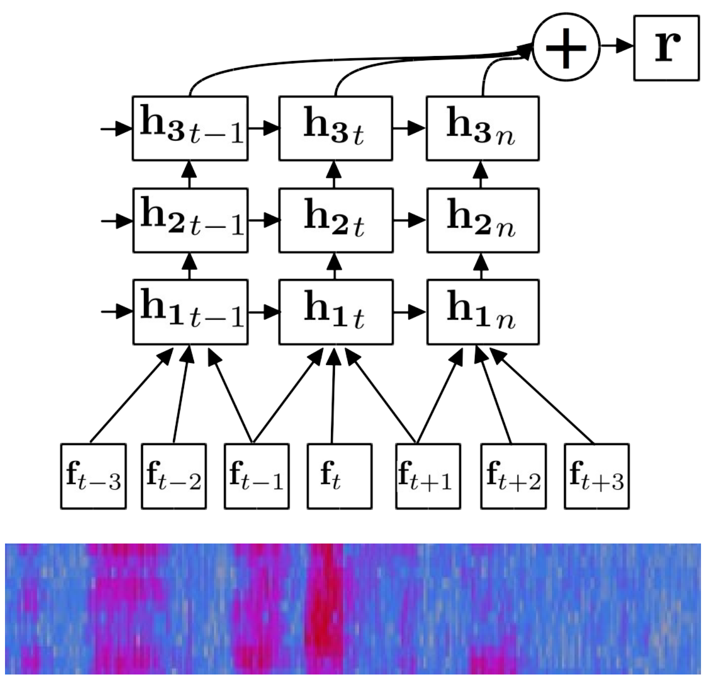
Image encoder



Encoders

- Image
  - Fixed CNN
  - Linear projection
- Speech
  - CNN layer
  - GRU RNN layers





Text encoder

- Text
  - Embedding layer (symbol lookup)
  - GRU RNN layers
- Text = sequence of characters

Evaluation metrics

- Image retrieval
  - Encode an image into joint speech/image space
  - Rank images by distance
  - Check how good the ranking is
    - Recall@K (higher better)
    - Median rank of correct image (lower better)
- Speaker identity decodability (lower better)
  - Logistic regression model on encoded speech

Experimental conditions

- Vary number of tasks (1-3)
- Vary which layers are shared
- Vary whether tasks are trained on same or different data
  - Flickr8K – speech, text, image
  - Libri – speech, text

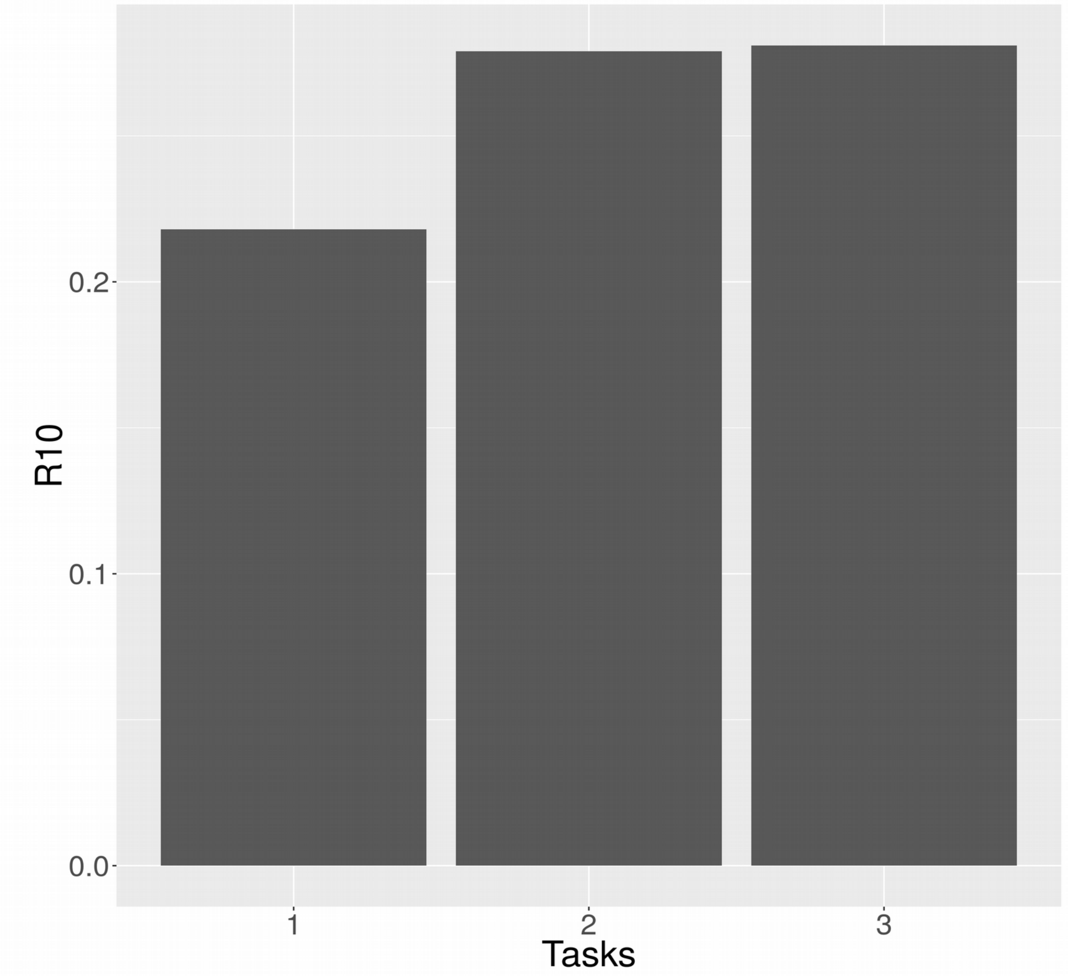
Results on validation data

Data	Tasks	S	T	S2I	S2T	T2S	T2I	R@10	Medr	Spkr
NA	1	2	.	2	.	.	.	0.218	63.8	0.297
Joint	2	2	1	2	0	0	.	0.279	42.3	0.101
Disjoint	2	2	1	2	0	0	.	0.280	41.3	0.177
Joint	3	2	1	2	0	0	1	**0.281**	**39.7**	**0.085**
Joint	3	4	1	0	0	0	0	0.248	46.3	0.211
Disjoint	3	2	1	2	0	0	1	**0.280**	**41.7**	**0.177**
Disjoint	3	4	1	0	0	0	0	0.223	59.3	0.282

1-task vs 2-task vs 3-task

Speech/Text helps.
Quite a bit.

Text/Image doesn't.

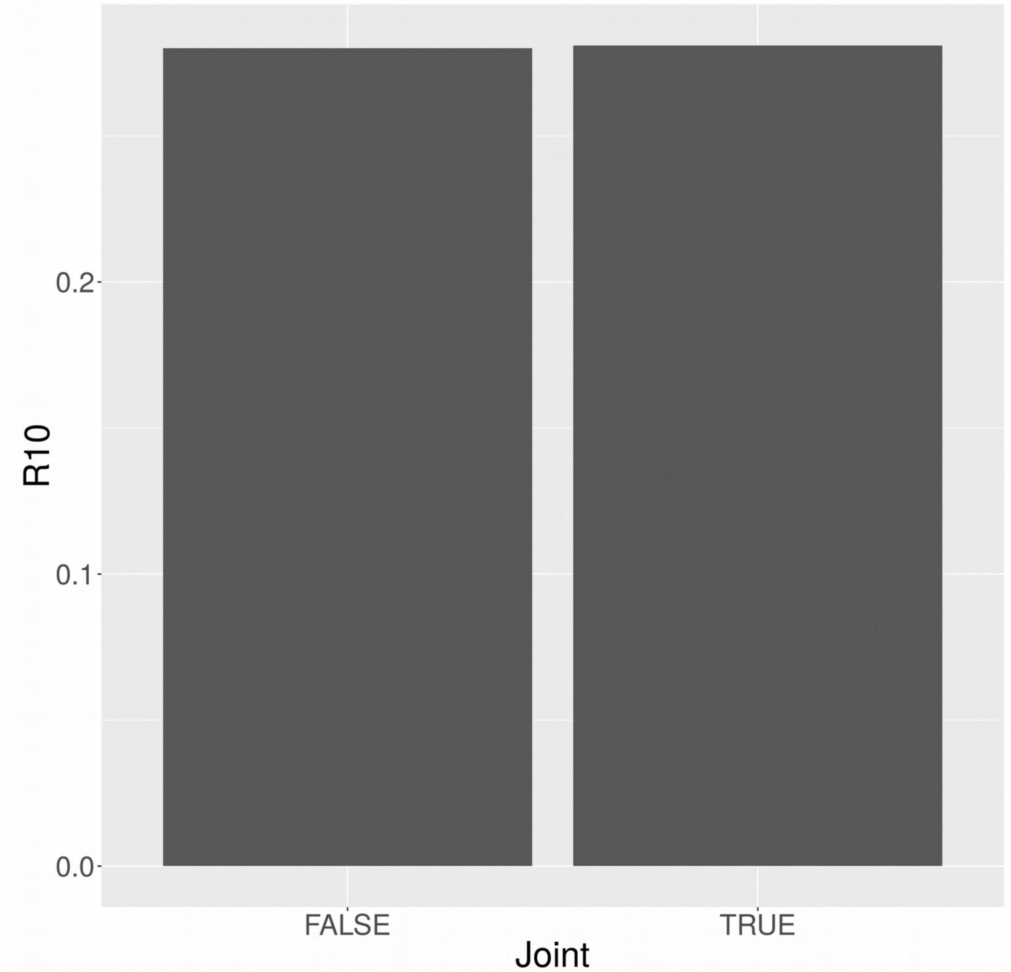


(joint)

Joint vs disjoint

Same/different data makes no difference →

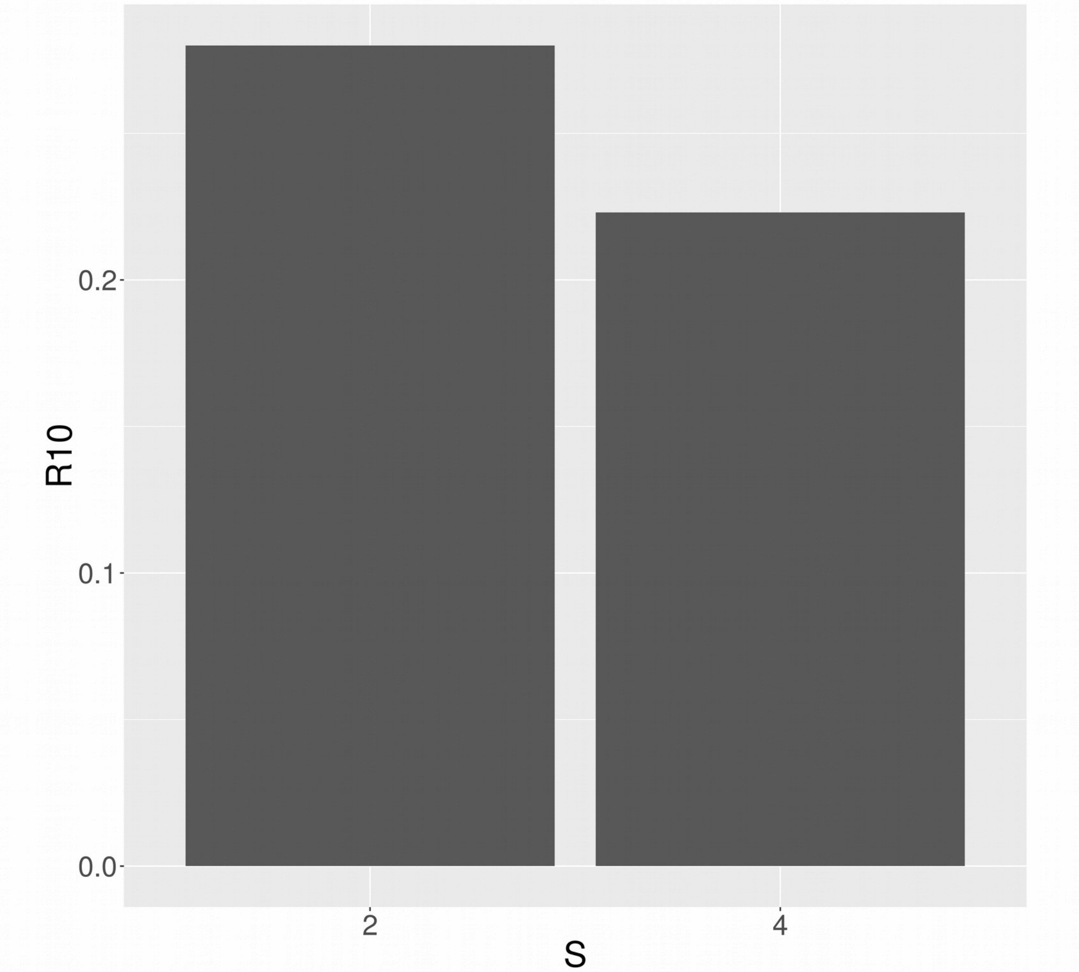
MTL helps because of inductive bias



(3-task)

Full vs partial sharing

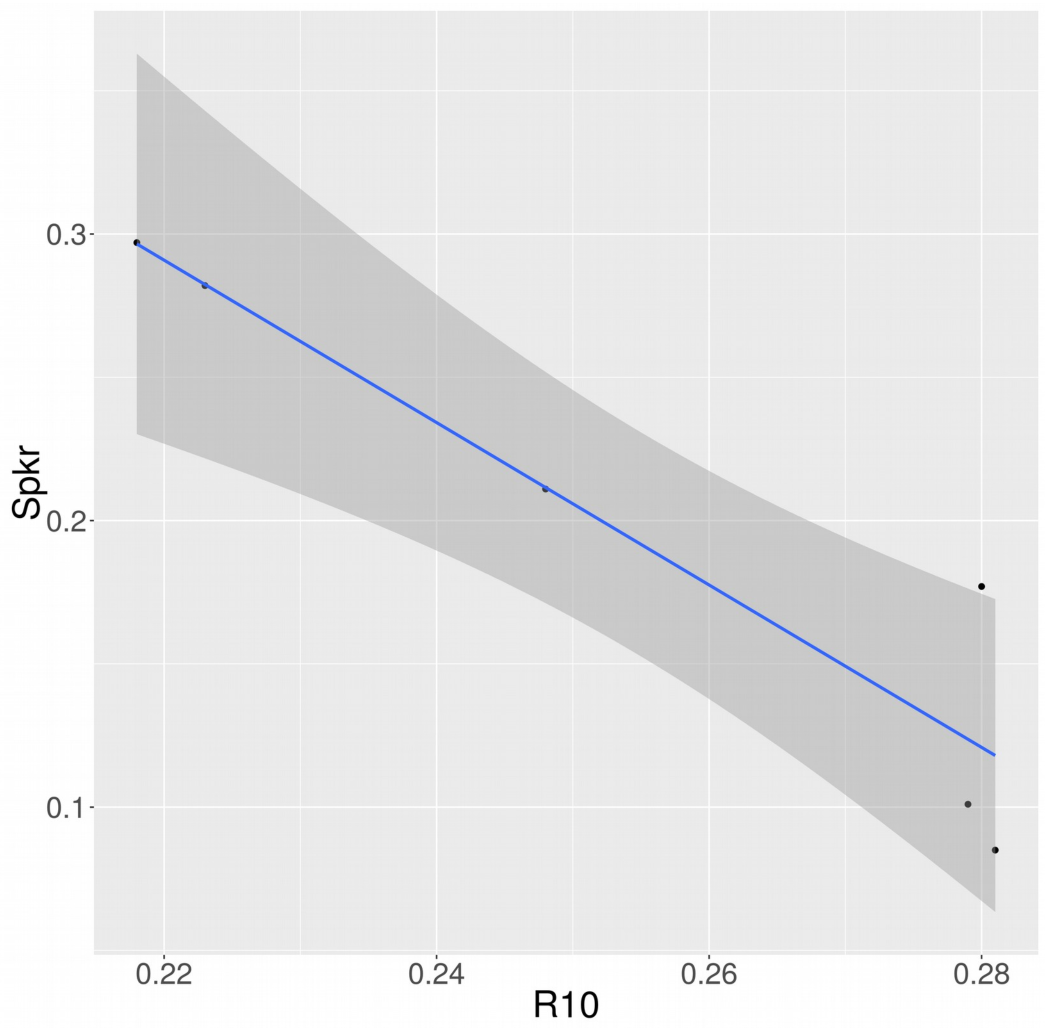
Sharing 2 bottom layers of speech encoder works better than sharing all 4 layers.



(disjoint)

Speaker decodability

Better models: more speaker-invariant.



Compared to previous work

Compared to previous single task approaches

Data	Tasks	S	T	S2I	S2T	T2S	T2I	R@10	Medr	Spkr
NA	1	Harwath and Glass 2015						0.179	-	-
NA	1	Chrupała et al 2017						0.253	48	-
NA	1	2	.	2	.	.	.	0.244	51	0.312
Joint	3	2	1	2	0	0	1	0.296	34	0.096

(test set)

Current and future work

- Speech transcription in addition to current Speech/Text task
- Compare against pipeline architecture in a controlled fashion
- Evaluate with varying data size for auxiliary task