Empirical Study on Transfer Learning for Text Classification

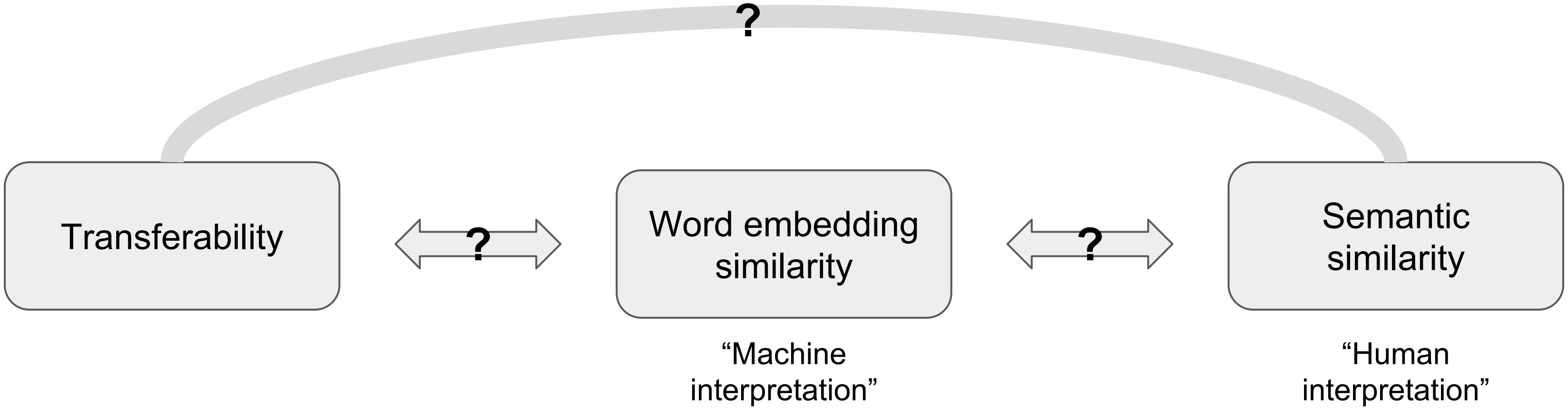
Yunshu Du
Supervised by Nidhi Hegde
Borealis AI internship
Sept 13, 2018. Edmonton, AB

Motivation

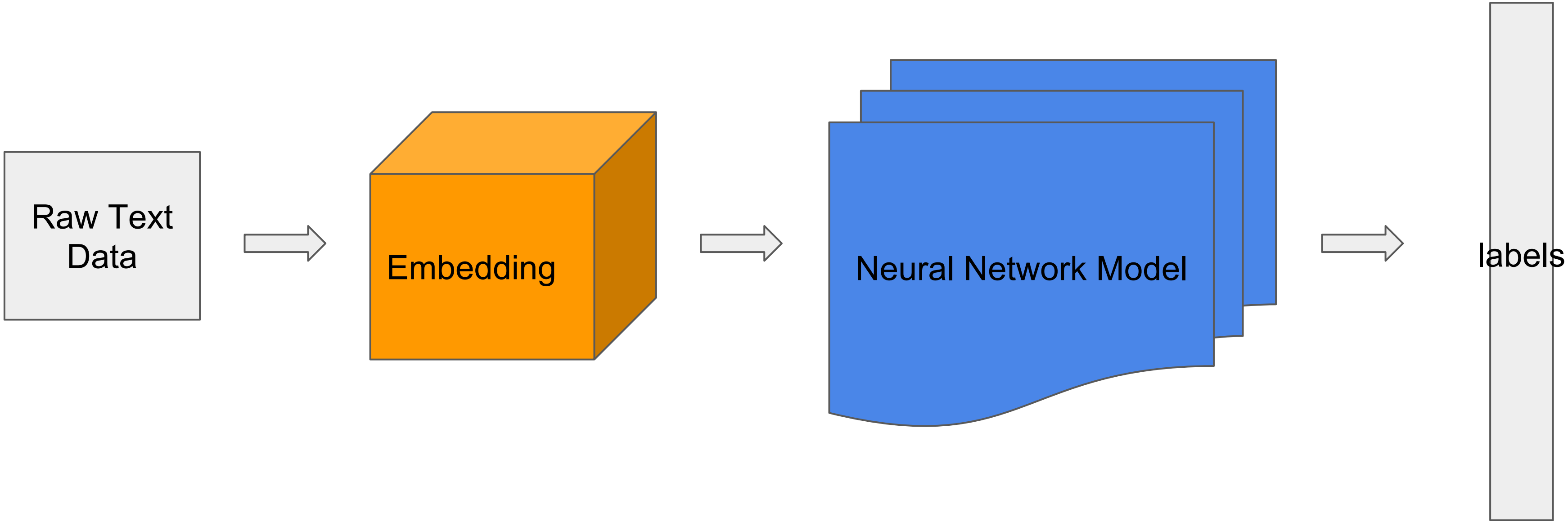
- Apollo
  - Entity classification based on news articles
  - Need rich historical data for training
  - How to make prediction if new entities or a new articles present?

- Transfer learning is important in many domains
  - Computer vision, reinforcement learning, natural language processing...

We are interested in...



Text classification task



Data

- StackExchange Questions (SE)

id	**title**	content	**tag**

- Sentence Involving Compositional Knowledge (SICK)

id	sent_1	sent_2	sim	label
1	A group of kids is playing in a yard and an old man...	A group of boys in a yard is playing and a man...	4.5	NETURAL
2	An elderly man is sitting on a bench	An old person is sitting on a bench	4.6	ENTAILMENT
3	A cat is stuck on a moving ceiling fan	There is no cat swinging on a fan	2.8	CONTRADUCTION

Embedding Methods

- W2V
  - Google pre-trained model (*google-w2v*)
  - Self-pretrained (*self-w2v*)
- GloVe
  - The smallest Stanford pre-trained model: glove.6b (*stanford-glove*)
  - Self-pretrained (*self-glove*)
- Random
  - Uniformly random initialize embedding weights (*rand*)

Others: ELMO, InferSent, USE

Models

- CNN
    - Embedding layer
    - 3 CNN layers
    - 1 fully connected layer + softmax
- RNN
    - Embedding layer
    - 128 hidden cells
        - Vanilla
        - LSTM
    - 1 fully connected layer + softmax
- FastText (*FT*)
    - 1 fully connected layer + softmax

Training options

- Transfer and freeze embedding
- Transfer and finetune embedding
- Finetune everything else

- Adam optimizer
- Cross-entropy loss
- Batch size: 32 (smaller seems better)

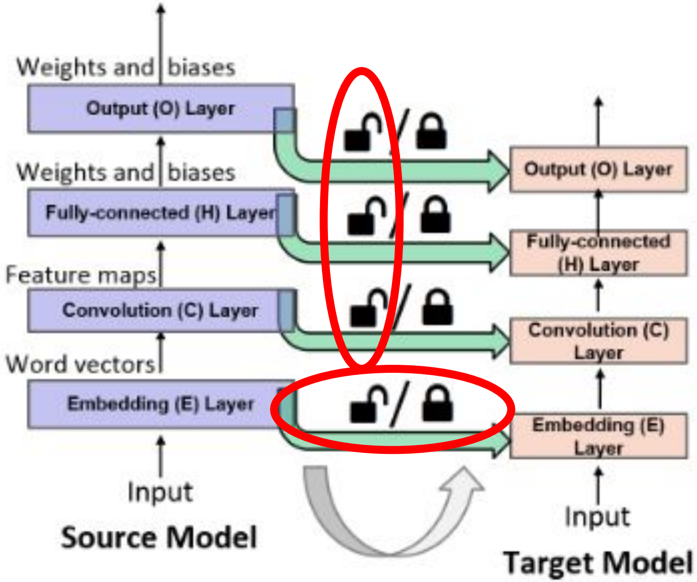
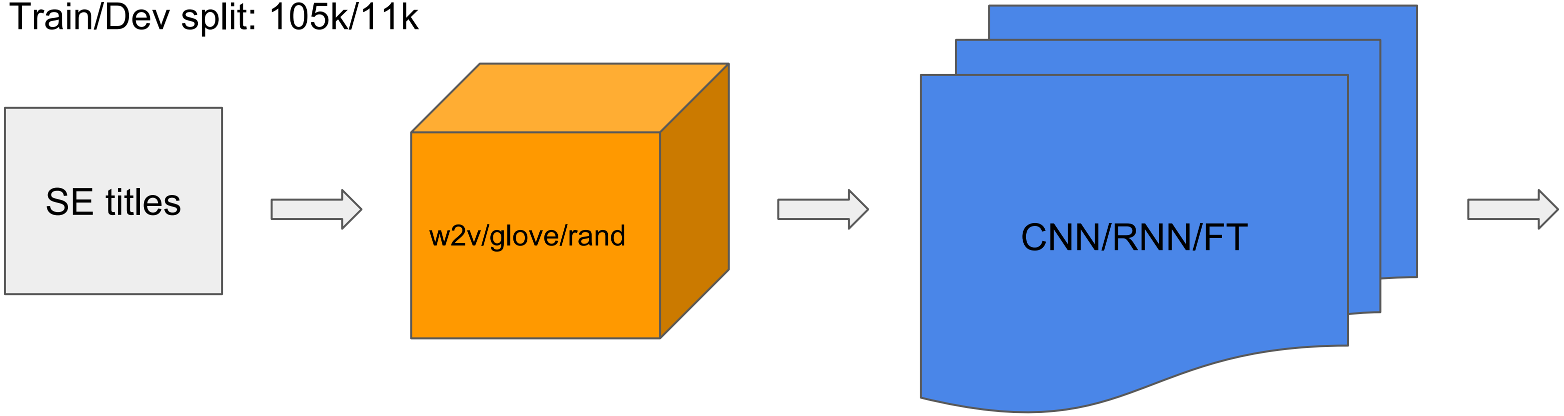


Figure from: [A Practitioners' Guide to Transfer Learning for Text Classification using Convolutional Neural Networks]

Step 0: build a baseline classifier

Vocabulary Size: 37056
Train/Dev split: 105k/11k



id	**title**	content	tag
1	How important is the minimum layer height on a 3d printer?	...	quality, resolution

More details in bitbucket: https://bitbucket.org/rbcmllab/nlp-transfer/src/master/Milestone%20%231%20(Jun19-Jul10).md

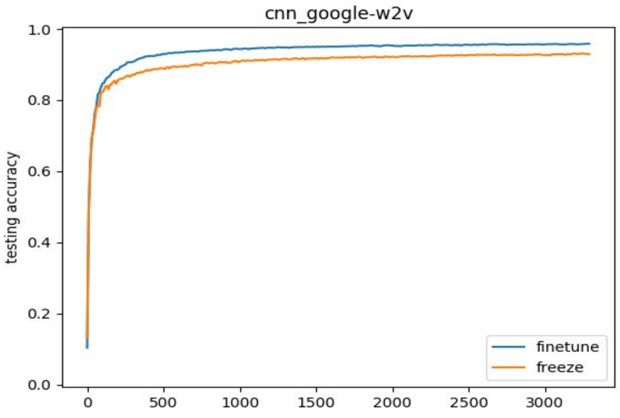
Step 0: build a baseline classifier

- How does each component affect learning
  - Freeze vs. finetune?
  - Embedding methods?
  - Model architectures?

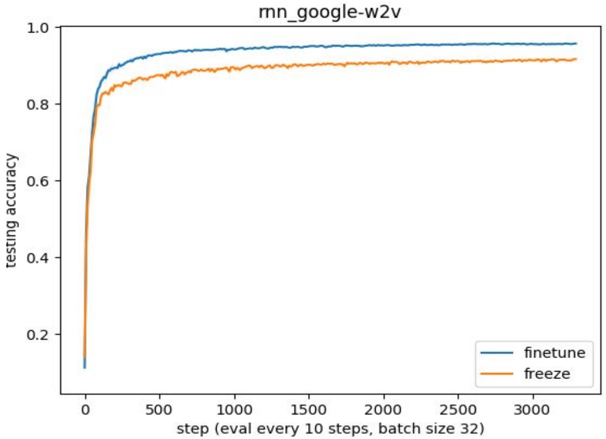
More details in bitbucket: https://bitbucket.org/rbcmllab/nlp-transfer/src/master/Milestone%20%231%20(Jun19-Jul10).md

How does each component affect learning

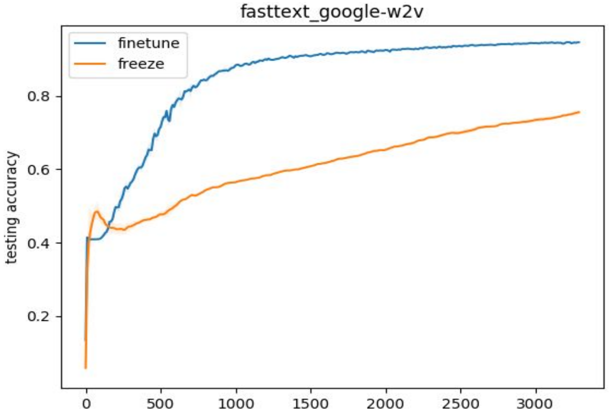
- Freeze vs. **Finetune** (google-w2v embedding)



CNN



RNN-LSTM

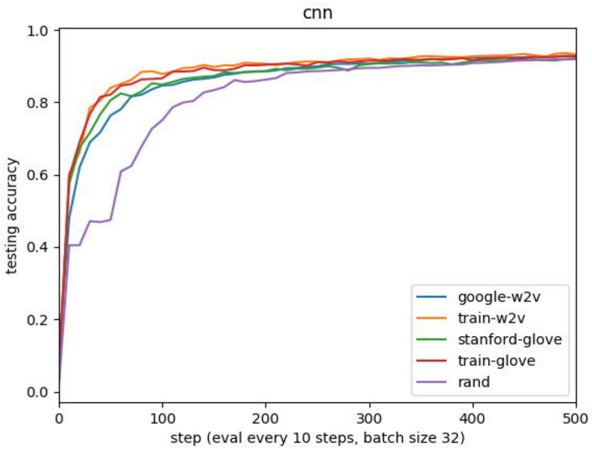


FT

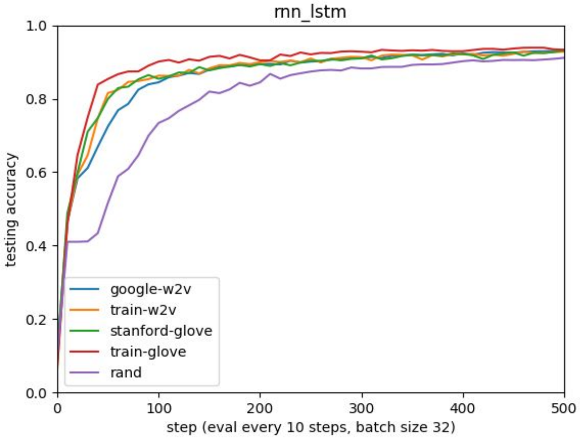
More details in bitbucket: https://bitbucket.org/rbcmllab/nlp-transfer/src/master/Milestone%20%231%20(Jun19-Jul10).md

How does each component affect learning

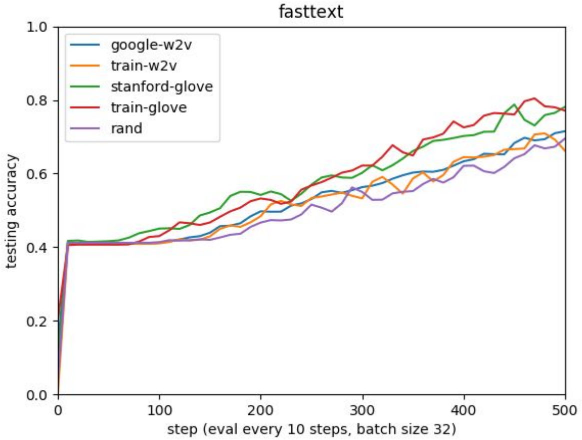
- Which embedding is better?
  - The two “self-train” embeddings seem to be slightly better



CNN



RNN-LSTM

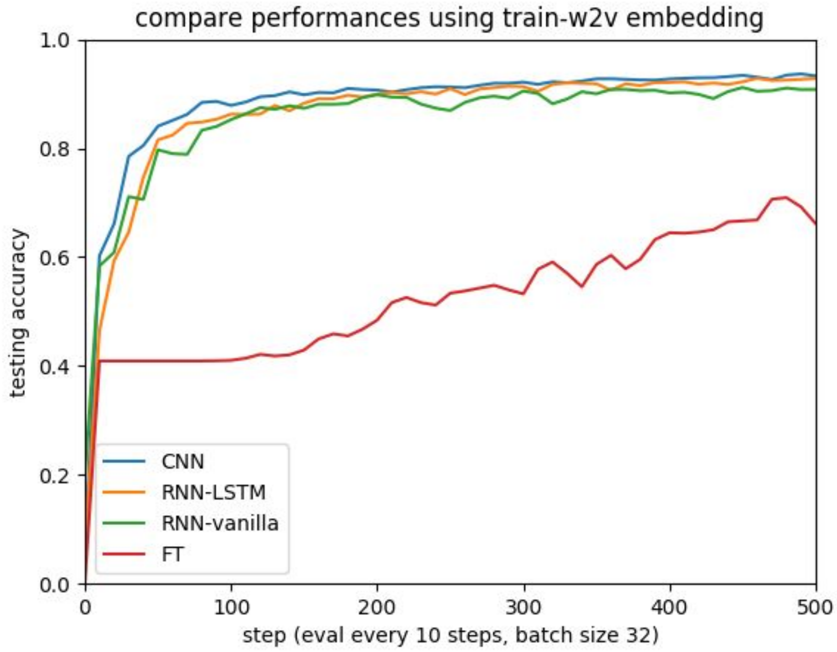


FT

More details in bitbucket: https://bitbucket.org/rbcmllab/nlp-transfer/src/master/Milestone%20%231%20(Jun19-Jul10).md

How does each component affect learning

- Which model architecture is better?
  - Compared across models with train-w2v embedding



More details in bitbucket: https://bitbucket.org/rbcmllab/nlp-transfer/src/master/Milestone%20%231%20(Jun19-Jul10).md

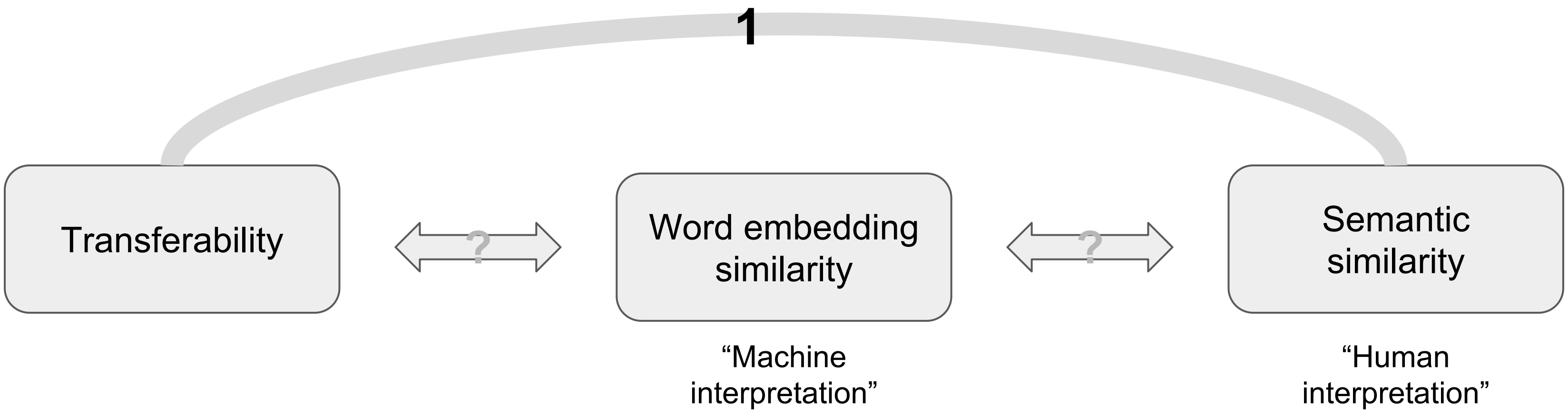
Summary on Step 0

- Always finetune the embedding layers
- When your dataset is big enough, self-trained embedding could be better
- CNN seems better in short sentences; consider RNN if sentences are long; one layer model could also achieve comparable results

- What else to try?
  - freeze/finetune other layers
  - Finetune learning rate (Existing framework: ULMFiT)
  - Try larger scale classification

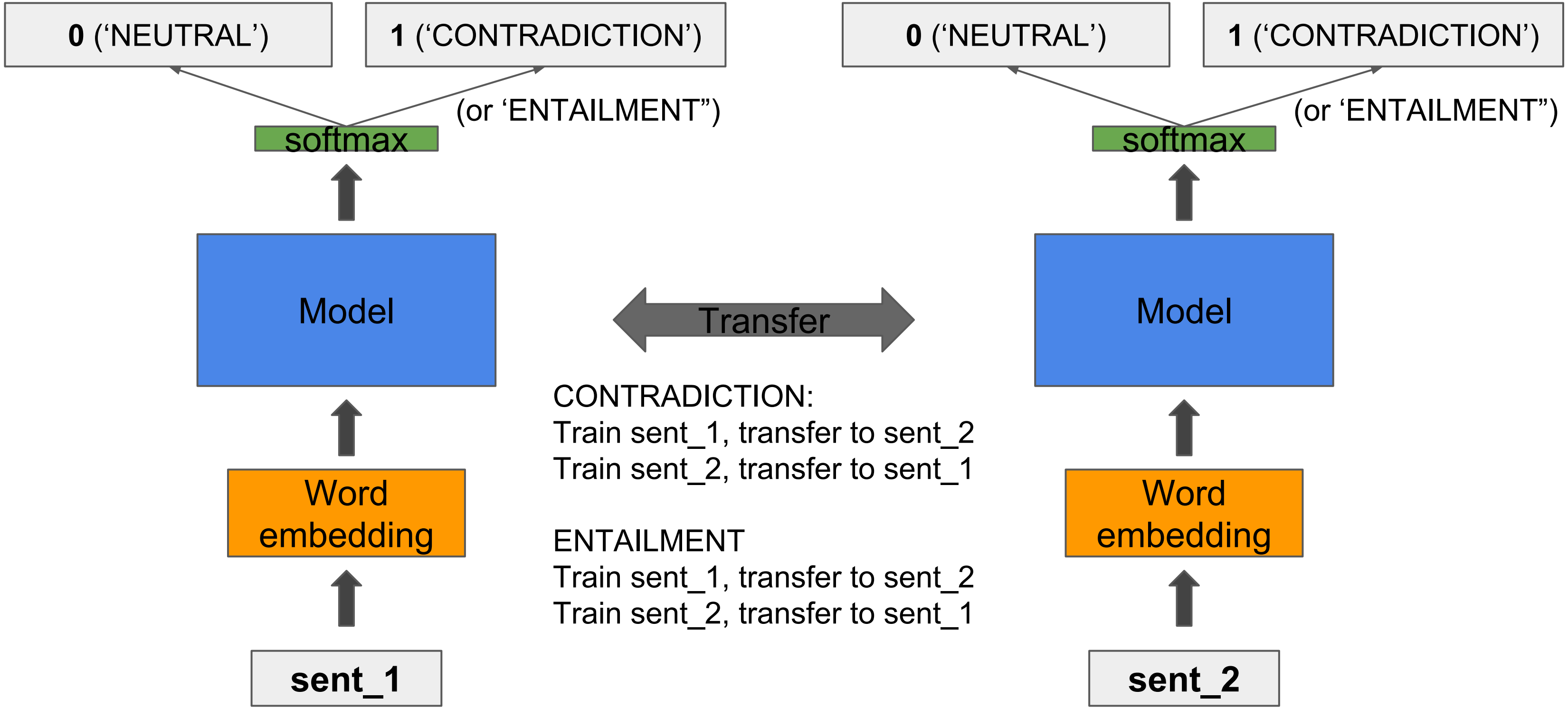
More details in bitbucket: https://bitbucket.org/rbcmllab/nlp-transfer/src/master/Milestone%20%231%20(Jun19-Jul10).md

Step 1: transferability ⟷ semantic similarity



SICK data "classification"

id	**sent_1**	**sent_2**	sim	**label**



SICK data "classification"

id	**sent_1**	**sent_2**	sim	**label**

Hypothesis:

"ENTAILMENT" should transfer better than "CONTRADICTION" if we assume that sentences that are "entailed" based on human interpretation means more "similarity"

SICK data “classification”

id	**sent_1**	**sent_2**	sim	**label**

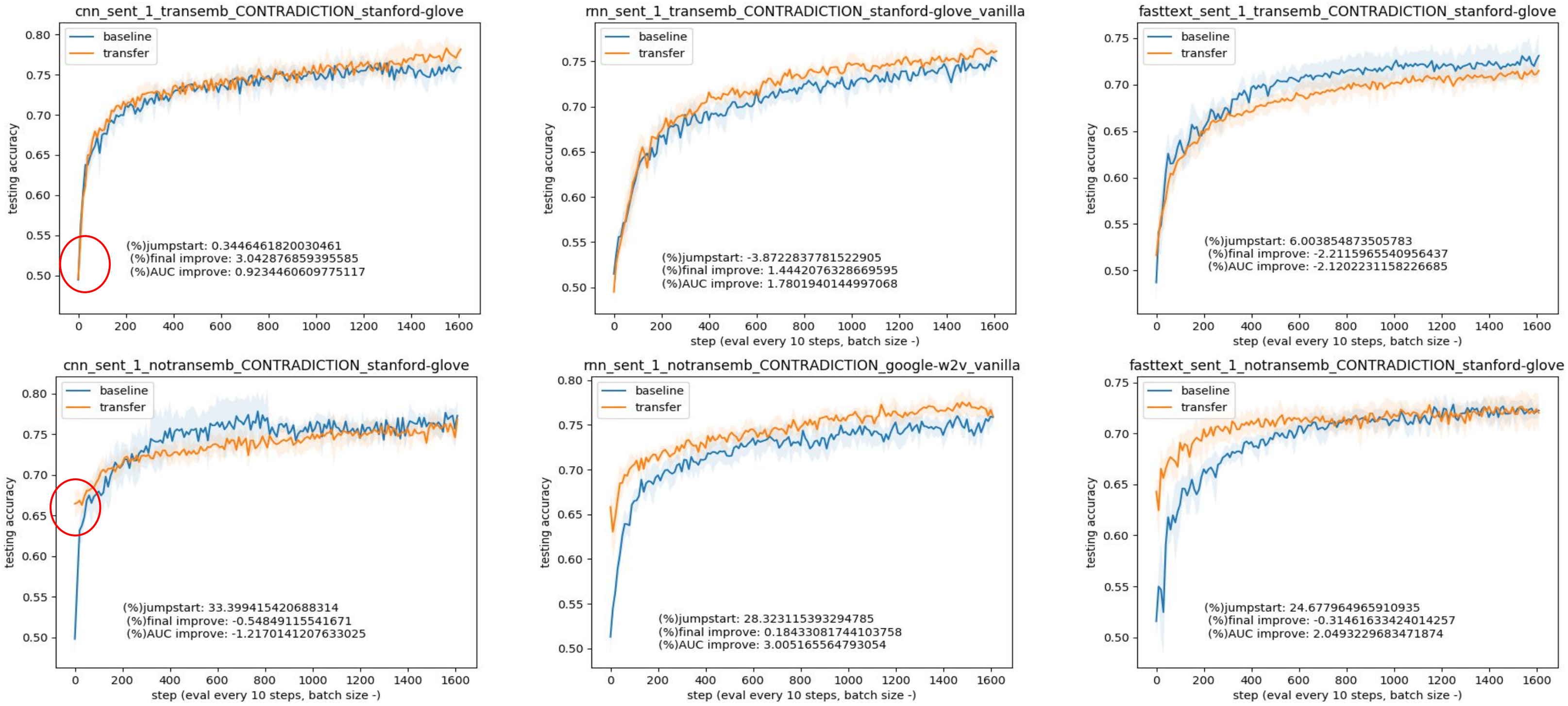
- Evaluate baseline vs. transfer:
  - Jumpstart: accuracy differences at step 0
  - Final improve: accuracy differences at the last step
  - AUC (area under the curve) improve: accumulated accuracy differences

SICK data “classification”: observations

- Do not transfer the embedding layer (stanford-glove, CONTRADICTION)

Transfer emb

Not Transfer emb

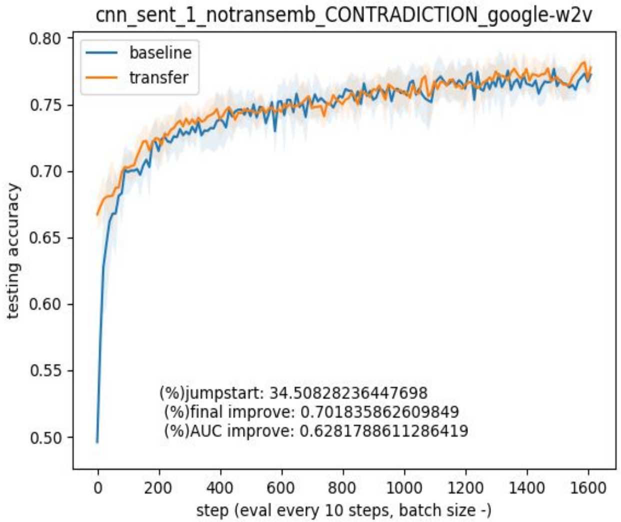


CNN RNN-vanilla FT

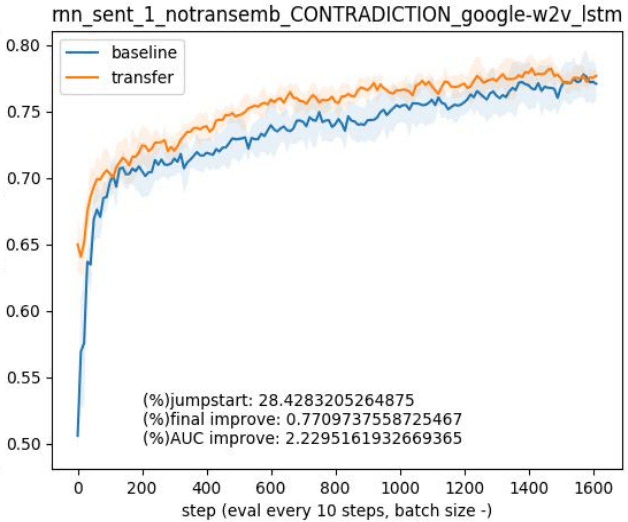
SICK data “classification”: observations

- Distinct behaviours among models;

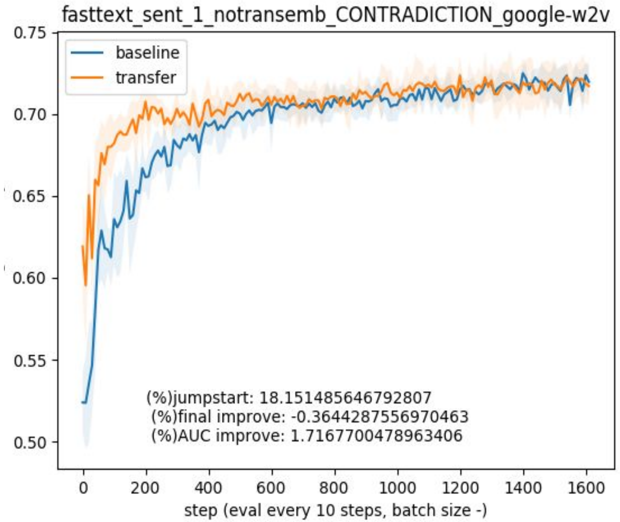
(google-w2v, CONTRADICTION)



CNN



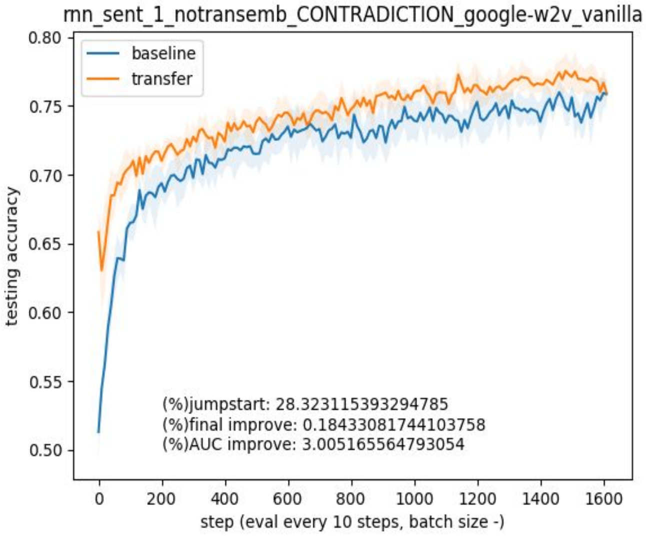
RNN-LSTM



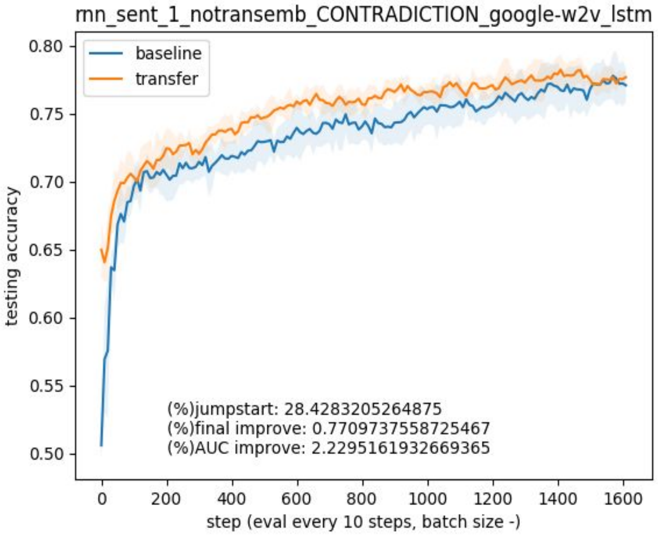
FT

SICK data "classification": observations

- Distinct behaviours among models; while RNN seems to be the most suitable for this transfer task. (google-w2v, CONTRADICTION)



RNN-vanilla



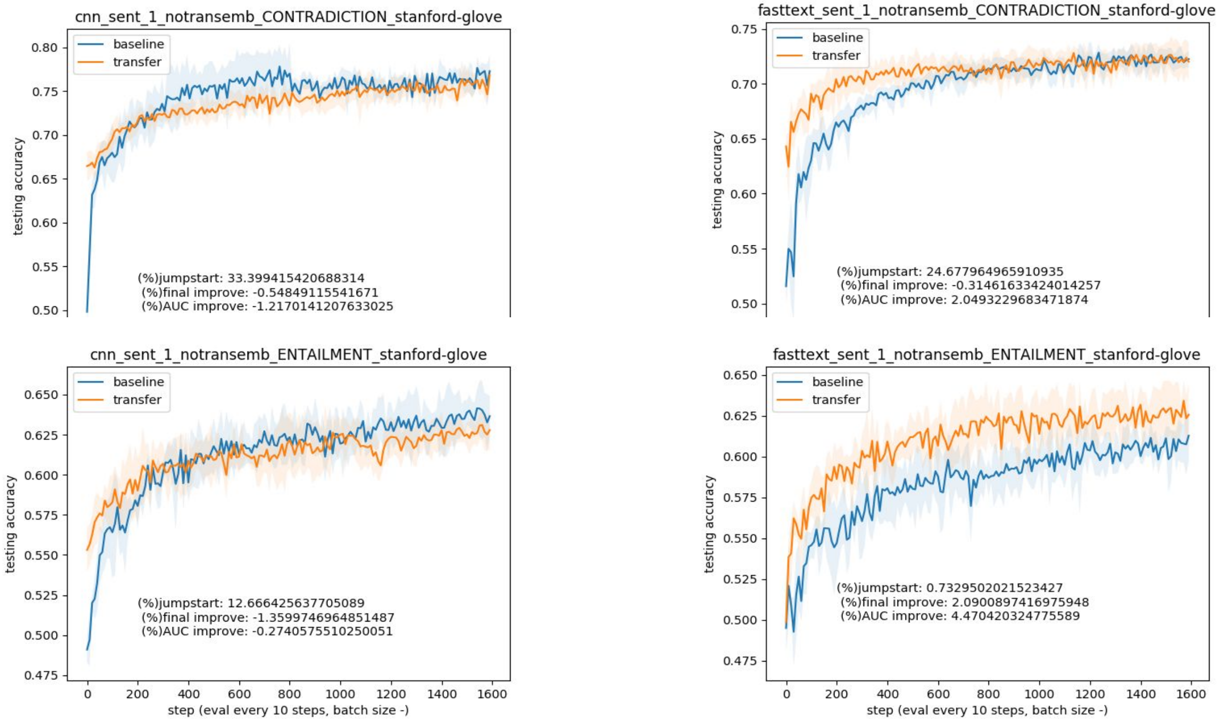
RNN-LSTM

SICK data “classification”: observations

- Overall, “CONTRADICTION” transfers slightly better than “ENTAILMENT”;
- In a few runs “ENTAILMENT” performed well (stanford-glove)

CONTRA

ENTAIL



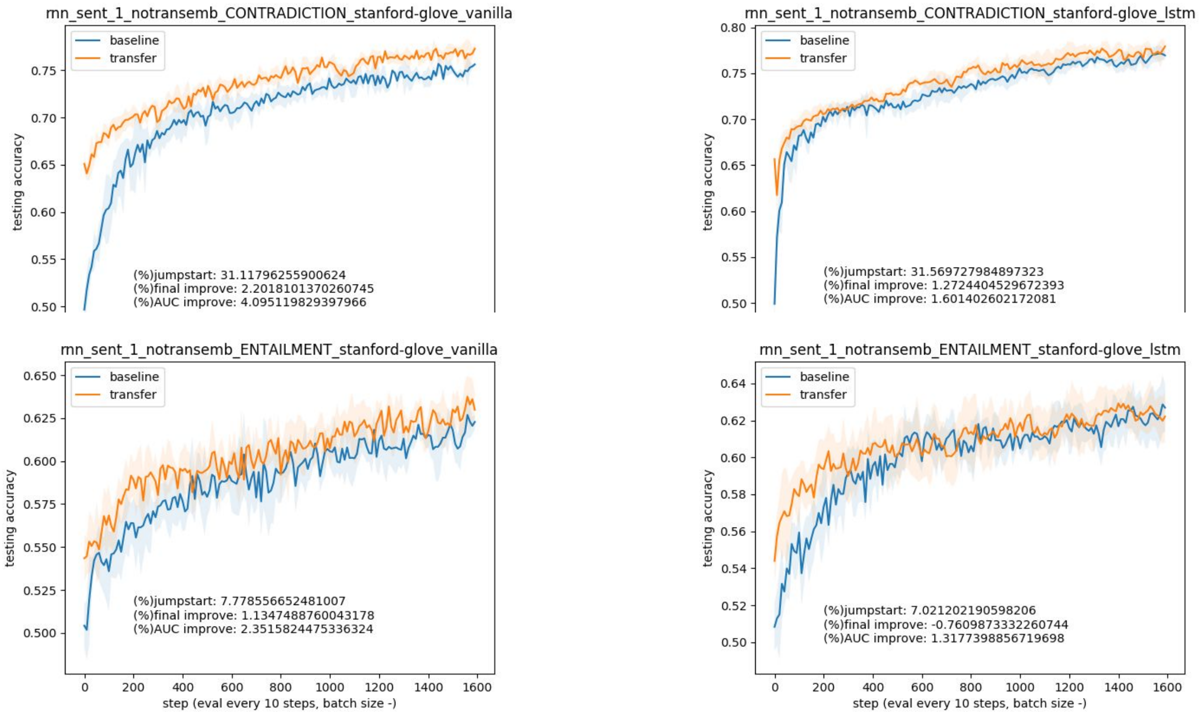
CNN FT

SICK data “classification”: observations

- Overall, “CONTRADICTION” transfers slightly better than “ENTAILMENT”;
- In a few runs “ENTAILMENT” performed well (stanford-glove)

CONTRA

ENTAIL



RNN-vanilla RNN-LSTM

SICK data “classification”

id	**sent_1**	**sent_2**	sim	**label**

Hypothesis:

“ENTAILMENT” should transfer better than “CONTRADICTION” if we assume that sentences that are “entailed” based on human interpretation means more “similarity”

SICK data "classification"

id	**sent_1**	**sent_2**	sim	**label**

Hypothesis:

"ENTAILMENT" should transfer better than "CONTRADICTION" if we assume that sentences that are "entailed" based on human interpretation means more "similarity"

However…

ENTAILMENT

- *If A is true, then B is true*:
  - A: A dog is running in a field
  - B: An animal is running in a field

NETURAL

- *If A is true, then B cannot be said to be true or false*:
  - A: A man is breaking three eggs in a bowl
  - B: A girl is pouring some milk in a bowl

CONTRADICTION

- *If A is true, then B is false*:
  - A: A man is playing golf
  - B: No man is playing golf

Lesson learned in Step 1

- Experiments were not-so-appropriately designed due to a misunderstanding in the dataset --- understand your data is important

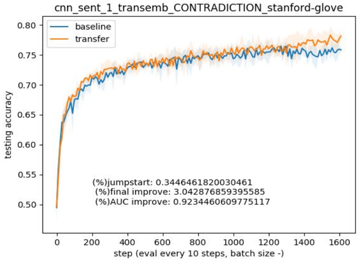
Lesson learned in Step 1

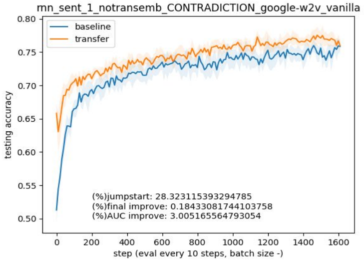
- Experiments were not-so-appropriately designed due to a misunderstanding in the dataset --- understand your data is important
- Directly connect transferability with human interpretation of text is too big of a step, need more bridges.

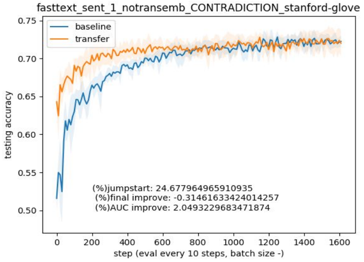
Lesson learned in Step 1

- Experiments were not-so-appropriately designed due to a misunderstanding in the dataset --- understand your data is important
- Directly connect transferability with human interpretation of text is too big of a step, need more bridges.

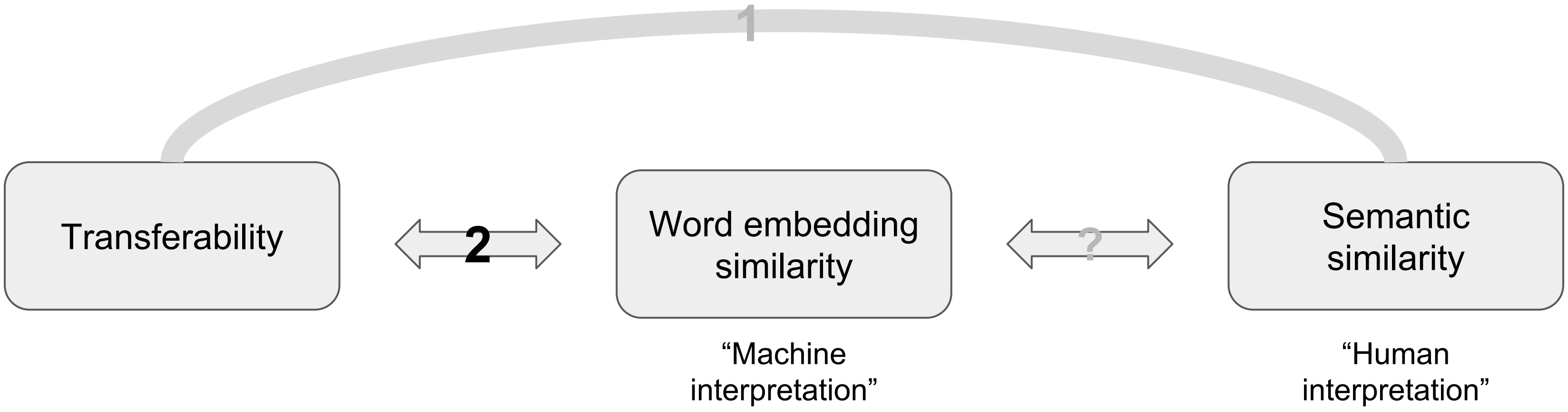
- But...still observed transfer patterns in different model/embedding methods.



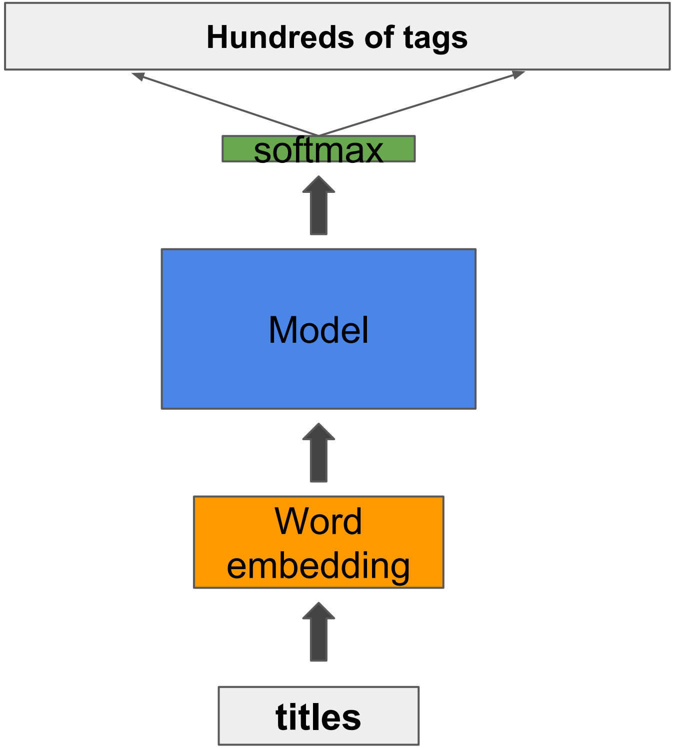




Step 2: transferability ⟷ embedding similarity



Going back to SE data



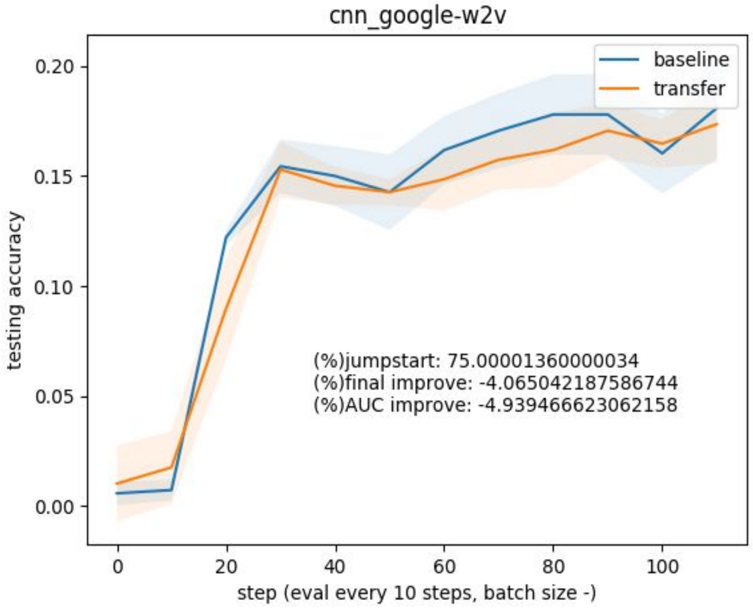
id	**title**	content	**tag**
1	What is the right approach to write the spin controller for a soccer robot?	...	soccer control

- Source: biology, cooking, crypto, diy, robotics, travel
  - Vocabulary Size: 33398
  - Train/Dev split: 78300/8700
  - #tags: 4268
- Target: bioinformatics
  - Vocabulary Size: 2705
  - Train/Dev split: 1228/136
  - #tags: 379

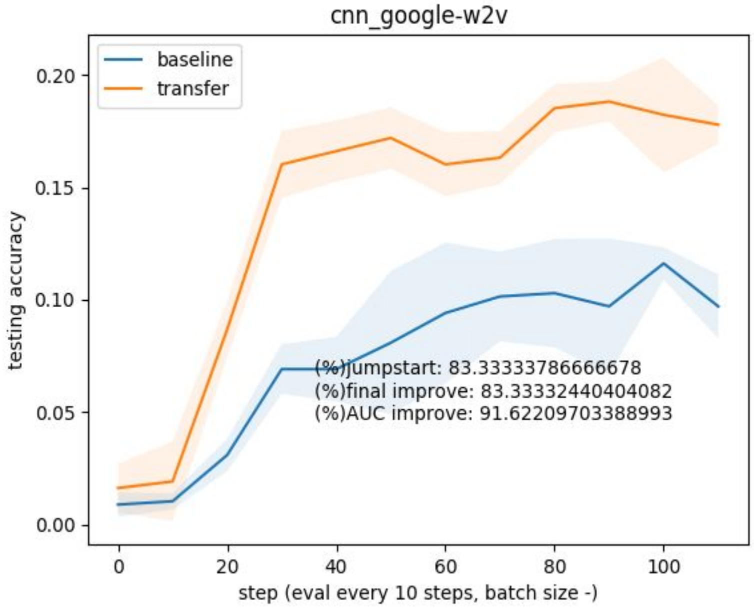
1. Perform transfer
2. Compute embedding distances
3. Is transferability correlated with embedding distances?

Transfer Results

- All sources to “bioinformatics” vs. “biology” to “bioinformatics”
  - CNN model with google-w2v embedding



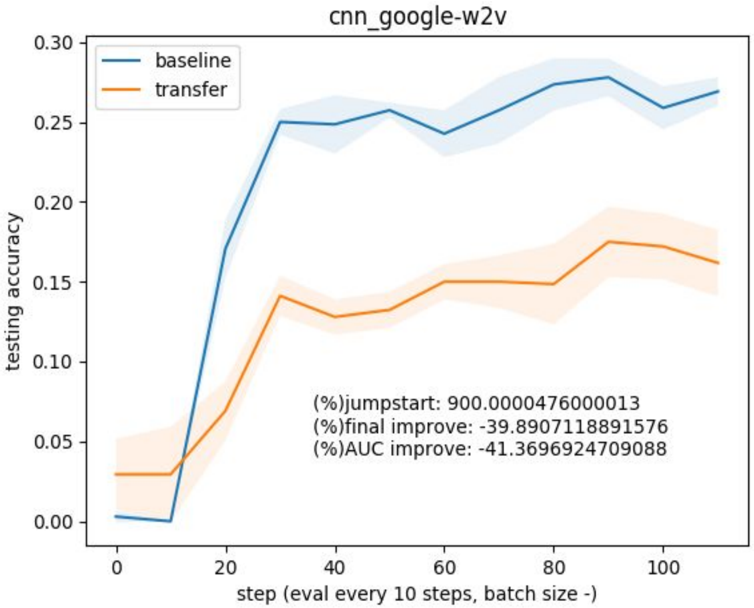
All to “bioinformatics”



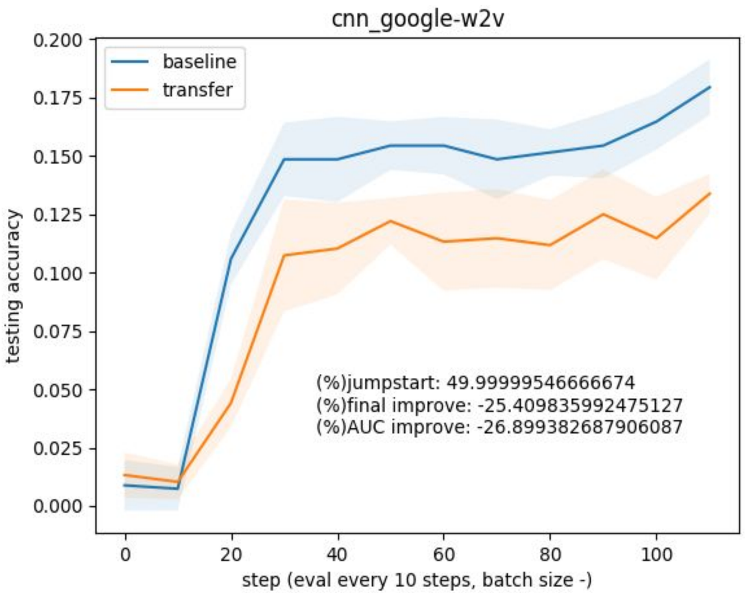
“biology” to “bioinformatics”

Transfer Results

- "diy" to "bioinformatics" vs. "robotics" to "bioinformatics"
  - CNN model with google-w2v embedding



"diy" to "bioinformatics"

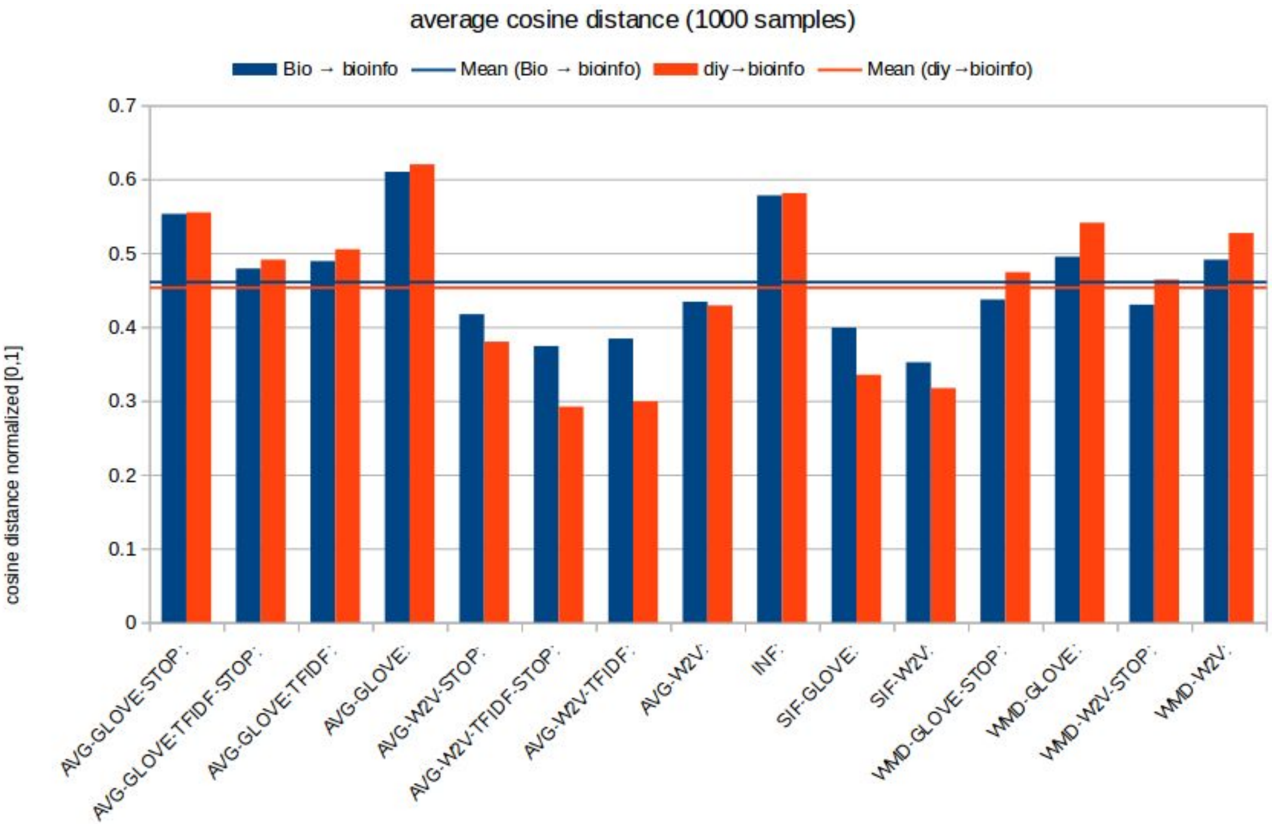


"robotics" to "bioinformatics"

Look at embedding distances

- *m* to *n* pairwise comparison on cosine similarity over the entire source and target sentences, then take the average score (normalized to [0, 1])
    - To save time, we do random sample of 1000 sentences from source and target and average over 5 iterations. (1000 x 1000 x 5 comparisons)
    - Score -> 0: near; Score -> 1: far

Look at embedding distances



Biology -> Bioinformatics: avg 0.46

Diy -> Bioinformatics: avg 0.45

Revisit SICK data

- Is “CONTRADICTION” actually closer in embedding distance? Does that explain why its transfer performed better than “ENTAILMENT”?

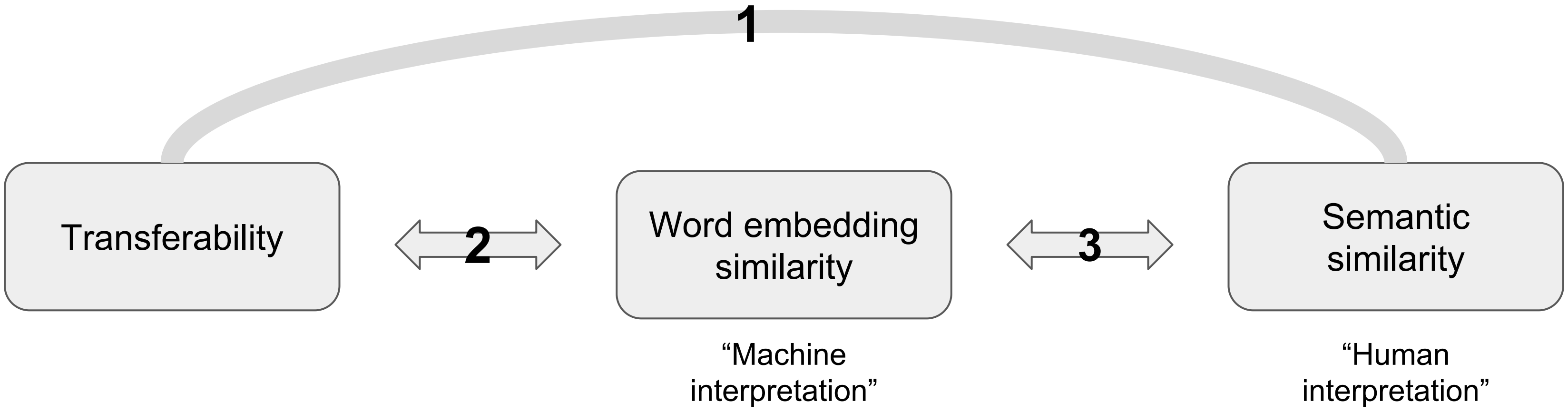
Revisit SICK data

- Is "CONTRADICTION" actually closer in embedding distance? Does that explain why its transfer performed better than "ENTAILMENT"?
- Based on 1 measurement using USE, CONTRADICTION and ENTAILMENT shows the same embedding distance of 0.3

sent_1	**sent_2**	**label**
...	**...**	ENTAILMENT

sent_1	**sent_2**	**label**
...	**...**	CONTRADICTION

Step 3: embedding ⟷ semantic similarity



An existing recent work: [Evaluation of sentence embeddings in downstream and linguistic probing tasks]

Takeaway

- Always finetune a pretrained model with your data (at least in the embedding)
    - But do not transfer the embedding layer
    - If your data is big enough, consider training an embedding from scratch
- Model/embedding selection is still task-dependent
- There are some patterns in transferability vs. “similarity”, but
    - One will need to define a similarity measurement accordingly, multiple measurements should be evaluated
    - In this work we looked at “semantic similarity” as a measurement for transferability, but no solid conclusion on the correlation

Thank you!

Empirical Study on Transfer Learning for Text Classification

Yunshu Du
Supervised by Nidhi Hegde
Borealis AI internship
Sept 13, 2018. Edmonton, AB