– IN5550 –

Neural Methods in Natural Language Processing

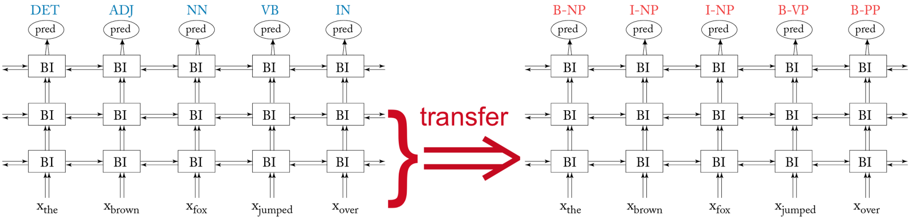
Ensembles, transfer and multi-task learning: Part 3

Erik Velldal

Language Technology Group (LTG)
University of Oslo

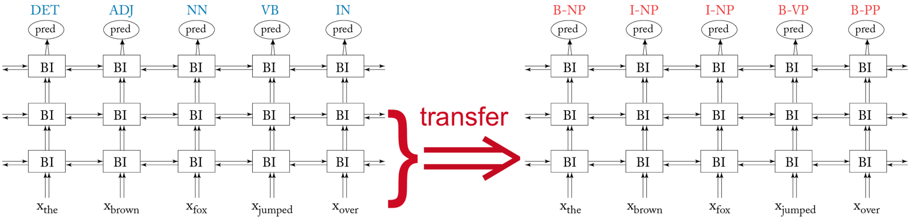
Transfer learning

- Learn a model M1 for task A, and re-use (parts of) M1 in another model M2 to be (re-)trained for task B.
- Example: Transfer learning with tagging as the source task and chunking as the target (destination) task.



Transfer learning

- Learn a model M1 for task A, and re-use (parts of) M1 in another model M2 to be (re-)trained for task B.
- Example: Transfer learning with tagging as the source task and chunking as the target (destination) task.



- Can you think of any examples of transfer learning we've seen so far?

Related notions

- Self-supervised learning:
- Making use of unlabeled data while learning in a supervised manner.
- E.g. word embeddings, trained by predicting words in context.
- Pretrained LMs most widely used instance of transfer in NLP.

Related notions

- Self-supervised learning:
- Making use of unlabeled data while learning in a supervised manner.
- E.g. word embeddings, trained by predicting words in context.
- Pretrained LMs most widely used instance of transfer in NLP.

- Transfer sometimes applied for domain adaptation:
- same task but different domains or genres.

Related notions

- Self-supervised learning:
- Making use of unlabeled data while learning in a supervised manner.
- E.g. word embeddings, trained by predicting words in context.
- Pretrained LMs most widely used instance of transfer in NLP.

- Transfer sometimes applied for domain adaptation:
- same task but different domains or genres.
- Also used in cross-lingual approaches,

Related notions

- Self-supervised learning:
- Making use of unlabeled data while learning in a supervised manner.
- E.g. word embeddings, trained by predicting words in context.
- Pretrained LMs most widely used instance of transfer in NLP.

- Transfer sometimes applied for domain adaptation:
- same task but different domains or genres.
- Also used in cross-lingual approaches,
- and as part of distillation.

TL/MTL and regularization

- MTL can be seen as a regularizer in its own right; keeps the weights from specializing too much to just one task.
- With transfer on the other hand, there is often a risk of unlearning too much of the pre-trained information:
- 'Catastrophic forgetting' (McCloskey & Cohen, 1989; Ratcliff, 1990).

TL/MTL and regularization

- MTL can be seen as a regularizer in its own right; keeps the weights from specializing too much to just one task.
- With transfer on the other hand, there is often a risk of unlearning too much of the pre-trained information:
- 'Catastrophic forgetting' (McCloskey & Cohen, 1989; Ratcliff, 1990).
- May need to introduce regularization for the transfered layers.
- Extreme case: frozen weights (infinite regularization)
- Not unusual to only re-train selected parameters / higher layers.
- Other strategies: gradual unfreezing, reduced or layer-specific learning rates (in addition to early stopping, dropout, L2, etc.)

When is TL/MTL most useful

- When low-level features learned for task A could be helpful for learning task B.
- When you have limited labeled data for your main/target task and want to tap into a larger dataset for some other related aux/source task.

TL/MTL in NLP

- TL/MTL is particularly well-suited for neural models:
- Representation learners! With a modular design.

TL/MTL in NLP

- TL/MTL is particularly well-suited for neural models:
- Representation learners! With a modular design.

- Intuitively very well-suited for NLP too:
- Due to the complexity of the overall task of NLP (understanding language), it has been split up into innumerable sub-tasks.
- Typically have rather small labeled data sets, but closely related tasks.

TL/MTL in NLP

- TL/MTL is particularly well-suited for neural models:
- Representation learners! With a modular design.

- Intuitively very well-suited for NLP too:
- Due to the complexity of the overall task of NLP (understanding language), it has been split up into innumerable sub-tasks.
- Typically have rather small labeled data sets, but closely related tasks.

- We've unfortunately not seen huge boosts (unlike e.g. computer vision).
- Exception: Transfer of pre-trained embeddings or LMs for input representations.
- TL/MTL still a very active area of research.
- Lots of research currently on the representational transferability of different encoding architectures and objectives.