Multi-Task Learning

A lecture for Statistical Methods for Natural Language Semantics

by

Verna Dankers

April 29, 2019

Contents

Multi-Task Learning = MTL
Single-Task Learning = STL

Main task vs. Auxiliary task

Motivation

1. Improve the performance of specific tasks by introducing inductive biases.

   E.g. POS tags correlate with dependencies, so improve dependency parsing using POS labelling.

2. Move towards a unified natural language processing architecture.

   E.g. Frame any NLP task as question answering for one in DecaNLP model (McCann et al., 2018).

MTL Mechanisms

How can MTL with tasks A and B improve performance on A (Caruana, 1997)?

1. Data amplification

   Introducing B means adding data and introducing regularisation.

2. Representation bias
3. Attribute selection
4. Eavesdropping

MTL Mechanisms

How can MTL with tasks A and B improve performance on A (Caruana, 1997)?

1. Data amplification
2. Representation bias

   Introducing B may lead to finding different local minima, i.e. lead to exploring different representations in the hypothesis space.

3. Attribute selection
4. Eavesdropping

MTL Mechanisms

How can MTL with tasks A and B improve performance on A (Caruana, 1997)?

1. Data amplification
2. Representation bias
3. Attribute selection

   Task B can help the model focus its attention on the input features that are most relevant.

4. Eavesdropping

MTL Mechanisms

How can MTL with tasks A and B improve performance on A (Caruana, 1997)?

1. Data amplification
2. Representation bias
3. Attribute selection
4. Eavesdropping

Features useful for both A and B may be easier to learn on task B.

MTL Mechanisms

The mechanisms at work:

1. Unsupervised tasks such as language modelling, autoencoding and SkipThought can improve sequence-to-sequence (Luong et al., 2015) and sequence labelling tasks (Rei, 2017).
2. Training attention modules with human eye movement data can improve sequence classification (Barret et al., 2018).

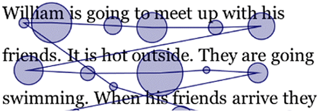

Figure 1: Example of eye movement behaviour through movements and fixations.

Architectures

Hard parameter sharing – 'Vanilla'

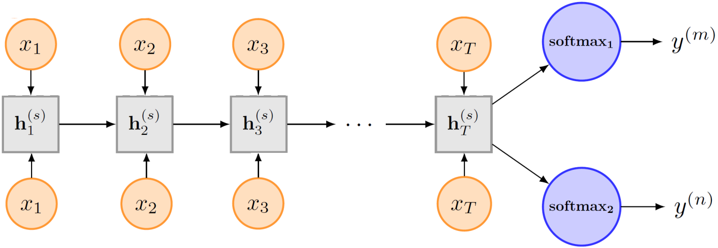

Figure 2: Shared recurrent layer, followed by task-specific classification layers (image adapted from Liu et al. (2016)).

Architectures

Hard parameter sharing – Share encoder *and* decoder

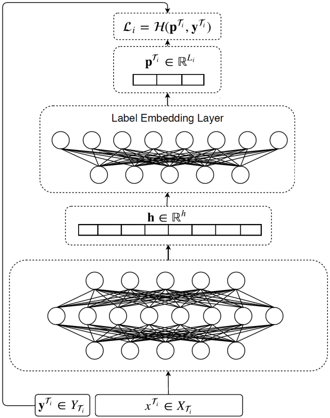

Figure 3: Joint label embedding space of Augenstein et al. (2018).

Architectures

Hard parameter sharing – Share encoder *and* decoder

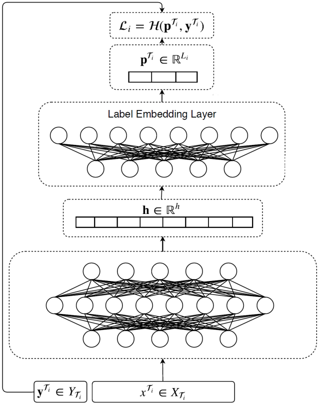

Figure 3: Joint label embedding space of Augenstein et al. (2018).

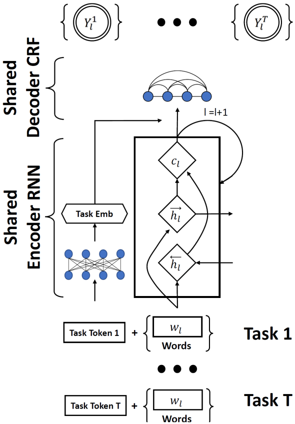

Figure 4: Task embeddings of Changpinyo et al. (2018).

Architectures

Hard parameter sharing – Hierarchical setup

- Predicting two different tasks can be more accurate when performed in different layers than in the same layer.

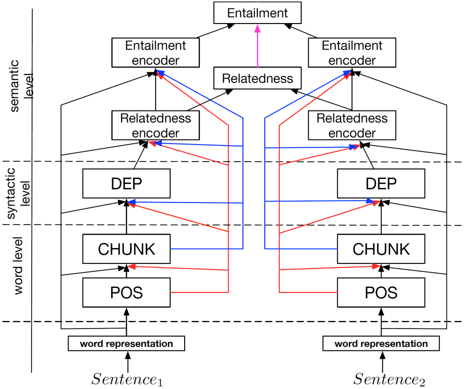

Figure 5: Joint-many model of Hashimoto et al. (2017).

Architectures

Soft parameter sharing – Gated network

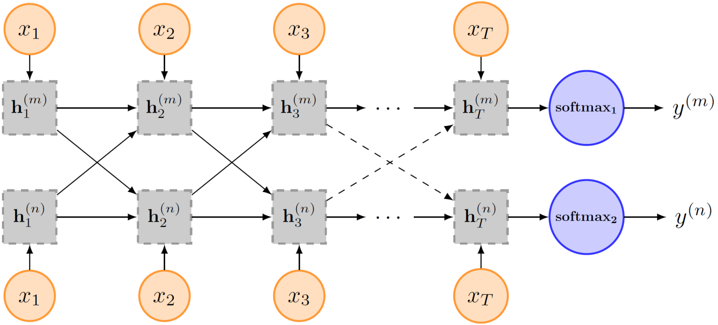

Figure 6: Soft parameter sharing setup, networks connected through gates (Liu et al, 2016).

Architectures

Soft parameter sharing – Shared-private network

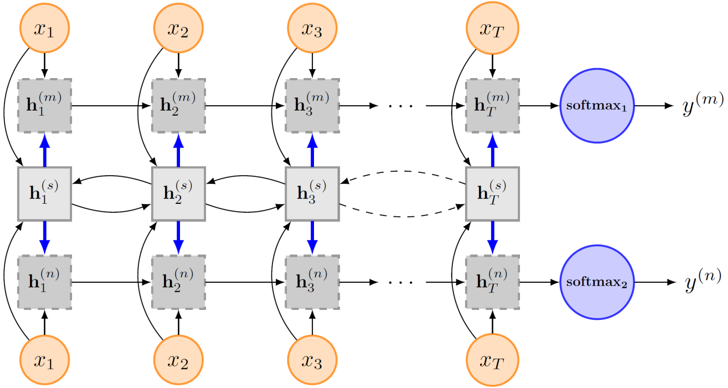

Figure 7: Tasks connected indirectly through shared network (Liu et al, 2016).

Architectures

Soft parameter sharing – Cross-stitch Network

- Presented in multi-task computer vision architecture (Misra et al., 2016);
- Units linearly combine hidden states from two tasks.

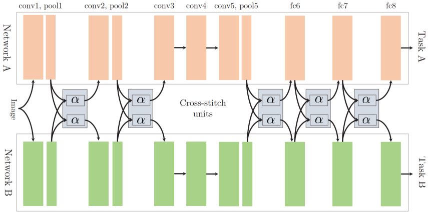

Figure 8: Cross-stitch network of (Misra et al., 2016).

$$\begin{bmatrix} \tilde{h}_A \\ \tilde{h}_B \end{bmatrix} = \begin{bmatrix} \alpha_{AA} & \alpha_{AB} \\ \alpha_{BA} & \alpha_{BB} \end{bmatrix} \begin{bmatrix} h_A^\top & h_B^\top \end{bmatrix}$$

Architectures

Soft parameter sharing – Sluice Network

- Cross-Stitch Units with more α parameters ($4 \rightarrow 16$);
- Orthogonality constraint on subspaces in recurrent layer;
- Skip-connections with corresponding β parameters.

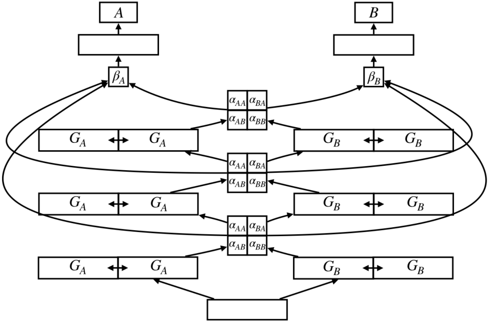

Figure 9: Sluice network presented by Ruder et al. (2019).

Training Strategies

1. Consecutive training (Hashimoto et al., 2017)
   - In one epoch, iterate over the datasets in order of complexity;
   - Introduce successive regularisation to avoid forgetting.

$$J_5(\theta_{\mathrm{ent}}) = \overbrace{-\sum_{(s,s')} \log p(y^{(5)}_{(s,s')} = \alpha | h^{(5)}_s, h^{(5)}_{s'})}^{\text{task objective}}$$
$$+ \underbrace{\lambda \|W_{\mathrm{ent}}\|^2}_{\text{task weight decay}} + \underbrace{\delta \|\theta_{\mathrm{rel}} - \theta'_{\mathrm{rel}}\|^2}_{\text{successive regularisation}},$$

Figure 10: Training objective for the entailment layer of the Joint-Many model of Hashimoto et al. (2017).

Training Strategies

2. Curriculum learning (Bengio et al., 2009)
   - When training machine learning models, start with easier subtasks and gradually increase the difficulty level of the tasks;
   - Motivation from humans and animals who learn better when trained with a curriculum-like strategy.
3. Anti-curriculum learning
   - Despite the motivation curriculum learning does not always work best;
   - McCann et al. (2018) start training using only 'difficult' tasks (e.g. NLI) in phase one and add 'easy' tasks in phase two (e.g. sentiment analysis) with DecaNLP.

Training Strategies

4. Randomised training
   - Uniform Task Selection (Søgaard and Goldberg 2016);
   - Proportional Task Selection: according to dataset size (Sahn et al., 2018).
5. Periodic task alternations
   - Dong et al. (2015) use periodic task alternations with equal training ratios for every task.
6. Alternative training algorithms
   - E.g. recently proposed teaching distillation from teacher (STL architectures) to student (MTL architecture) (Keskar et al., 2019) ← *For inspiration, not generally recommended strategy.*

General Guidelines

Selection taken from Ruder (2017):

1. Related tasks
   - Classical choice: choose a strongly related task as auxiliary task;
   - E.g. auxiliary task of sentiment analysis with main task emotion prediction (Yu and Jiang, 2016);
   - This is the guideline underlying most of your research project choices.
2. Representation learning
   - Autoencoding;
   - Language Modelling

General Guidelines

3. Eavesdropping
   - Learn features that are harder to learn using the main task;
   - E.g. Cheng et al., (2015) perform name error detection (main task) and include sentence level name detection (auxiliary task).

Reference	my name is captain <u>rodriguez</u>
Hypothesis	my name is captain <u>road radios</u>

Table 1: Example from the name error detection task.

4. Attribute selection
   - Learn what to focus on in the input, such as attention learning discussed in the Introduction.

General Guidelines

5. Adversarial training objective
   - Remember the shared-private model;
   - Nothing prevents interference of shared and private information;
   - $\rightarrow$ Introduce adversarial loss that prevents the shared space from performing the individual tasks.

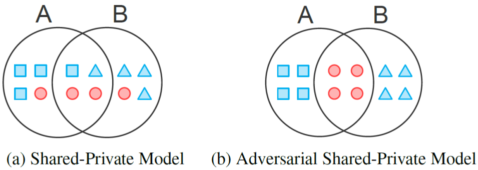

Figure 11: Illustration of the effect of introducing the adversarial loss in the shared-private network (Liu et al., 2017).

Task Relations Study (1)

- Bingel and Søgaard (2017) perform a systematic study of *when* and *why* MTL works for sequence labelling;
- Glove embeddings, hard parameter sharing bi-LSTM and task-specific output layers;
- Random selection training strategy.

Task Relations Study (1)

- Logical type tagging (CCG)
- Chunking (CHU)
- Sentence compression (COM)
- Semantic frames (FNT)
- POS tagging (POS)
- Hyperlink prediction (HYP)
- Keyphrase detection (KEY)
- Multi-word-expression detection (MWE)
- Super-sense tagging 1 (SEM)
- Super-sense tagging 2 (STR)

Task Relations Study (1)

- Logical type tagging (CCG)
- Chunking (CHU)
- Sentence compression (COM)
- Semantic frames (FNT)
- POS tagging (POS)
- Hyperlink prediction (HYP)
- Keyphrase detection (KEY)
- Multi-word-expression detection (MWE)
- Super-sense tagging 1 (SEM)
- Super-sense tagging 2 (STR)

	CCG	CHU	COM	FNT	POS	HYP	KEY	MWE	SEM	STR
CCG		1.4	0.45	0.58	1.8	0.24	0.3	0.45	1.4	0.84
CHU	-0.052		-0.15	-0.12	-0.45	-0.5	-0.22	-0.27	-0.099	-0.32
COM	-5	1.3		1.3	-1.4	-2.4	-4.8	0.82	-3	-0.63
FNT	-5.8	-1	-6.1		-9.4	-5.7	-3.6	-9.4	-3	-0.68
POS	4.9	2.9	1.9	0.9		-0.85	-0.26	1.3	3.4	2.9
HYP	12	4	-11	9.2	22		1.5	-7.7	23	8.1
KEY	5.7	3.2	-1	-0.43	-1.3	-2.6		-4.7	0.59	0.69
MWE	18	20	7.4	5.5	1.6	-3.8	-5.8		16	8.6
SEM	-5	-0.76	-1.2	-0.81	-0.85	-1.3	-0.83	-1.1		-1.7
STR	-1.7	1.5	-0.26	-0.72	0.037	-1.5	-1.4	-1.6	1.7	

Figure 12: Relative gains and losses (%) in F1 for including auxiliary tasks (columns) with main tasks (rows).

Task Relations Study (1)

- Logical type tagging (CCG)
- Chunking (CHU)
- Sentence compression (COM)
- Semantic frames (FNT)
- POS tagging (POS)
- Hyperlink prediction (HYP)
- Keyphrase detection (KEY)
- Multi-word-expression detection (MWE)
- Super-sense tagging 1 (SEM)
- Super-sense tagging 2 (STR)

magenta = benefit most,
brown = most beneficial

	CCG	CHU	COM	FNT	POS	HYP	KEY	MWE	SEM	STR
CCG		1.4	0.45	0.58	1.8	0.24	0.3	0.45	1.4	0.84
CHU	-0.052		-0.15	-0.12	-0.45	-0.5	-0.22	-0.27	-0.099	-0.32
COM	-5	1.3		1.3	-1.4	-2.4	-4.8	0.82	-3	-0.63
FNT	-5.8	-1	-6.1		-9.4	-5.7	-3.6	-9.4	-3	-0.68
POS	4.9	2.9	1.9	0.9		-0.85	-0.26	1.3	3.4	2.9
HYP	12	4	-11	9.2	22		1.5	-7.7	23	8.1
KEY	5.7	3.2	-1	-0.43	-1.3	-2.6		-4.7	0.59	0.69
MWE	18	20	7.4	5.5	1.6	-3.8	-5.8		16	8.6
SEM	-5	-0.76	-1.2	-0.81	-0.85	-1.3	-0.83	-1.1		-1.7
STR	-1.7	1.5	-0.26	-0.72	0.037	-1.5	-1.4	-1.6	1.7	

Figure 12: Relative gains and losses (%) in F1 for including auxiliary tasks (columns) with main tasks (rows).

Task Relations Study (1)

- Logical type tagging (CCG)
- Chunking (CHU)
- Sentence compression (COM)
- Semantic frames (FNT)
- POS tagging (POS)
- Hyperlink prediction (HYP)
- Keyphrase detection (KEY)
- Multi-word-expression detection (MWE)
- Super-sense tagging 1 (SEM)
- Super-sense tagging 2 (STR)

blue and red = symbiotic relations

	CCG	CHU	COM	FNT	POS	HYP	KEY	MWE	SEM	STR
CCG		1.4	0.45	0.58	1.8	0.24	0.3	0.45	1.4	0.84
CHU	-0.052		-0.15	-0.12	-0.45	-0.5	-0.22	-0.27	-0.099	-0.32
COM	-5	1.3		1.3	-1.4	-2.4	-4.8	0.82	-3	-0.63
FNT	-5.8	-1	-6.1		-9.4	-5.7	-3.6	-9.4	-3	-0.68
POS	4.9	2.9	1.9	0.9		-0.85	-0.26	1.3	3.4	2.9
HYP	12	4	-11	9.2	22		1.5	-7.7	23	8.1
KEY	5.7	3.2	-1	-0.43	-1.3	-2.6		-4.7	0.59	0.69
MWE	18	20	7.4	5.5	1.6	-3.8	-5.8		16	8.6
SEM	-5	-0.76	-1.2	-0.81	-0.85	-1.3	-0.83	-1.1		-1.7
STR	-1.7	1.5	-0.26	-0.72	0.037	-1.5	-1.4	-1.6	1.7	

Figure 12: Relative gains and losses (%) in F1 for including auxiliary tasks (columns) with main tasks (rows).

Task Relations Study (1)

- Using logistic regression, they try to predict MTL gains from dataset statistics (e.g. size, label distribution entropy) and STL model characteristics (e.g. loss curve values), and find good predictors:
    1. Multi-task gains are more likely for main tasks that quickly plateau with non-plateauing auxiliary tasks;
    2. Label entropy of the auxiliary task.
- But also bad ones:
    3. Contrary to earlier research: dataset sizes.

Task Relations Study (2)

- Changpinyo et al. (2018) move beyond pairwise comparisons;
- Extensive empirical studies on 11 sequence tagging tasks;
- Multiple architectures:
    1. Hard-parameter sharing with task-specific output layers;
    2. Hard-parameter sharing of all layers, but with task embeddings.
- Uniform selection training strategy.

Task Relations Study (2)

- POS tagging (UPOS, XPOS)
- Chunking (CHUNK)
- Named entity recognition (NER)
- Multi-word expression identification (MWE)
- Supersense tagging (SEM)
- Semantic trait tagging (SEMTR)
- Supersense tagging (SUPSENSE)
- Sentence compression (COM)
- Semantic frame prediction (FRAME)
- Hyperlink detection (HYP)

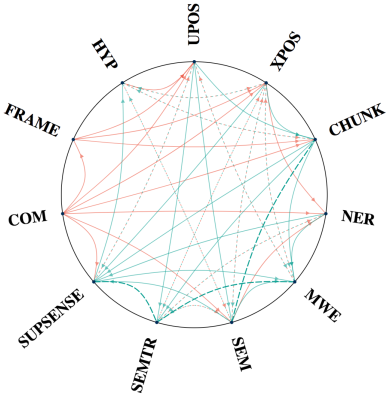

Figure 13: Pairwise MTL relations, green is beneficial, red is harming, dotted is asymmetric.

Task Relations Study (2)

- POS tagging (UPOS, XPOS)
- Chunking (CHUNK)
- Named entity recognition (NER)
- Multi-word expression identification (MWE)
- Supersense tagging (SEM)
- Semantic trait tagging (SEMTR)
- Supersense tagging (SUPSENSE)
- Sentence compression (COM)
- Semantic frame prediction (FRAME)
- Hyperlink detection (HYP)

magenta = always benefit

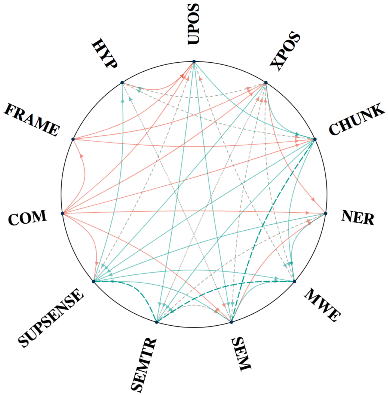

Figure 13: Pairwise MTL relations, green is beneficial, red is harming, dotted is asymmetric.

Task Relations Study (2)

- POS tagging (UPOS, XPOS)
- Chunking (CHUNK)
- Named entity recognition (NER)
- Multi-word expression identification (MWE)
- Supersense tagging (SEM)
- Semantic trait tagging (SEMTR)
- Supersense tagging (SUPSENSE)
- Sentence compression (COM)
- Semantic frame prediction (FRAME)
- Hyperlink detection (HYP)

blue = beneficial, red = harmful

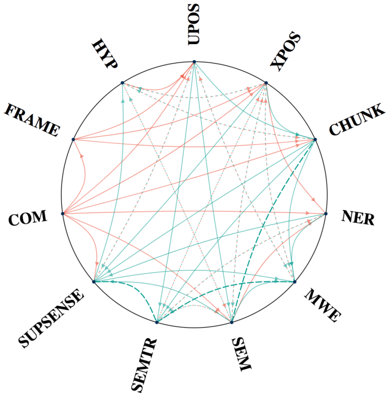

Figure 13: Pairwise MTL relations, green is beneficial, red is harming, dotted is asymmetric.

Task Relations Study (2)

1. STL vs. Oracle

   Oracle outperforms or is not worse than STL.

2. All/Oracle vs. Pairwise

   Oracle almost always better than Pairwise, All in half of the cases.

3. All vs. Oracle

   Generally Oracle is better than All.

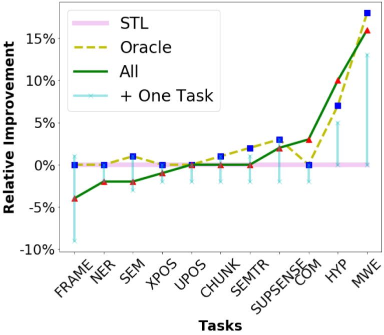

Figure 14: Summary of all results.

Task Relations Study (2)

1. Clusters of syntactic and semantic tasks (COM vs. HYP/MWE);
2. Tasks trained on the same data are not neighbours;
3. Label set entropy not indicative of distance in this space.

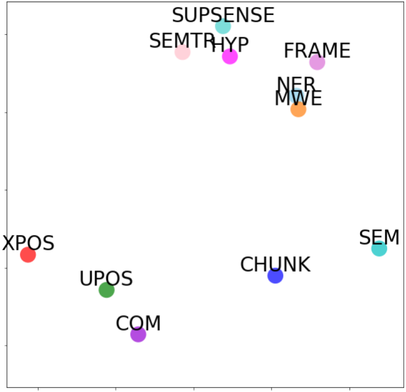

Figure 15: t-SNE visualisation of task embeddings.

Questions & Discussion

- We have discussed *why* you may want to use MTL and *how* MTL could provide performance gains;
- *which* architectures exist and *how* you can train them;
- *how* to choose the tasks to combine;
- *which* tasks go well together and *how* you can systematically research performance gains and losses.

[ARS18] Isabelle Augenstein, Sebastian Ruder, and Anders Søgaard. "Multi-Task Learning of Pairwise Sequence Classification Tasks over Disparate Label Spaces". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 2018, pp. 1896–1906.

[Bar+18] Maria Barrett et al. "Sequence classification with human attention". In: *Proceedings of the 22nd Conference on Computational Natural Language Learning*. 2018, pp. 302–312.

[Ben+09] Yoshua Bengio et al. "Curriculum learning". In: *Proceedings of the 26th annual international conference on machine learning*. ACM. 2009, pp. 41–48.

[BS17] Joachim Bingel and Anders Søgaard. "Identifying beneficial task relations for multi-task learning in deep neural networks". In: *arXiv preprint arXiv:1702.08303* (2017).

[Car97] Rich Caruana. "Multitask learning". In: *Machine learning* 28.1 (1997), pp. 41–75.

[CFO15] Hao Cheng, Hao Fang, and Mari Ostendorf. "Open-domain name error detection using a multi-task rnn". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 2015, pp. 737–746.

[CHS18] Soravit Changpinyo, Hexiang Hu, and Fei Sha. "Multi-Task Learning for Sequence Tagging: An Empirical Study". In: *Proceedings of the 27th International Conference on Computational Linguistics*. 2018, pp. 2965–2977.

[Don+15] Daxiang Dong et al. "Multi-task learning for multiple language translation". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Vol. 1. 2015, pp. 1723–1732.

[H+17] Kazuma Hashimoto, Yoshimasa Tsuruoka, Richard Socher, et al. "A Joint Many-Task Model: Growing a Neural Network for Multiple NLP Tasks". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2017, pp. 1923–1933.

[Kes+19] Nitish Shirish Keskar et al. "Unifying Question Answering and Text Classification via Span Extraction". In: *arXiv preprint arXiv:1904.09286* (2019).

[LQH16] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. "Recurrent neural network for text classification with multi-task learning". In: *arXiv preprint arXiv:1605.05101* (2016).

[LQH17] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. "Adversarial Multi-task Learning for Text Classification". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2017, pp. 1–10.

[Luo+15] Minh-Thang Luong et al. "Multi-task sequence to sequence learning". In: *arXiv preprint arXiv:1511.06114* (2015).

[McC+18] Bryan McCann et al. "The Natural Language Decathlon: Multitask Learning as Question Answering". In: *arXiv preprint arXiv:1806.08730* (2018).

[Mis+16] Ishan Misra et al. "Cross-stitch networks for multi-task learning". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 3994–4003.

[Rei17] Marek Rei. "Semi-supervised Multitask Learning for Sequence Labeling". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2017, pp. 2121–2130.

[Rud+19] Sebastian Ruder et al. "Latent Multi-task Architecture Learning". In: *Thirty-Third AAAI Conference on Artificial Intelligence*. 2019.

[Rud17] Sebastian Ruder. "An overview of multi-task learning in deep neural networks". In: *arXiv preprint arXiv:1706.05098* (2017).

[SG16] Anders Søgaard and Yoav Goldberg. "Deep multi-task learning with low level tasks supervised at lower layers". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Vol. 2. 2016, pp. 231–235.

[SWR18] Victor Sanh, Thomas Wolf, and Sebastian Ruder. "A hierarchical multi-task approach for learning embeddings from semantic tasks". In: *Thirty-Second AAAI Conference on Artificial Intelligence*. 2018.

[YJ16] Jianfei Yu and Jing Jiang. "Learning sentence embeddings with auxiliary tasks for cross-domain sentiment classification". In: *Proceedings of the 2016 conference on empirical methods in natural language processing*. 2016, pp. 236–246.