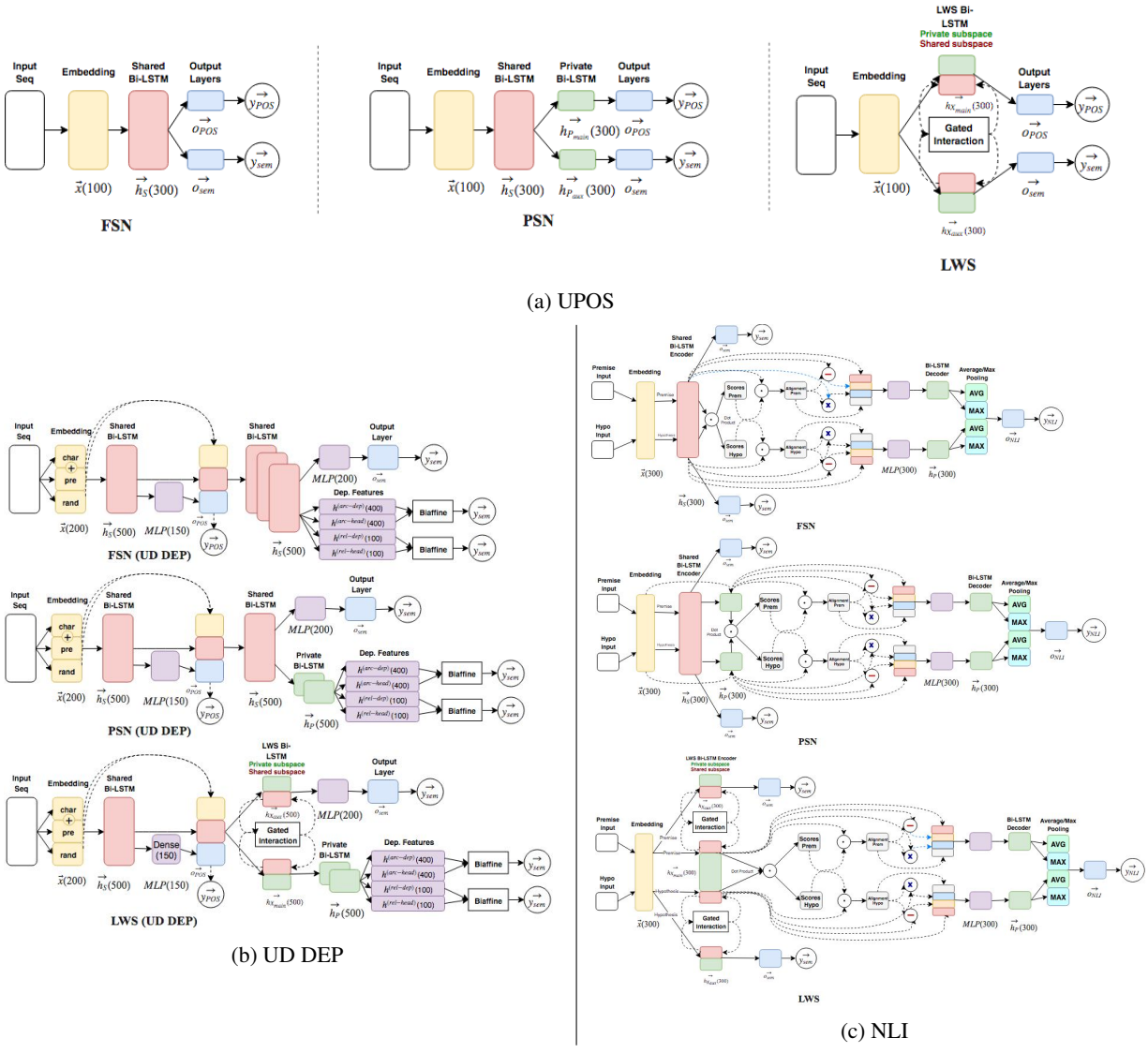
A MTL setting Diagrams, Preprocessing, and Hyperparameters

UPOS Tagging

Figure 3a shows the three MTL models used for UPOS. All hyperparameters were tuned with respect to loss on the English UD 2.0 UPOS validation set. We trained for 20 epochs with a batch size of 128 and optimized using Adam (Kingma and Ba, 2014) with a learning rate of 0.0001. We weight the auxiliary semantic tagging loss with λ = 0.1. The pre-trained word embeddings we used are GloVe embeddings (Pennington et al., 2014) of dimension 100 trained on 6 billion tokens of Wikipedia 2014 and Gigaword 5. We applied dropout and recurrent dropout with a probability of 0.3 to all bi-LSTMs.

UD DEP

Figure 3b shows the three MTL models for UD DEP. We use the gold tokenization. All hyperparameters were tuned with respect to loss on the English UD 2.0 UD validation set. We trained for 15 epochs with a batch size of 50 and optimized using Adam with a learning rate of $2e-3$. We weight the auxiliary semantic tagging loss with λ = 0.5. The pre-trained word embeddings we use are GloVe embeddings of dimension 100 trained on 6 billion tokens of Wikipedia 2014 and Gigaword 5. We applied dropout with a probability of 0.33 to all bi-LSTM, embedding layers, and non-output dense layers.



(a) UPOS

(b) UD DEP

(c) NLI

Figure 3: The three MTL settings for each task. Layers dimensions are displayed in brackets.

NLI

Figure 3c shows the three MTL models for NLI. All hyperparameters were tuned with respect to loss on the SNLI and SICK-E validation datasets (separately). For the SNLI experiments, we trained for 37 epochs with a batch size of 128. For the SICK-E experiments, we trained for 20 epochs with a batch size of 8. Note that the ESIM model was designed for the SNLI dataset, therefore performance is non-optimal for SICK-E. For both sets of experiments: we optimized using Adam with a learning rate of 0.00005; we weight the auxiliary semantic tagging loss with $\lambda = 0.1$; the pre-trained word embeddings we use are GloVe embeddings of dimension 300 trained on 840 billion tokens of Common Crawl; and we applied dropout and recurrent dropout with a probability of 0.3 to all bi-LSTM, and non-output dense layers.

B SNLI model output analysis

Table 2 shows demonstrative examples from the SNLI test set on which the *Learning What to Share* (LWS) model outperforms the single-task (ST) model. The examples cover all possible combinations of entailment classes. Table 3 explains the relevant part of the semantic tagset. Table 4 shows the per-label precision and recall scores.

Tag category	Semantic tag with examples
Anaphoric	DEF: definite; *the, lo^{IT}, der^{DE}* HAS: possessive pronoun; *my, her*
Attribute	COL: colour; *red, crimson, light_blue, chestnut_brown* QUC: concrete quantity; *two, six_million, twice* IST: intersective; *open, vegetarian, quickly* REL: relation; *in, on, 's, of, after*
Unnamed entity	CON: concept; *dog, person*
Logical	ALT: alternative & repetitions; *another, different, again* DIS: disjunction & exist. quantif.; *a, some, any, or*
Discourse	SUB: subordinate relations; *that, while, because*
Events	ENS: present simple; *we walk, he walks* EPS: past simple; *ate, went* EXG: untensed progessive; *is running* EXS: untensed simple; *to walk, is eaten, destruction*
Tense & aspect	NOW: present tense; *is skiing, do ski, has skied, now*

Table 3: The list of semantic tags found in Table 2.

Model	Label		
	Entailment	Contradiction	Neutral
FSN	80.64/93.23	91.64/83.63	83.97/77.63
ST	84.86/91.54	90.10/88.04	84.74/79.71
PSN	84.08/92.70	91.17/88.63	85.96/79.15
LWS	84.45/92.87	91.74/88.91	85.95/79.65

Table 4: Per-label precision (left) and recall (right) for all models.

Premise-hypothesis pairs	ST	LWS/GOLD
P: The^{DEF} gentleman^{CON} is^{NOW} speaking^{EXS} while^{SUB} the^{DEF} others^{ALT} are^{NOW} listening^{EXS} **H**: The^{DEF} man^{CON} is^{NOW} being^{EXS} given^{EXS} respect^{CON}	N	E
P: Men^{CON} wearing^{EXG} hats^{CON} walk^{EXS} on^{REL} the^{DEF} street^{CON} **H**: The^{DEF} men^{CON} having^{EXS} hats^{CON} on^{REL} their^{HAS} head^{CON}	C	E
P: Three^{QUC} men^{CON} in^{REL} orange^{IST} suits^{CON} are^{NOW} doing^{EXG} street^{CON} repairs^{CON} at^{REL} night^{CON} **H**: Three^{QUC} men^{CON} in^{REL} orange^{IST} suits^{CON} escaped^{EPS} from^{REL} prison^{CON}	N	C
P: A^{DIS} toddler^{CON} sits^{ENS} on^{REL} a^{DIS} stone^{CON} wall^{CON} surrounded^{EXS} by^{REL} fallen^{EXS} leaves^{CON} **H**: An^{DIS} child^{CON} is^{NOW} throwing^{EXG} stones^{CON} at^{REL} a^{DIS} leaf^{CON} wall^{CON}	E	C
P: An^{DIS} old^{IST} shoemaker^{CON} in^{REL} his^{HAS} factory^{CON} **H**: The^{DEF} shoemaker^{CON} is^{NOW} wealthy^{IST}	C	N
P: A^{DIS} kid^{CON} slides^{CON} down^{IST} a^{DIS} yellow^{COL} slide^{CON} into^{REL} a^{DIS} swimming^{CON} pool^{CON} **H**: The^{DEF} kid^{CON} is^{NOW} playing^{EXS} at^{REL} the^{DEF} waterpark^{CON}	E	N

Table 2: Examples of the entailment problems from SNLI which are incorrectly classified by the ST model but correctly classified by the LWS model. Automatically assigned semantic tags are in superscript.