A Derivation for Stochastic Pruning

To re-parametrize the discrete binary Bernoulli variable $m^l_{i,j} \sim B(\sigma(g^l_{i,j}))$, denote the approaximate differentiable variable as $\tilde{m}^l_{i,j} = \sigma(\frac{g^l_{i,j} + \log U - \log(1-U)}{\tau})$ where τ is a real-valued temperature value, we have the following deriavation holds for arbitrary $\epsilon \in (0, 0.5)$:

$$P(m^l_{i,j} = 1) - P(\tilde{m}^l_{i,j} \geq 1 - \epsilon) \leq (\frac{\tau}{4}) \log \frac{1}{\epsilon} \tag{1}$$

Specifically, when temperature τ approaches 0, $\tilde{m}^l_{i,j} = m^l_{i,j}$.

Lemma 1: $\sigma^{-1}(x) = \log \frac{x}{1-x}$.

Lemma 2: $\frac{\sigma(x) - \sigma(y)}{x - y} \leq \frac{1}{4}$.

Proof:

$$P(\tilde{m}^l_{i,j} \geq 1 - \epsilon) \tag{2}$$

$$= P(\sigma(\frac{g^l_{i,j} + \log U - \log(1-U)}{\tau}) \geq 1 - \epsilon) \tag{3}$$

$$= P(\frac{g^l_{i,j} + \log U - \log(1-U)}{\tau} \geq \log(\frac{1}{\epsilon} - 1)) \tag{4}$$

$$= P(g^l_{i,j} - \tau \log(\frac{1}{\epsilon} - 1) \geq \log(\frac{1}{U} - 1)) \tag{5}$$

$$= P(e^{g^l_{i,j} - \tau \log(\frac{1}{\epsilon} - 1)} \geq \frac{1}{U} - 1) \tag{6}$$

$$= P(U \geq \frac{1}{1 + e^{g^l_{i,j} - \tau \log(\frac{1}{\epsilon} - 1)}}) \tag{7}$$

$$= \sigma(g^l_{i,j} - \tau \log(\frac{1}{\epsilon} - 1)) \tag{8}$$

Then:

$$P(m^l_{i,j} = 1) - P(\tilde{m}^l_{i,j} \geq 1 - \epsilon) \tag{9}$$

$$= \sigma(g^l_{i,j}) - \sigma(g^l_{i,j} - \tau \log \frac{1}{\epsilon} - 1) \tag{10}$$

$$\leq \frac{\tau}{4} \log(\frac{1}{\epsilon} - 1) \tag{11}$$

$$\leq \frac{\tau}{4} \log \frac{1}{\epsilon} \tag{12}$$

The process for deriving $P(m^l_{i,j} = 0) - P(\tilde{m}^l_{i,j} \leq \epsilon) \leq (\frac{\tau}{4}) \log \frac{1}{\epsilon}$ can be analogously obtained. $\square$

B Knowledge Combination for Fine-tuning, Zero-shot and Triple Classification on Commonsense Reasoning

B.1 Notation for Knowledge Type

HasSubevent: 0
MadeOf: 1
HasPrerequisite: 2
MotivatedByGoal: 3
AtLocation: 4
CausesDesire: 5
IsA: 6
NotDesires: 7
Desires: 8
CapableOf: 9
PartOf: 10
HasA: 11
UsedFor: 12
ReceivesAction: 13
Causes: 14
HasProperty: 15

In the remainder of this section, we use $\cup$ to indicate mask union operation upon multiple commonsense knowledge types.

B.2 Fine-tuning

For fine-tuning on commonsense reasonging tasks, we only experiments with BERT-BASE due and perform hyper-parameter search only in terms of batch size in the range of $\{8, 16, 32\}$ and learning rate in the range of $\{3e^{-5}, 4e^{-5}, 5e^{-5}\}$ due to computational budget. We also adopt early stopping based on accuracy on the devlopment set. The combination achieving highest accuracy is shown in Table 1.

B.3 Zero-shot

In constrast with fine-tuning, zero-shot evaluation is deterministic as long as the model does not involve any stochastic module, thereby averting extensive hyperparameter tuning. Instead we perform exaustive search over knowledge combinations for each pretrained language model with number of knowledge types in $\{3, 4, 5\}$. The ConceptNet-grounded knowledge type combination achieving highest accuracy is listed in Table 2.

B.4 Triple Classification

In analogy with zero-shot evaluation, here we show the optimal knowledge type combination of each

Task	RTE	COPA	CSQA	SWAG	HellaSWAG	aNLI	CosmosQA
BERT	$0 \cup 6 \cup 14$	$5 \cup 8 \cup 14$	$3 \cup 4 \cup 8 \cup 12 \cup 14$	$1 \cup 6 \cup 10 \cup 11$	$0 \cup 3 \cup 5 \cup 8 \cup 14$	$0 \cup 3 \cup 5 \cup 8 \cup 14$	$0 \cup 3 \cup 5 \cup 8 \cup 14$

Table 1: Optimal fine-tuning knowledge type combination for BERT-BASE on commonsense reasoning tasks.

Task	COPA (Dev.)	CSQA	CA	WSC	SM	ARCT1	ARCT2
DistilBERT	$1 \cup 6 \cup 14$	$2 \cup 3 \cup 13$	$0 \cup 1 \cup 7 \cup 9$	$6 \cup 7 \cup 10$	$2 \cup 8 \cup 13$	$2 \cup 3 \cup 14$	$1 \cup 2 \cup 7$
BERT	$4 \cup 11 \cup 15$	$1 \cup 2 \cup 15$	$6 \cup 8 \cup 12$	$2 \cup 9 \cup 14$	$6 \cup 12 \cup 15$	$1 \cup 9 \cup 10$	$1 \cup 5 \cup 8$
RoBERTa	$2 \cup 3 \cup 8$	$0 \cup 2 \cup 5$	$0 \cup 1 \cup 8$	$1 \cup 2 \cup 4 \cup 5 \cup 11$	$8 \cup 11 \cup 12$	$2 \cup 5 \cup 11 \cup 13$	$0 \cup 8 \cup 11 \cup 13$
MPNet	$1 \cup 6 \cup 8 \cup 10$	$6 \cup 12 \cup 13$	$2 \cup 3 \cup 10$	$1 \cup 3 \cup 4 \cup 9$	$6 \cup 10 \cup 13 \cup 15$	$2 \cup 5 \cup 6 \cup 11$	$5 \cup 6 \cup 7 \cup 11$

Table 2: Optimal zero-shot knowledge type combination for each PLM on each commonsense reasoning tasks.

Model	P@1	P@2	P@3	Sparsity	$l_b - l_t$	# Param.
BERT-LARGE w/o pruning	15.1	20.9	24.6	0%	-	336M
BERT-LARGE w/ stochastic pruning	22.1	30.1	35.4	∼30%	17-24	336M
BERT-LARGE w/ deterministic pruning	69.2	74.1	76.3	∼50%	17-24	284M

Table 3: Macro-averaged precision metrics of BERT-LARGE on the ConceptNet subset of LAMA.

PLM for triple classification task on ConceptNet-100K [1].

DistilBERT-BASE: $3 \cup 4 \cup 12$
BERT-BASE: $9 \cup 13 \cup 14$
RoBERTa-BASE: $0 \cup 4 \cup 9 \cup 13$
MPNet-BASE: $1 \cup 4 \cup 9$

C Additional Pruning Results

C.1 BERT-LARGE

We also apply our pruning procedure upon BERT-LARGE, the rank-based metrics on LAMA [2] is shown in Table 3.

D Extracted Commonsense Triples

Applying the pruned DistilBERT-BASE model to predict missing objects for triples in ConceptNet-100K test set, we obtain commonsense triples deemed to be novel by three human annotators with Flessi's Kappa score κ of 0.65. We further filtered out triples that are included in the training or development set of ConceptNet-100K. Here we show some representative cases categorized by their relations:

CapableOf:
$(computer, crash)$, $(computer, communicate)$
IsA:
$(sex, relationship)$, $(submarine, weapon)$, $(submarine, vessel)$
AtLocation:
$(knife, war)$, $(knife, dinner)$, $(crab, dinner)$
UsedFor:
$(stage, fun)$, $(stage, performance)$, $(literature, education)$, $(literature, research)$
HasA:
$(book, index)$, $(book, information)$
HasProperty:
$(music, loud)$

Future work involves using seed triples beyond ConceptNet-100K dataset, e.g., the whole ConceptNet knowledge graph , and mining more novel and plausible commonsense knowledge.

[1] https://ttic.uchicago.edu/~kgimpel/commonsense.html
[2] https://github.com/facebookresearch/LAMA