UniDrop: A Simple yet Effective Technique to Improve Transformer without Extra Cost

Zhen Wu[1], Lijun Wu[2], Qi Meng[2], Yingce Xia[2], Shufang Xie[2], Tao Qin[2], Xinyu Dai[1], Tie-Yan Liu[2]
[1]National Key Laboratory for Novel Software Technology, Nanjing University
[2]Microsoft Research Asia
wuz@smail.nju.edu.cn, daixinyu@nju.edu.cn
{Lijun.Wu, meq, yingce.xia, shufxi, taoqin, tyliu}@microsoft.com

Background & Motivation

- Background
  - Great success of Transformer in NLP tasks.
  - The over-parameterization of Transformer.
  - Dropout is a popular regularization method.
- Goal
  - Achieve stronger Transformer or even SOTA results with various dropout techniques only.
- Advantages
  - Free of extra model architecture design, save computational costs.
  - Free of knowledge enhancement, don't require extra resources.

Proposed UniDrop

- UniDrop: unites three different-level dropout techniques from fine-grain to coarse-grain into Transformer models
  - **Feature dropout (FD):** conventional dropout (Srivastava et al., 2014), applied on hidden representations of networks.
  - **Structure dropout (SD):** randomly drops some entire substructures or components from the whole model.
  - **Data dropout(DD):** randomly drops out some tokens in an input sequence.
- Feature Dropout
  - **FD-1** (attention dropout): applied to the attention weight A, $A = QK^T$.
  - **FD-2** (activation dropout): applied after the activation function between the two linear transformations of FFN sub-layer.
  - **FD-3** (query, key, value dropout): adds dropout to query Q, key K, and value V before calculating attention.
  - **FD-4** (output dropout): applies dropout to the output features before linear transformation for softmax classification.

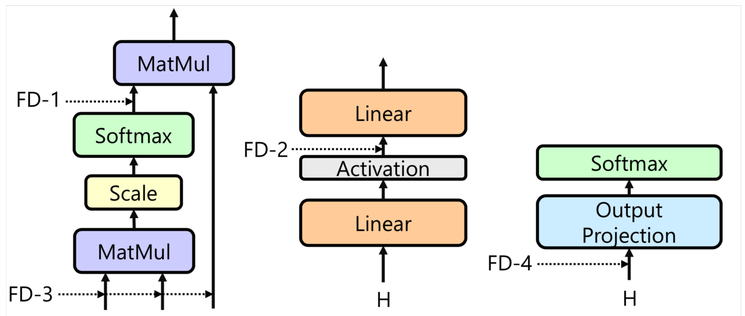

Figure 1: Structure and overview of feature dropout

- Structure Dropout
  - Adopt LayerDrop (Fan et al., 2020) as the structure dropout, which drops some entire layers at training time and directly reduces the Transformer model size.
- Data Dropout
  - Direct data dropout brings the risk of losing high-quality training samples.
  - Propose a two-stage data dropout strategy:
    - Given a sequence, with probability p_k, we keep the original sequence and do not apply data dropout. If data dropout is applied, for each token, with another probability p, we will drop the token.
- UniDrop Integration
  - Theoretically demonstrate that the above three dropout techniques can regularize Transformer from different aspects.
  - They can work together to prevent Transformer from overfitting.

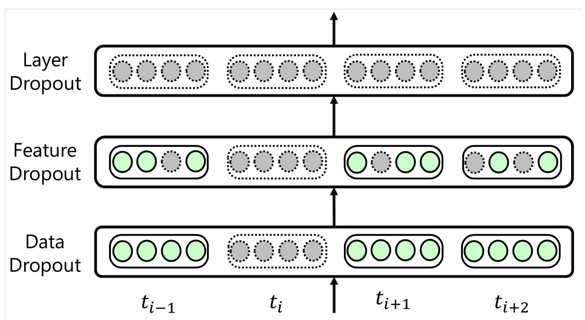

Figure 2: UniDrop illustration. The gray positions denote applying the corresponding dropout.

Experiments

- Neural Machine Translation
  - Metric: BLEU score
  - Datasets: IWSLT14 translation datasets

Table 1: Statistics IWSLT14 datasets

Datasets	Train	Dev	Test
En↔De	160k	7k	7k
En↔Ro	180k	4.7k	1.1k
En↔Nl	170k	4.5k	1.1k
En↔Pt-br	175k	4.5k	1.2k

Table 2: Comparison with existing works on IWSLT- 2014 De→En translation task.

Approaches	BLEU
Adversarial MLE (Wang et al., 2019b)	35.18
DynamicConv (Wu et al., 2019)	35.20
Macaron (Lu et al., 2019)	35.40
IOT (Zhu et al., 2021)	35.62
Soft Contextual Data Aug (Gao et al., 2019)	35.78
BERT-fused NMT (Zhu et al., 2020)	36.11
MAT (Fan et al., 2020b)	36.22
MixReps+co-teaching (Wu et al., 2020)	36.41
Transformer	34.84
+UniDrop	**36.88**

Table 3: Machine translation results

	En→De	De→En	En→Ro	Ro→En	En→Nl	Nl→En	Nn→Pt-br	Pt-br→En	Avg.	△
Transformer	28.67	34.84	24.74	32.14	29.64	33.28	39.08	43.63	33.25	-
+FD	29.61	36.08	25.45	33.12	30.37	34.50	40.10	44.74	34.24	+0.99
+SD	29.03	35.09	25.03	32.69	29.97	33.94	39.78	44.02	33.69	+0.44
+DD	28.83	35.26	24.98	32.76	29.72	34.00	39.50	43.71	33.59	+0.34
+UniDrop	**29.99**	**36.88**	**25.77**	**33.49**	**31.01**	**34.80**	**40.62**	**45.62**	**34.77**	+1.52
w/o FD	29.24	35.68	25.18	33.17	30.16	33.90	39.97	44.81	34.01	+0.76
w/o SD	29.92	36.70	25.59	33.26	30.55	34.75	40.45	45.60	34.60	+1.35
w/o DD	29.76	36.38	25.44	33.26	30.86	34.55	40.37	45.27	34.49	+1.24

- Text Classification
  - Metric: Accuracy
  - Datasets: GLUE tasks and Typical text classification datasets

Table 4: Statistics of text classification datasets

Datasets	Classes	Train	Dev
MNLI	3	393k	20k
QNLI	2	105k	5.5k
SST-2	2	67k	0.9k
MRPC	2	3.7k	0.4k
Datasets	**Classes**	**Train**	**Test**
IMDB	2	25k	25k
Yelp	5	650k	50k
AG's News	4	120k	76k
TREC	6	5.4k	0.5k

Table 5: Accuracy on GLUE tasks (dev set)

	MNLI	QNLI	SST-2	MRPC
BiLSTM+Attn, CoVe	67.9	72.5	89.2	72.8
BiLSTM+Attn, ELMo	72.4	75.2	91.5	71.1
$\text{BERT}_{\text{BASE}}$	84.4	88.4	92.9	86.7
$\text{BERT}_{\text{LARGE}}$	86.6	92.3	93.2	88.0
$\text{RoBERTa}_{\text{BASE}}$	87.1	92.7	94.7	89.0
+UniDrop	**87.8**	**93.2**	**95.5**	**90.4**

Table 6: Accuracy on the typical text classification datasets

	IMDB	Yelp	AG	TREC
Char-level CNN	-	62.05	90.49	-
VDCNN	-	64.72	91.33	-
DPCNN	-	69.42	93.13	-
ULMFiT	95.40	-	94.99	96.40
$\text{BERT}_{\text{BASE}}$	94.60	69.94	94.75	97.20
$\text{RoBERTa}_{\text{BASE}}$	95.7	70.9	95.1	97.6
+UniDrop	**96.0**	**71.4**	**95.5**	**98.0**

Analysis

- Overfitting

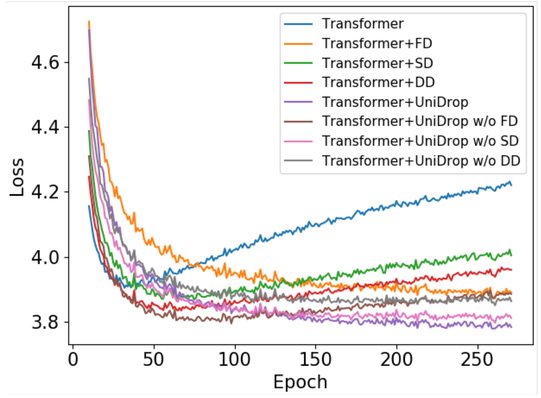

Figure 3: The dev loss of different models on IWSLT14 De→En translation task.

- Ablation Study

	De→En	En→De	Ro→En
Transformer	34.84	28.67	32.14
+UniDrop	36.88	29.99	33.49
w/o FD-1	36.72	29.84	33.33
w/o FD-2	36.57	29.76	33.28
w/o FD-3	36.59	29.83	33.31
w/o FD-4	36.65	29.59	33.24
w/o 2-stage DD	36.61	29.78	33.12

Table 7: Ablation study of data dropout and different feature dropout.

- Effects of Different Dropout Rates

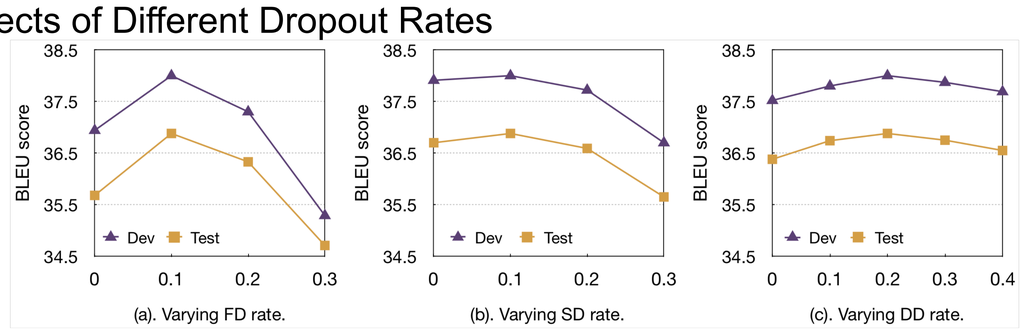

Figure 4: BLEU scores of Transformer+UniDrop on IWSLT14 De→En translation task when varying the rates on the setting (FD=0.1, SD=0.1, DD=0.2) .