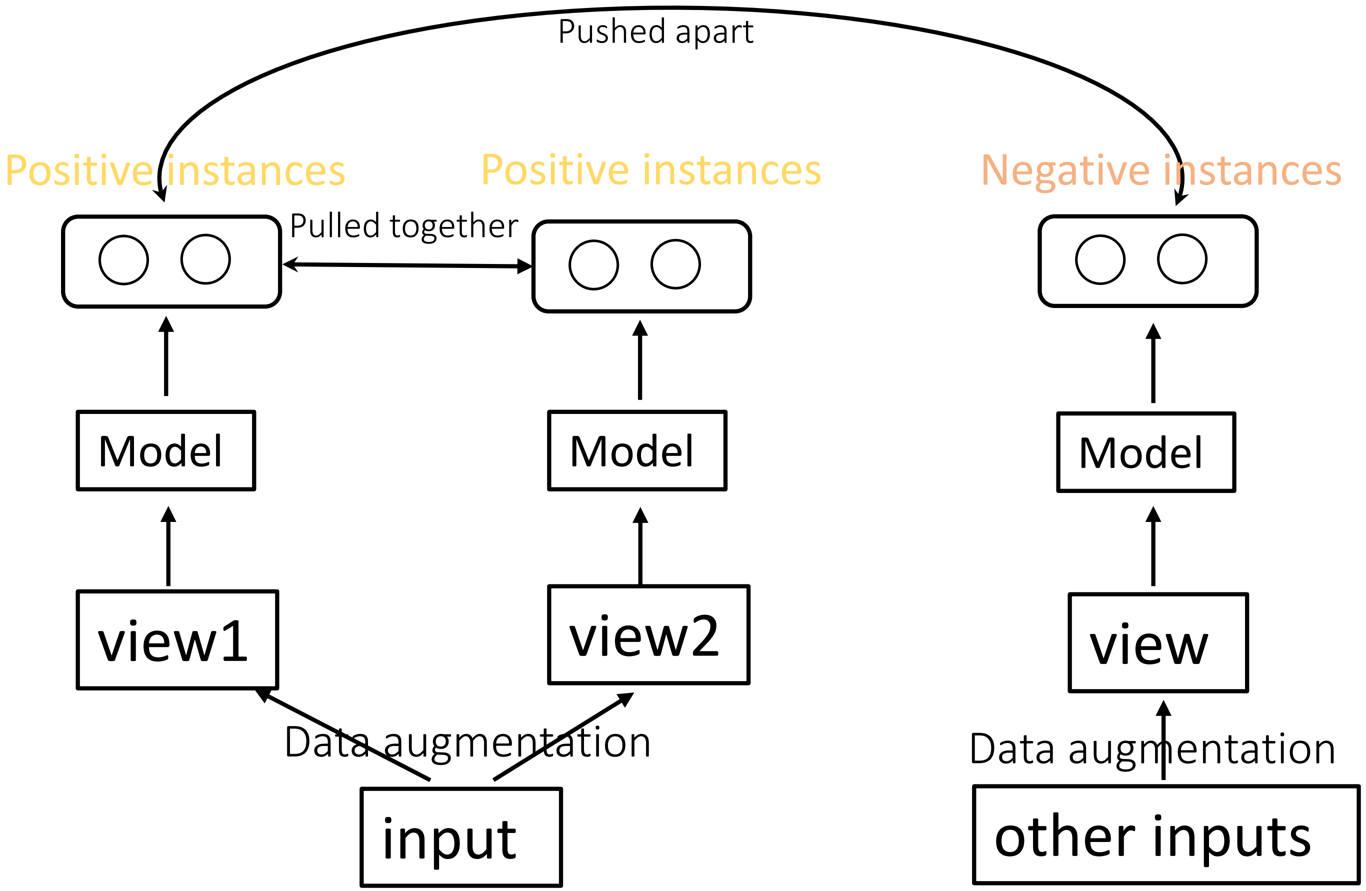
An Information Minimization Based Contrastive Learning Model for Unsupervised Sentence Embeddings Learning

Shaobin Chen[1], Jie Zhou[2], Yuling Sun[1], Liang He[1]

[1]East China Normal University [2]Fudan University

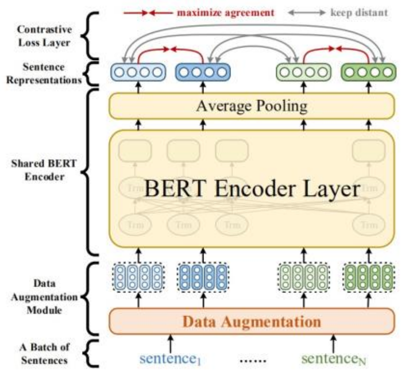
Code: https://github.com/Bin199/InforMin-CL

Introduction of Contrastive Learning

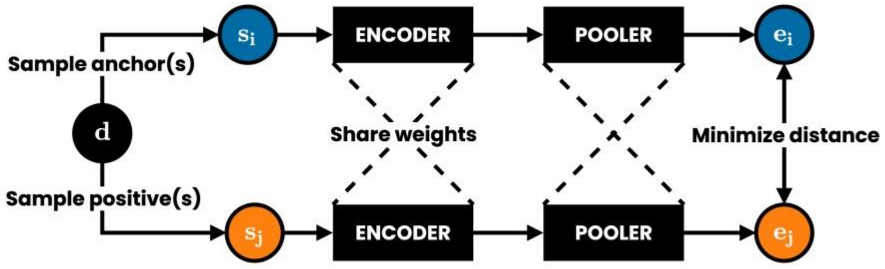


Previous Contrastive Learning Based Methods

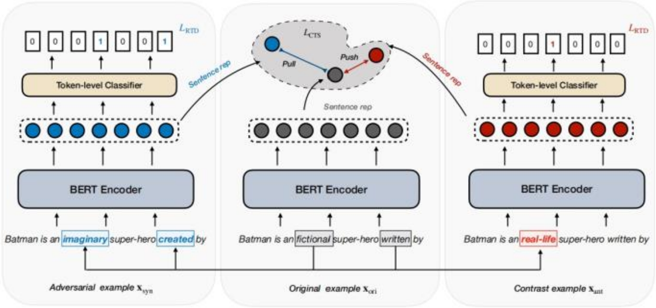
ConSERT



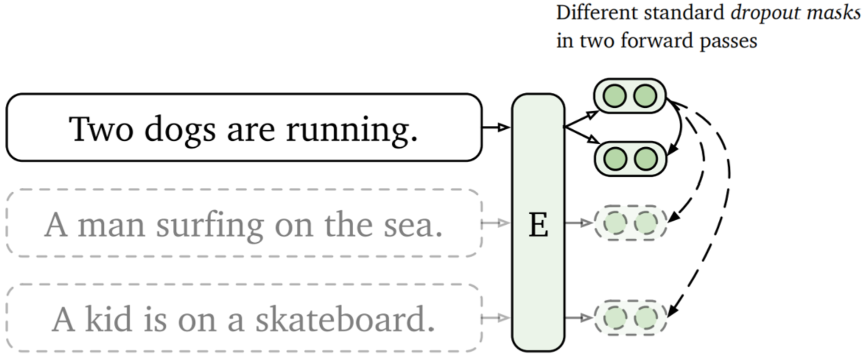
DeCLUTER



CLINE



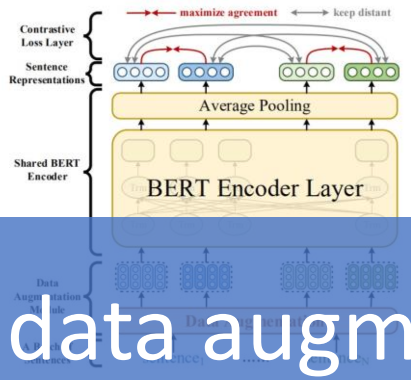
SimCSE



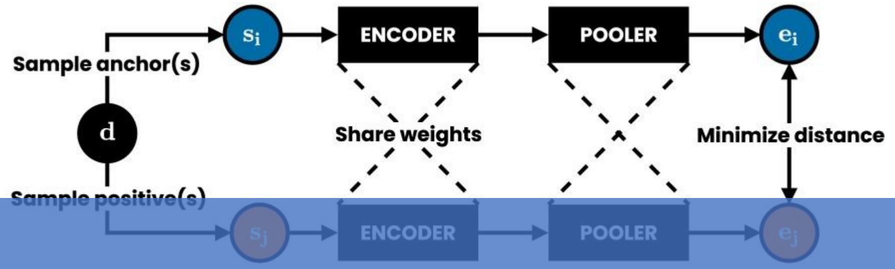
[*ConSERT: A Contrastive Framework for Self-Supervised Sentence Representation Transfer. Yuanmeng Yan et al. ACL 2021.*]
[*DeCLUTR: Deep Contrastive Learning for Unsupervised Textual Representations. John Giorgi et al. ACL 2021.*]
[*CLINE: Contrastive Learning with Semantic Negative Examples for Natural Language Understanding. Dong Wang et al. ACL 2021.*]
[*SimCSE: Simple Contrastive Learning of Sentence Embeddings. Tianyu Gao et al. EMNLP 2021.*]

Previous Contrastive Learning Based Methods

ConSERT



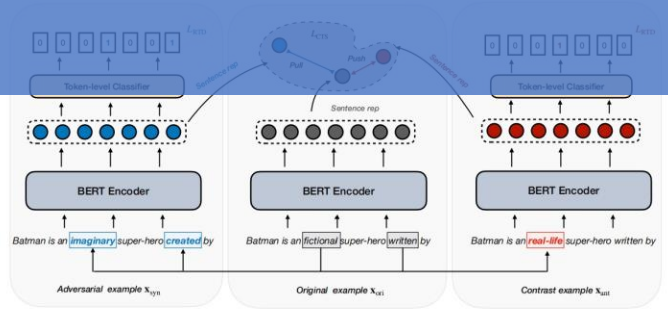
DeCLUTER



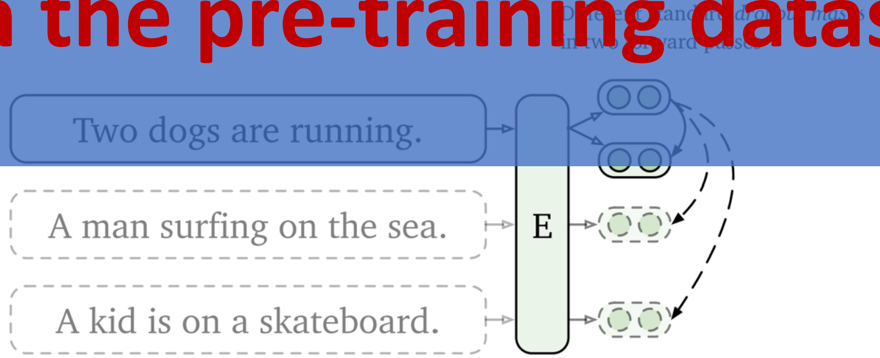
Focus on data augmentation skills, and model architecture

Ignore the redundant information in the pre-training datasets

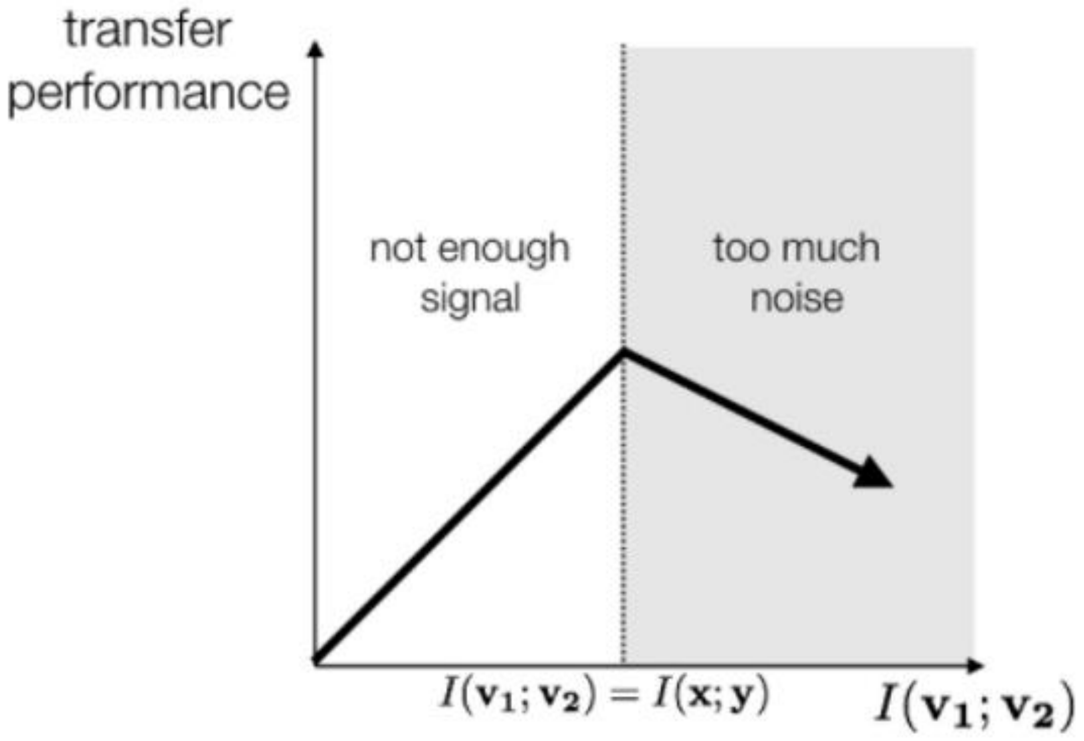
CLINE



SimCSE



Redundant Information



Redundant information leads to a drop in transfer performance!!!

[*What makes good views for contrastive learning? Yonglong Tian et al. NeurIPS 2020*.]

Redundant Information

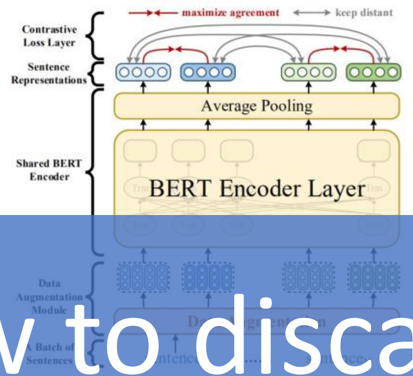
We care about two types of redundant information, stop words and **the style of the sentence** (e.g., restatement, capitalization, and hyphen.)

Original	Where is the party, it sounds great.
Stop words	Where is the party, it sounds great.
Restatement	The party sounds great, where is it.
Capitalization	Where Is The Party, It Sounds Great.
Hyphen	Where-is-the-party, it-sounds-great.

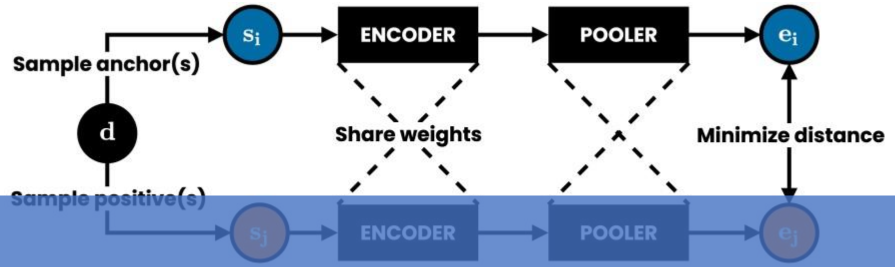
Less studied in previous work.

Question

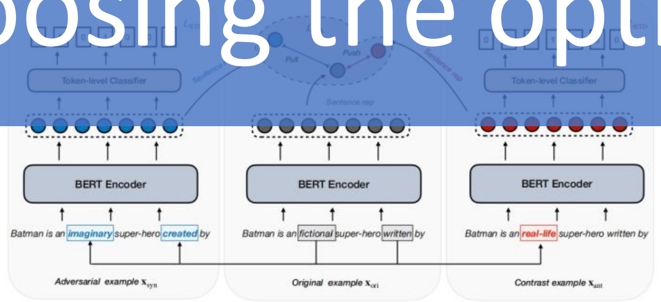
ConSERT

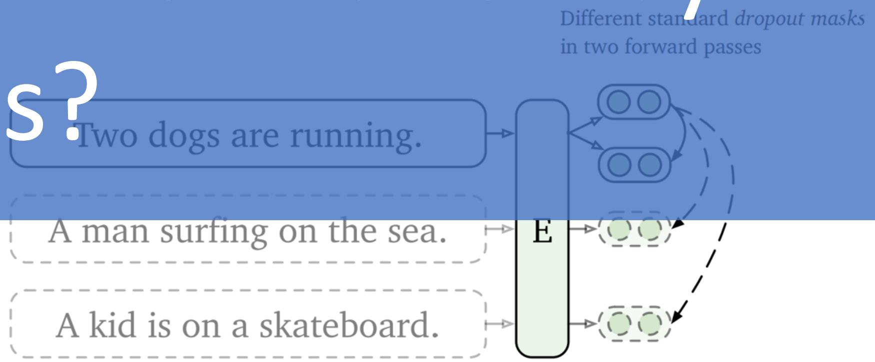


DeCLUTER



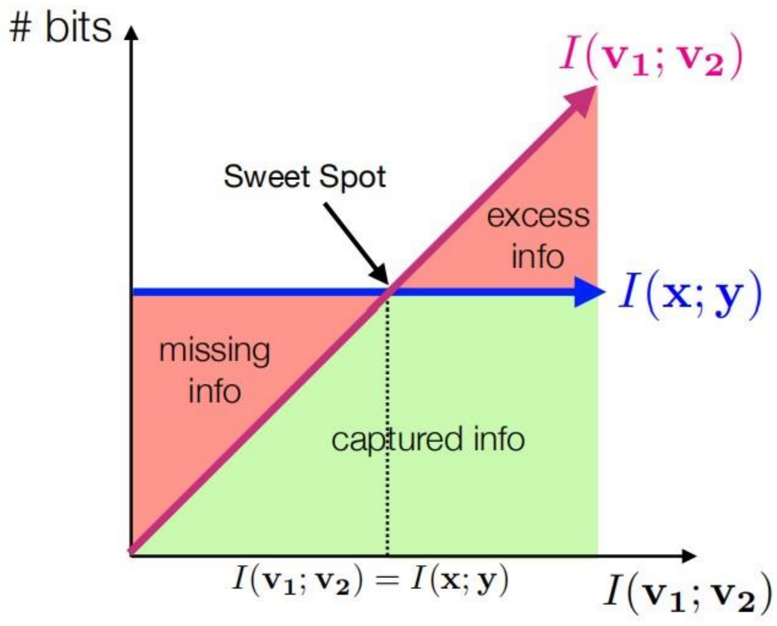
How to discard redundant information by choosing the optimal views?



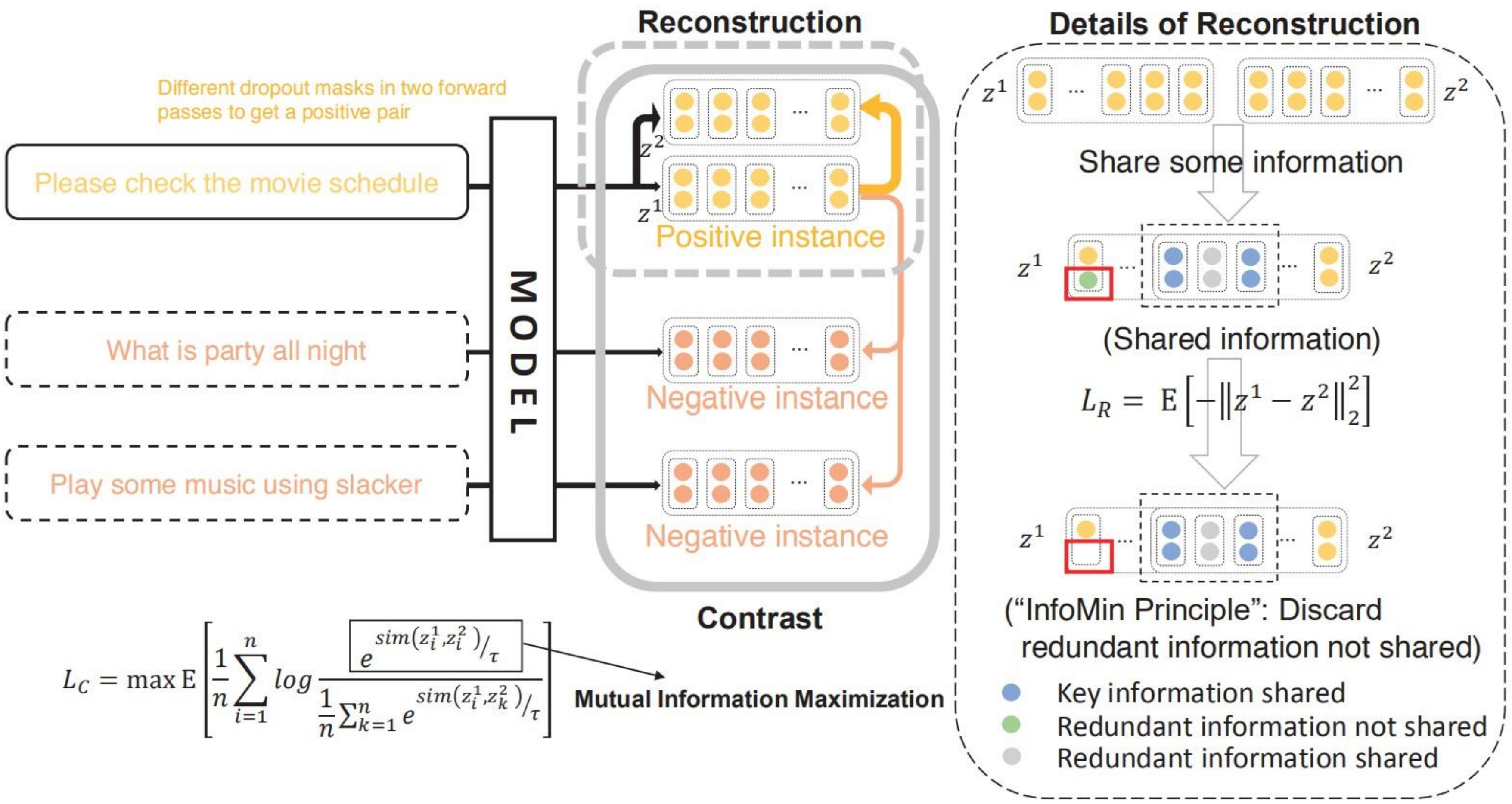


Solution

Draw inspiration from
<u>Information minimization principle</u>: A good set of views share the minimal information necessary to perform well at the downstream task.



Architecture of Proposed Model: InforMin-CL



Total Loss function L:

$$\mathcal{L} = \mathcal{L}_C + \lambda * \mathcal{L}_R$$

Contrast Keeps Almost All the Key Information

<u>Theorem 1:</u> The supervised learned representations contain all the key information in the input $I(X;T)$. The self-supervised representations contain all the key information in the input with a potential loss.

$$
\begin{aligned}
I(X;T) = I(Z^{sup};T) &= I(Z^{sup_{min}};T) \\
&\geq I\left(Z^{ssl};T\right) \\
&\geq I\left(Z^{ssl_{min}};T\right) \\
&\geq I(X;T) - \varepsilon
\end{aligned}
$$

X: input
Z: instance
S: self-supervised signal
T: key information
I: mutual information
H: information entropy

$$Z^{\text{sup}} = \arg\max_{Z} I(Z;T)$$

$$Z^{sup_{\min}} = \arg\min_{Z} H(Z|T)$$

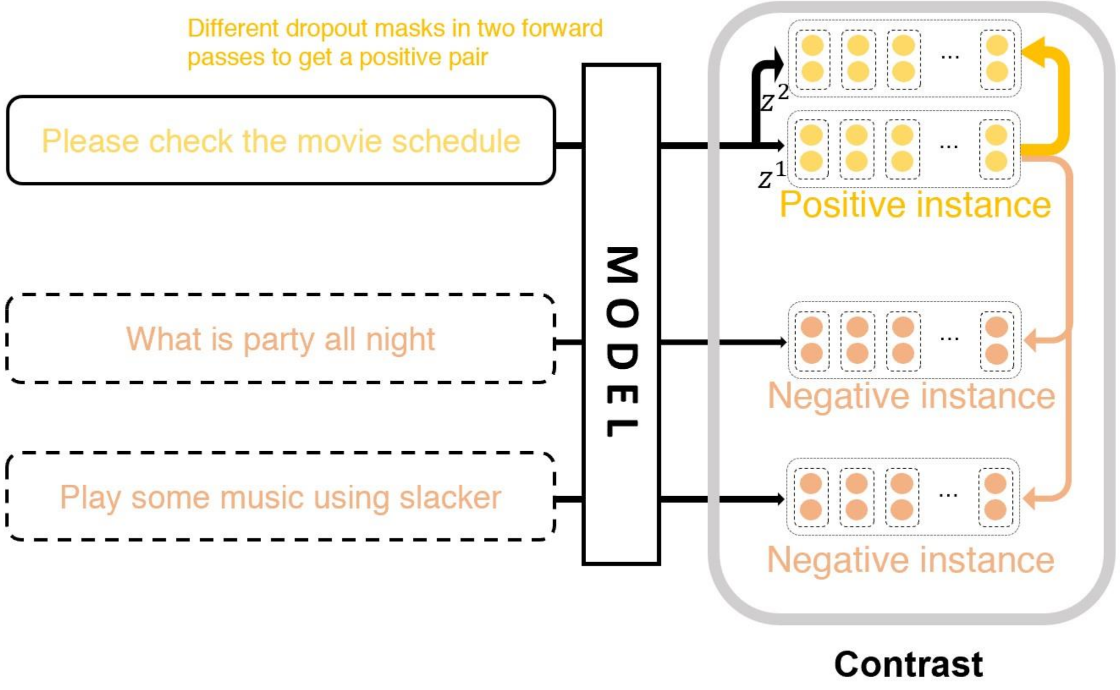
$$s.t.\ \ I(Z;T)\, is\, \text{maximized}$$

$$Z^{ssl} = \arg\max_{Z} I(Z;S)$$

$$Z^{ssl_{\min}} = \arg\min_{Z} H(Z|S)$$

$$s.t.\ \ I(Z;S)\, is\, \text{maximized}$$

Contrast Keeps Almost All the Key Information



Theorem 1 suggests maximizing $I(z^1, z^2)$ results in z^1 containing almost all the key information.

We minimize the following loss:

$$\mathcal{L}_C = \max \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{sim\left(z_i^1, z_i^2\right)}/\tau}{\frac{1}{N}\sum_{k=1}^{N} e^{sim\left(z_i^1, z_k^2\right)}/\tau}\right]$$

Reconstruction Discards the Redundant Information

<u>Theorem 2:</u> The sufficient self-supervised representation contains more redundant information in the input than the sufficient and minimal self-supervised representation. The latter contains an amount of the information, $I(X;S|T)$ that cannot be discarded from the input.

$$Z^{ssl} = \arg\max_{Z} I(Z;S)$$

$$Z^{ssl_{\min}} = \arg\min_{Z} H(Z|S)$$

$$s.t.\ \ I(Z;S)\, is \,\text{maximized}$$

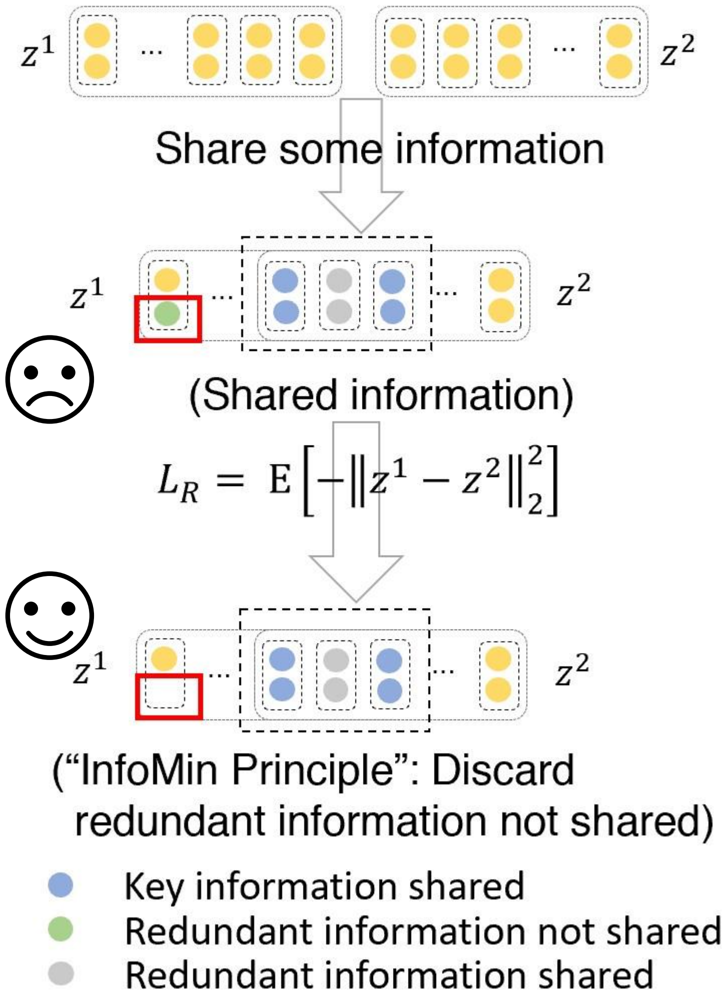
$$Z^{\text{sup}} = \arg\max_{Z} I(Z;T)$$

$$Z^{sup_{\min}} = \arg\min_{Z} H(Z|T)$$

$$s.t.\ \ I(Z;T)\, is \,\text{maximized}$$

$$\begin{aligned} I\left(Z^{ssl};X|T\right) &= I(X;S|T) + I\left(Z^{ssl};X|S,T\right) \\ &\geq I\left(Z^{ssl_{min}};X|T\right) = I(X;S|T) \\ &\geq I(Z^{sup_{min}};X|T) = 0 \end{aligned}$$

Reconstruction Discards the Redundant Information



Maximize $\mathbb{E}_{P_{Z^1,Z^2}}\left[\log P\left(Z^1|Z^2\right)\right]$ = Minimize $H\left(Z^1|Z^2\right)$

Reconstruct z^1 via z^2

maximize $\mathbb{E}_{P_{Z^1,Z^2}}\left[\log P\left(Z^1|Z^2\right)\right]$

under the constraint that

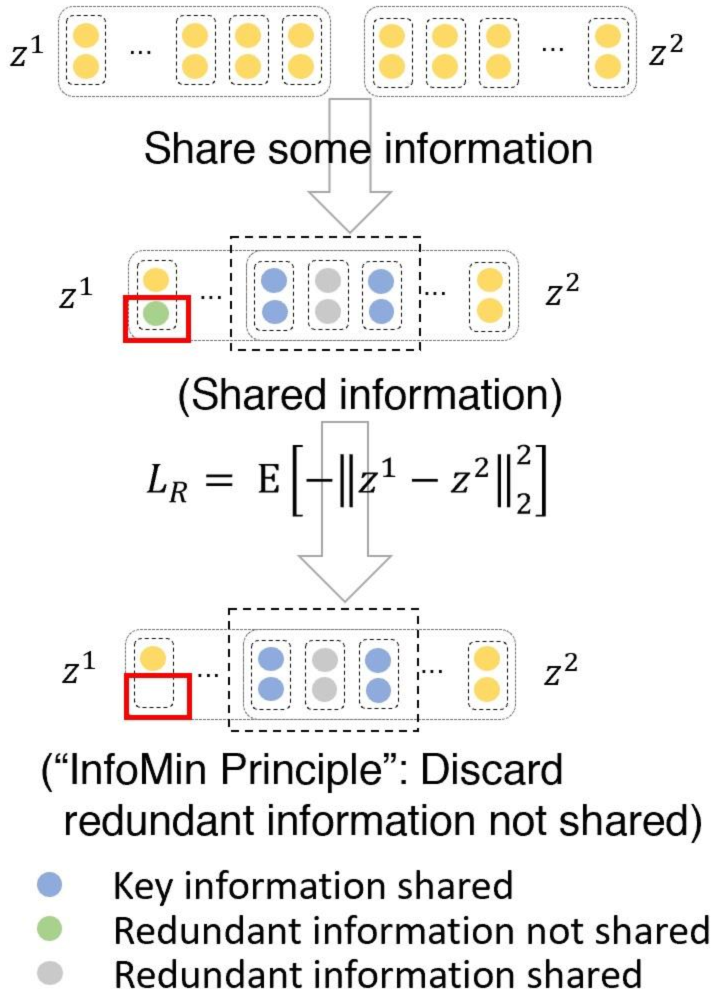
$I\left(z^1, z^2\right)$ is maximized: (by contrast operation)

We obtain:

$$z^{1^{ssl_{min}}}$$

Which contains the least redundant information according to *Theorem 2.*

Reconstruction Discards the Redundant Information



For eaiser optimization, we use

$$\mathbb{E}_{P_{Z^1,Z^2}}\left[\log Q_\Phi\left(Z^1|Z^2\right)\right]$$

As a lower bound of

$$\mathbb{E}_{P_{Z^1,Z^2}}\left[\log P\left(Z^1|Z^2\right)\right]$$

Where $Q_\Phi\left(Z^1|Z^2\right) \sim N\left(Z^1|Z^2, \sigma \mathrm{I}\right)$ (σI is a diagonal matrix)

We minimize the following loss:

$$\mathcal{L}_R = \mathbb{E}_{z^1,z^2 \sim P_{Z^1,Z^2}}\left[-\left\|z^1 - z^2\right\|_2^2\right]$$

Performance on Unsupervised (semantic textual similarity) Tasks

Model	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg.
GloVe embeddings (avg.)†	55.14	70.66	59.73	68.25	63.66	58.02	53.76	61.32
BERT_base (first – last avg.)†	39.70	59.38	49.67	66.03	66.19	53.87	62.06	56.70
BERT_base–flow†	58.40	67.10	60.85	75.16	71.22	68.66	64.47	66.55
BERT_base–whitening†	57.83	66.90	60.90	75.08	71.31	68.24	63.73	66.28
IS – BERT_base†	56.77	69.24	61.21	75.23	70.16	69.21	64.25	66.58
CT – BERT_base†	61.63	76.80	68.47	77.50	76.48	74.31	69.19	72.05
SCD – $\text{BERT}_\text{base}$$^\heartsuit$	66.94	78.03	69.89	78.73	76.23	76.30	**73.18**	74.19
SimCSE – BERT_base	67.01	82.14	73.76	80.49	79.01	77.04	69.94	75.63
InforMin-CL – BERT_base	**70.22**	**83.48**	**75.51**	**81.72**	**79.88**	**79.27**	71.03	**77.30**
$\text{RoBERTa}_\text{base}$ (first – last avg.)†	40.88	58.74	49.07	65.63	61.48	58.55	61.63	56.57
$\text{RoBERTa}_\text{base}$–whitening†	46.99	63.24	57.23	71.36	68.99	61.36	62.91	61.73
DeCLUTR – $\text{RoBERTa}_\text{base}$†	52.41	75.19	65.52	77.12	78.63	72.41	68.62	69.99
SCD – $\text{RoBERTa}_\text{base}$$^\heartsuit$	63.53	77.79	69.79	80.21	77.29	76.55	72.10	73.89
SimCSE – $\text{RoBERTa}_\text{base}$	**70.32**	82.48	**74.84**	**82.13**	**82.14**	**81.57**	68.62	**77.44**
InforMin-CL – $\text{RoBERTa}_\text{base}$	69.79	**82.57**	73.36	80.91	81.28	81.07	**70.30**	77.04
SimCSE – $\text{RoBERTa}_\text{large}$	**72.64**	83.78	**75.83**	**84.24**	**80.12**	**81.10**	69.81	**78.22**
InforMin-CL – $\text{RoBERTa}_\text{large}$	70.91	**84.20**	75.57	82.26	79.68	**81.10**	**72.81**	78.08

InforMin-CL outperforms all baselines significantly with BERT as an encoder.

Analysis of Experimental Result

Datasets	
BERT (16GB)	RoBERTa (160GB)
BooksCorpus	BooksCorpus
English Wikipedia	English Wikipedia
-	CC-NEWS
-	OPENWEB-TEXT
-	STORIES

The diverse large-scale high-quality pre-training datasets of RoBERTa contain less noise information, which results in InforMin-CL struggling to present its effects.

Performance on Supervised Tasks

Model	**MR**	**CR**	**SUBJ**	**MPQA**	**SST**	**TREC**	**MRPC**	**Avg.**
GloVe embeddings (avg.)†	77.25	78.30	91.17	87.85	80.18	83.00	72.87	81.52
Skip – thought†	76.50	80.10	93.60	87.10	82.00	92.20	73.00	83.50
Avg. BERT embeddings†	78.66	86.25	94.37	88.66	84.40	**92.80**	69.54	84.94
BERT– [CLS] embeddings†	78.68	84.85	94.21	88.23	84.13	91.40	71.13	84.66
IS – $\text{BERT}_{\text{base}}{}^{\dagger}$	81.09	**87.18**	94.96	88.75	85.96	88.64	74.24	85.83
SCD – $\text{BERT}_{\text{base}}{}^{\heartsuit}$	73.21	85.80	**99.56**	88.67	85.59	89.80	75.71	85.52
SimCSE – $\text{BERT}_{\text{base}}$	81.47	86.86	94.79	89.25	86.27	89.40	72.81	85.84
InforMin-CL – $\text{BERT}_{\text{base}}$	80.99	85.72	94.63	**89.47**	85.67	88.20	73.97	85.52
w/ MLM	**82.87**	87.05	95.22	88.43	**87.15**	92.20	**75.77**	**86.96**
SimCSE – $\text{RoBERTa}_{\text{base}}$	81.26	87.36	93.58	87.56	86.93	84.80	75.01	85.21
SCD – $\text{RoBERTa}_{\text{base}}{}^{\heartsuit}$	82.17	87.76	93.67	85.69	88.19	83.40	76.23	85.30
InforMin-CL – $\text{RoBERTa}_{\text{base}}$	82.22	88.08	93.57	**87.75**	87.59	86.60	76.99	86.11
w/ MLM	**83.49**	**88.69**	**94.79**	86.81	**88.30**	**89.40**	**77.57**	**87.01**
SimCSE – $\text{RoBERTa}_{\text{large}}$	80.85	85.99	93.08	87.65	86.33	89.00	72.46	85.05
InforMin-CL – $\text{RoBERTa}_{\text{large}}$	**82.50**	**88.32**	**93.81**	**89.38**	**87.64**	**90.80**	**74.49**	**86.71**

InforMin-CL outperforms all baselines with BERT or RoBERTa as the encoder.

Ablation Study

Influence of $\underline{\lambda}$ (the coefficient of reconstruction objective)

λ	**Avg. Sup**	**Avg. Unsup**
0.04	85.20	76.09
0.4	**85.52**	**77.30**
4	85.03	77.18

The performance of InforMin-CL on both unsupervised and supervised tasks rises first and falls later.

Ablation Study

Influence of $\underline{\beta}$ (the coefficient of MLM objectives)

Model	Avg. Sup	Avg. Unsup
w/o MLM	85.52	**77.30**
w/ MLM		
$\beta = 0.01$	86.46	**63.59**
$\beta = 0.1$ (ours)	86.96	63.25
$\beta = 1.0$	**87.04**	60.85

Consistently helps improve performance on supervised tasks but brings a significant drop on unsupervised tasks.

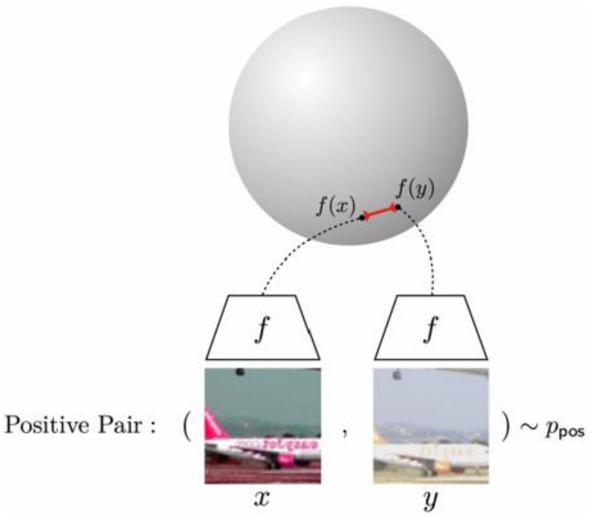
Ablation Study

Influence of Batch Size

Batch size	**64**	**128**	**256**
Avg. Sup	85.38	85.52	**85.77**
Avg. Unsup	76.64	**77.30**	76.14

Not sensitive to batch size

Why does InforMin-CL Work Well?



Alignment: How well positive pairs are aligned

Uniformity: How well the embeddings are uniformly distributed

[*Understanding Contrastive Representations Learning throuth Alignment and Uniformity on the Hyperspace. Wang et al. ICML 2020*.]

Why does InforMin-CL Work Well?

Qualitative Analysis:

$$\mathcal{L}_C = \max \left[\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\boxed{sim\left(z_i^1, z_i^2\right)}/\tau\right] - \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\log\frac{1}{N}\sum_{k=1}^{N}e^{sim\left(z_i^1, z_k^2\right)/\tau}\right]\right]$$

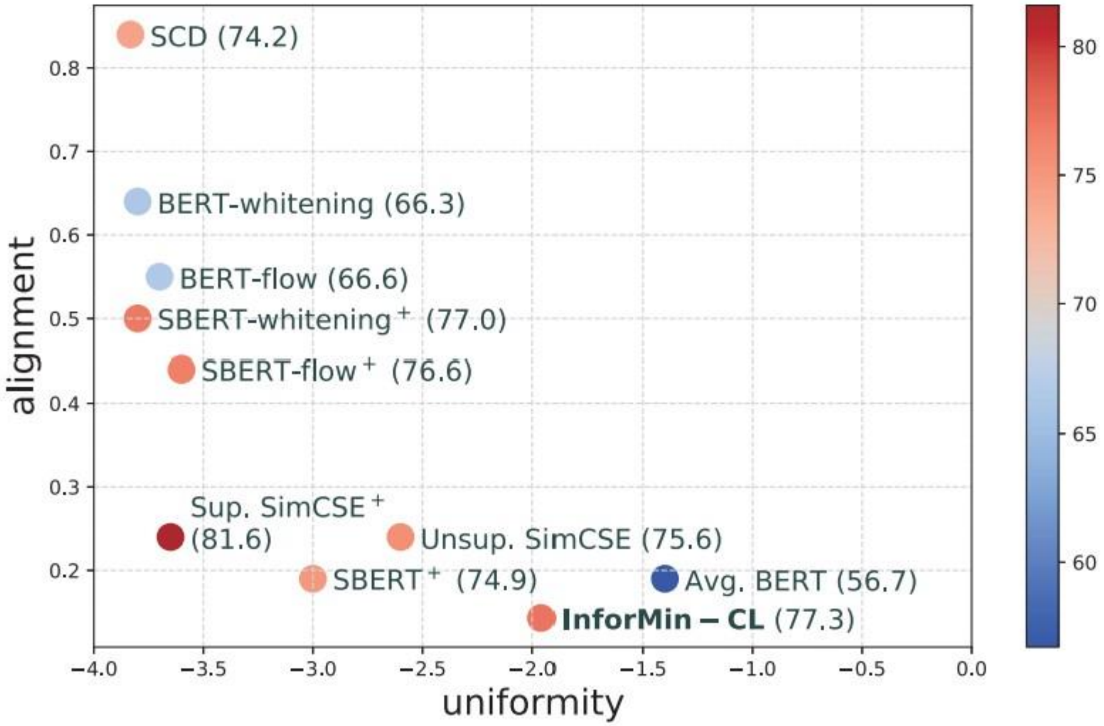
Optimizing L_R pulls z^1 and z^2 closer

$$\mathcal{L}_R = \mathbb{E}_{z^1, z^2 \sim P_{Z^1, Z^2}}\left[-\left\|z^1 - z^2\right\|_2^2\right]$$

Why does InforMin-CL Work Well?

Quantitative Analysis:

The smaller value, the better.

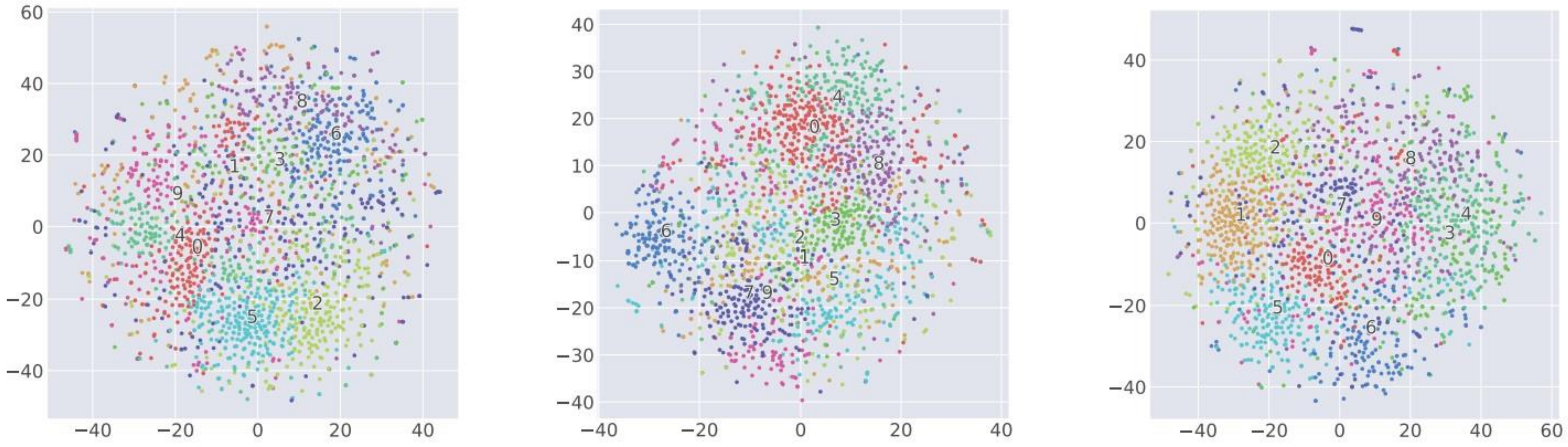


InforMin-CL achieves best in terms of *alignment*

Why does InforMin-CL Work Well?

Quantitative Analysis:

The t-SNE of sentence representations learned with models.



(a) SimCSE (b) SCD (c) InforMin-CL

Similar sentence pairs generated by InforMin-CL are more aligned.

Thanks!

Contact: chenshaobin000001@gmail.com

@shaobinchen3

知乎 临江仙