CROSS-MODAL KNOWLEDGE DISTILLATION IN MULTI-MODAL FAKE NEWS DETECTION

Zimian Wei, Hengyue Pan, Linbo Qiao, Xin Niu, Peijie Dong, Dongsheng Li

College of Computer, National University of Defense Technology
{weizimian16, hengyuepan, qiao.linbo, niuxin, dongpeijienudt, dsli}@nudt.edu.cn

ICASSP'22

220602 Chia-Chun Ho

Outline
of CMC

Introduction

Related Works

Problem Formulation

Methodology

Experiments

Conclusion

Comments

Introduction

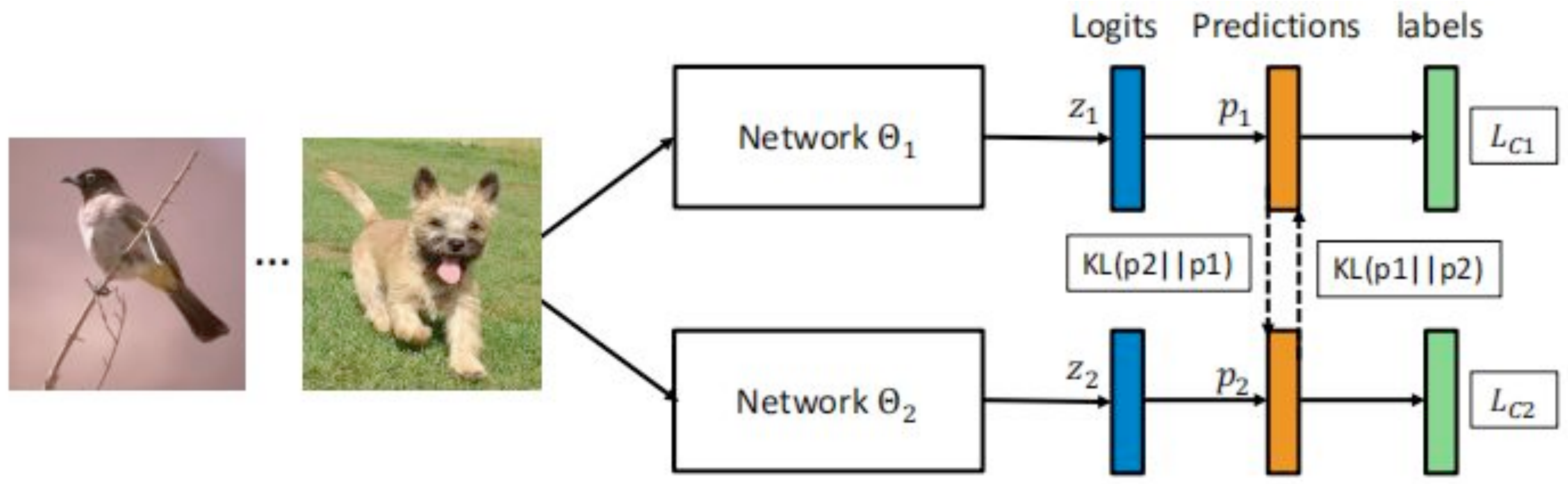
Fake News Detection

- Automatic fake news detection is important for normal society
  - To avoid the rampant dissemination of fake news on social media
  - To identify fake news according to extracted features
    - Textual contents, attached images, social contexts, etc
  - Mainly focus on textual & visual features in this paper

Introduction
Single/Multi-modal methods

- Single-modal methods
  - MVNN introduced a multi-branch CNN-RNN model to extract visual features
  - Some constructed ensemble classifier by 8 transformer-based pre-trained models
- Multi-modal methods
  - MVAE used a bi-model VAE to learn a shared representation between two networks
  - SpotFake+ integrated pre-trained LM & ImageNet models by multiple FCLs
    - They overlook cross-modal correlation knowledge to lead to sub-optimal results

Introduction
CMC



Deep mutual learning (CVPR'18)

- Inspired from Deep Mutual Learning (DML) that ensemble of networks
  - Learning collaboratively and and teach each other throughout the training process
- Propose a multi-modal fake news detector called CMC
  - To train two single-modal networks mutually
  - The distillation loss in CMC aims to exploit feature correlations between modalities
    - DML is to mimic the class posterior of each network with other peers

Introduction

Stages of CMC

- Mutual training stage
  - Single-modal networks are trained mutually in an ensemble learning paradigm
  - Capture cross-modal feature correlations by novel distillation loss
    - The positive pairs will be pulled together while the negative pairs will be separated
- Fusion mechanism training stage
  - The fusion mechanism based on BLOCK is trained
    - To further improve performance by better fitting discriminative information from modalities

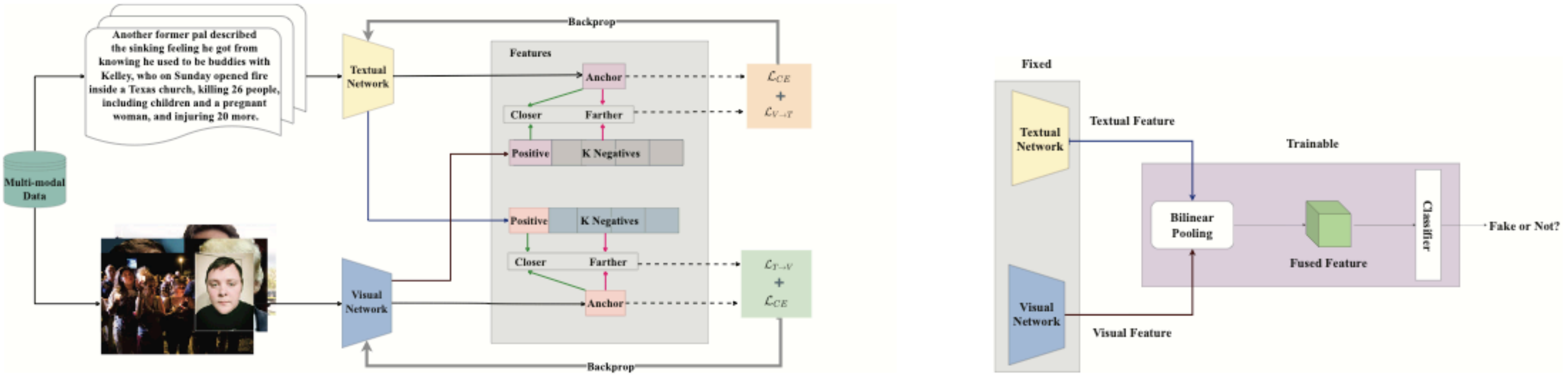
Introduction

Contributions

- Proposed a mutual learning strategy in multi-modal fake news detection
  - Collaboratively train the textual & visual networks to gain higher performance
    - Instead of integrating a shared representation between different modal networks
- Introduce a cross-modal distillation objective function as a soft target
  - To lead the single-modal network to learn feature correlations between modalities

Methodology

CMC

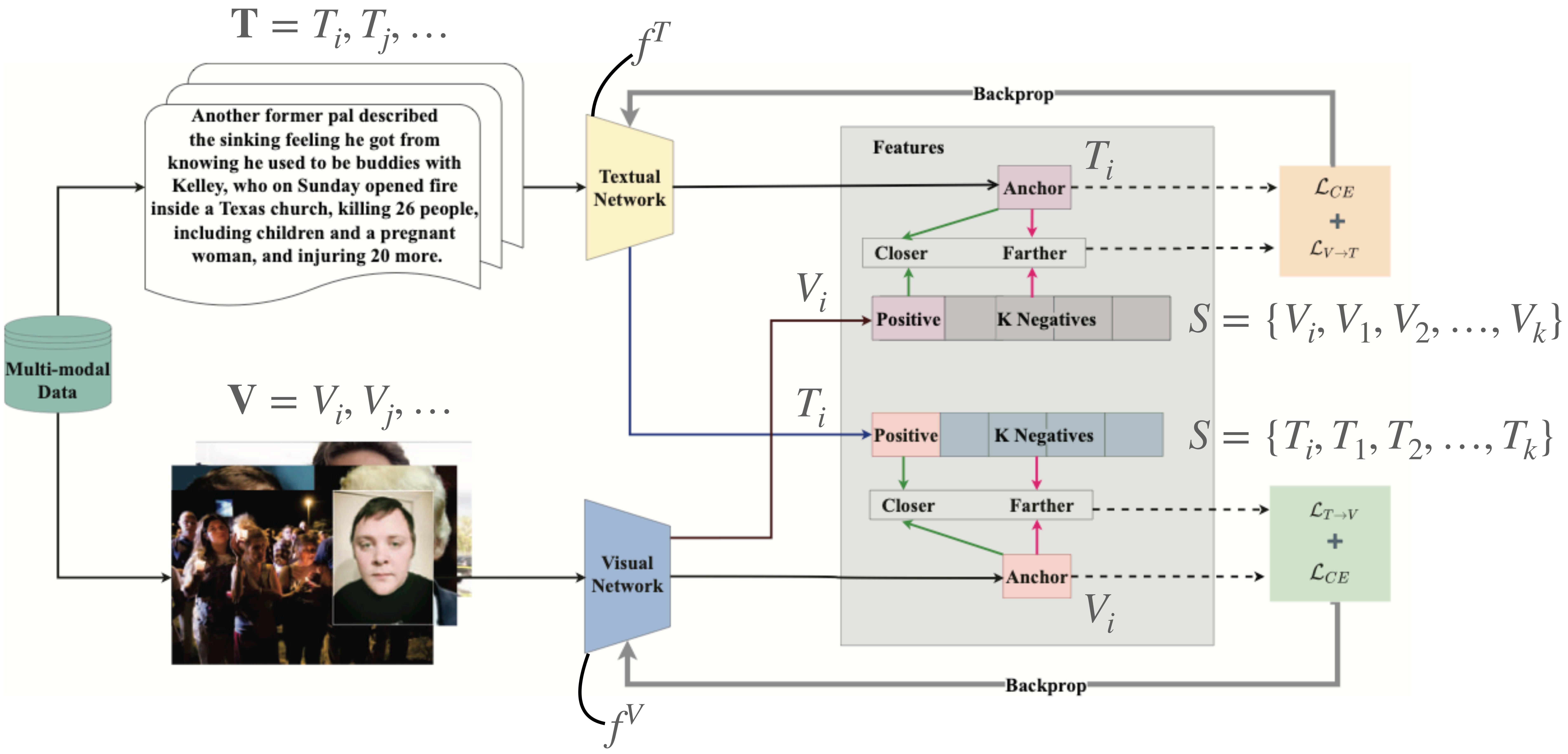


(a) Mutual training

(b) Fusion mechanism training

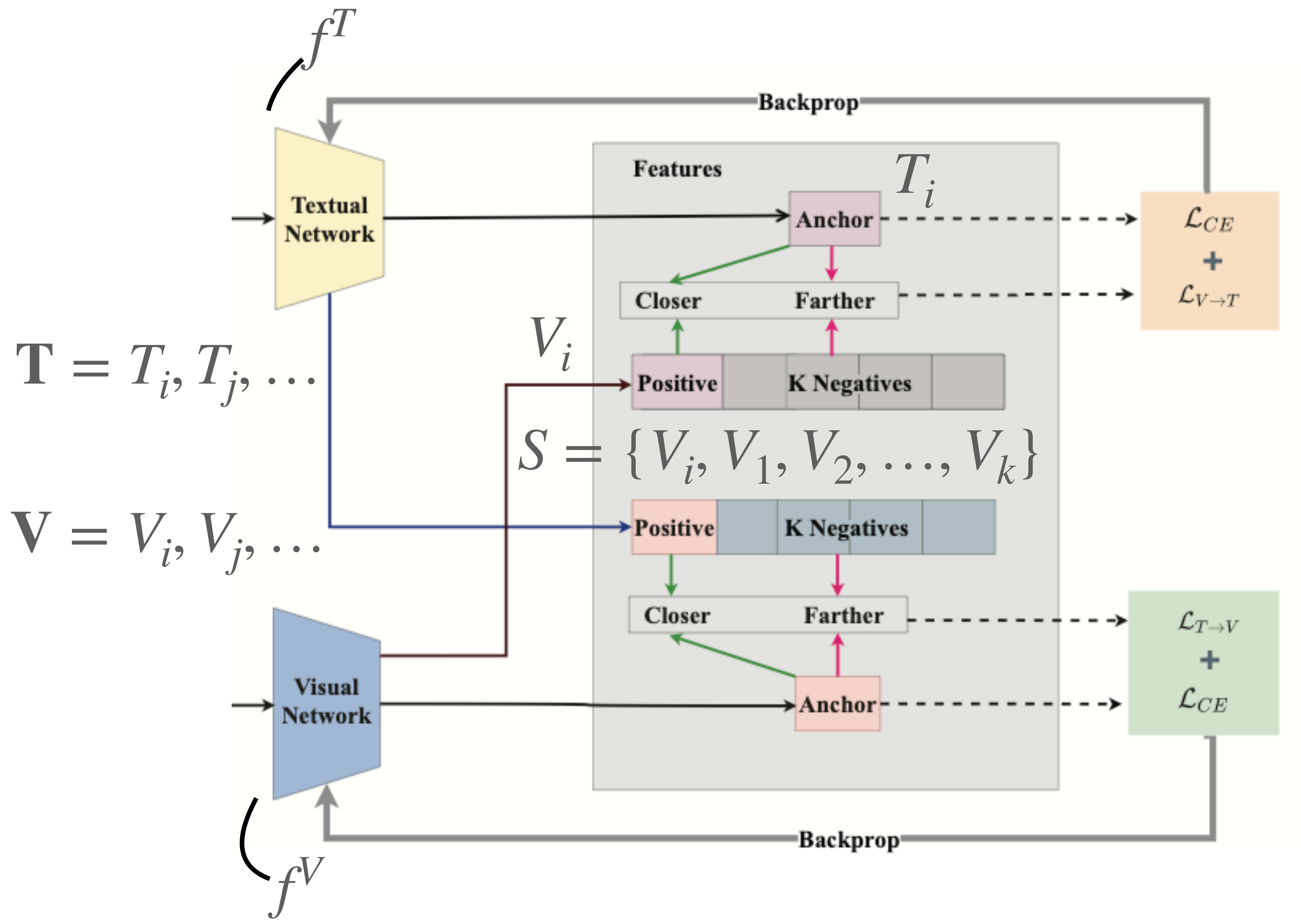
Methodology

Cross-modal Knowledge Distillation



Methodology

Cross-modal Knowledge Distillation

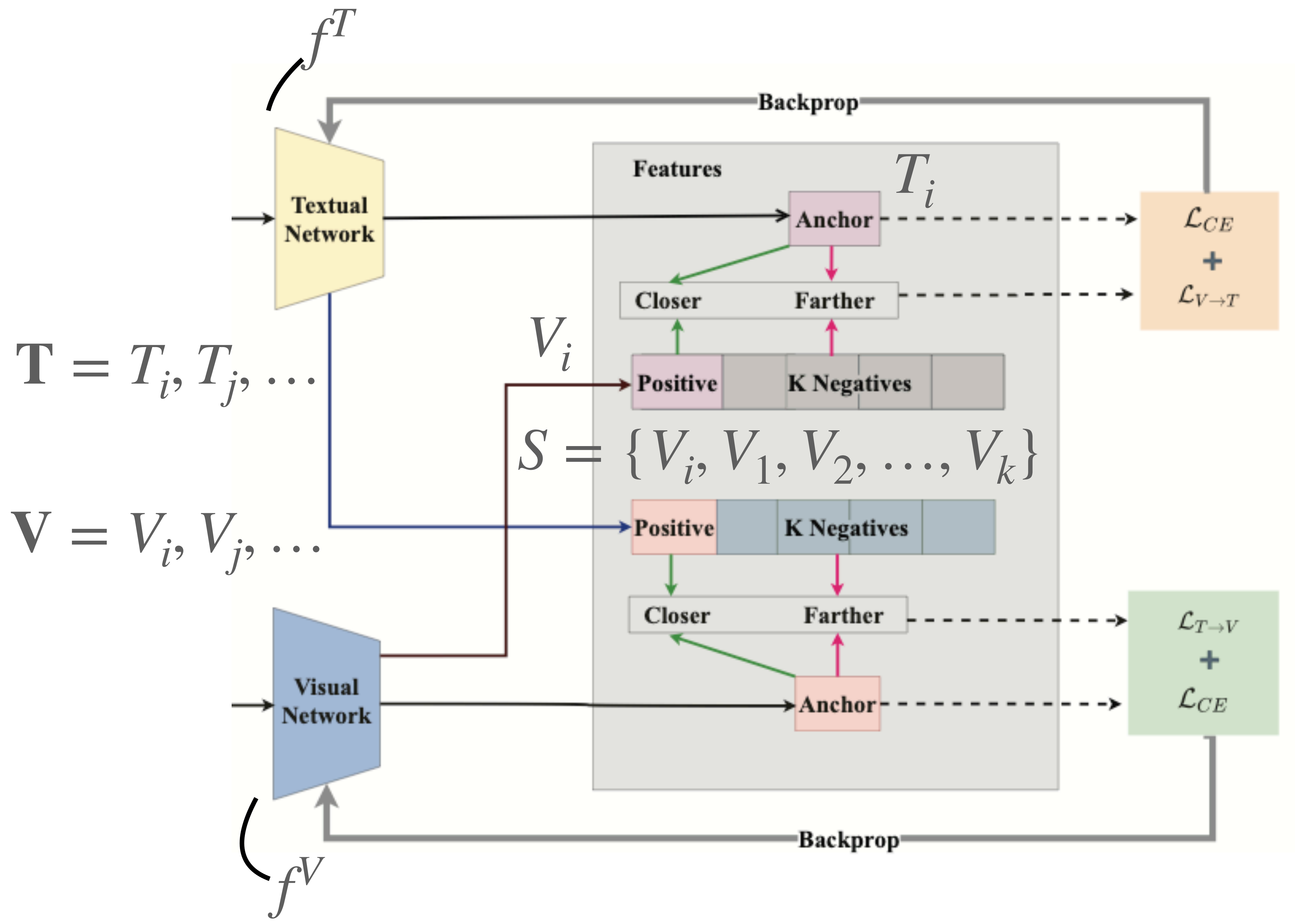


- The variable D decides whether V_j was drawn from the positive distribution ($D = 1$) or negative distribution ($D = 0$).

- The prior probabilities on D are as follows:
  - $p(D = 1) = \dfrac{1}{k + 1}$
  - $p(D = 0) = \dfrac{k}{k + 1}$

Methodology

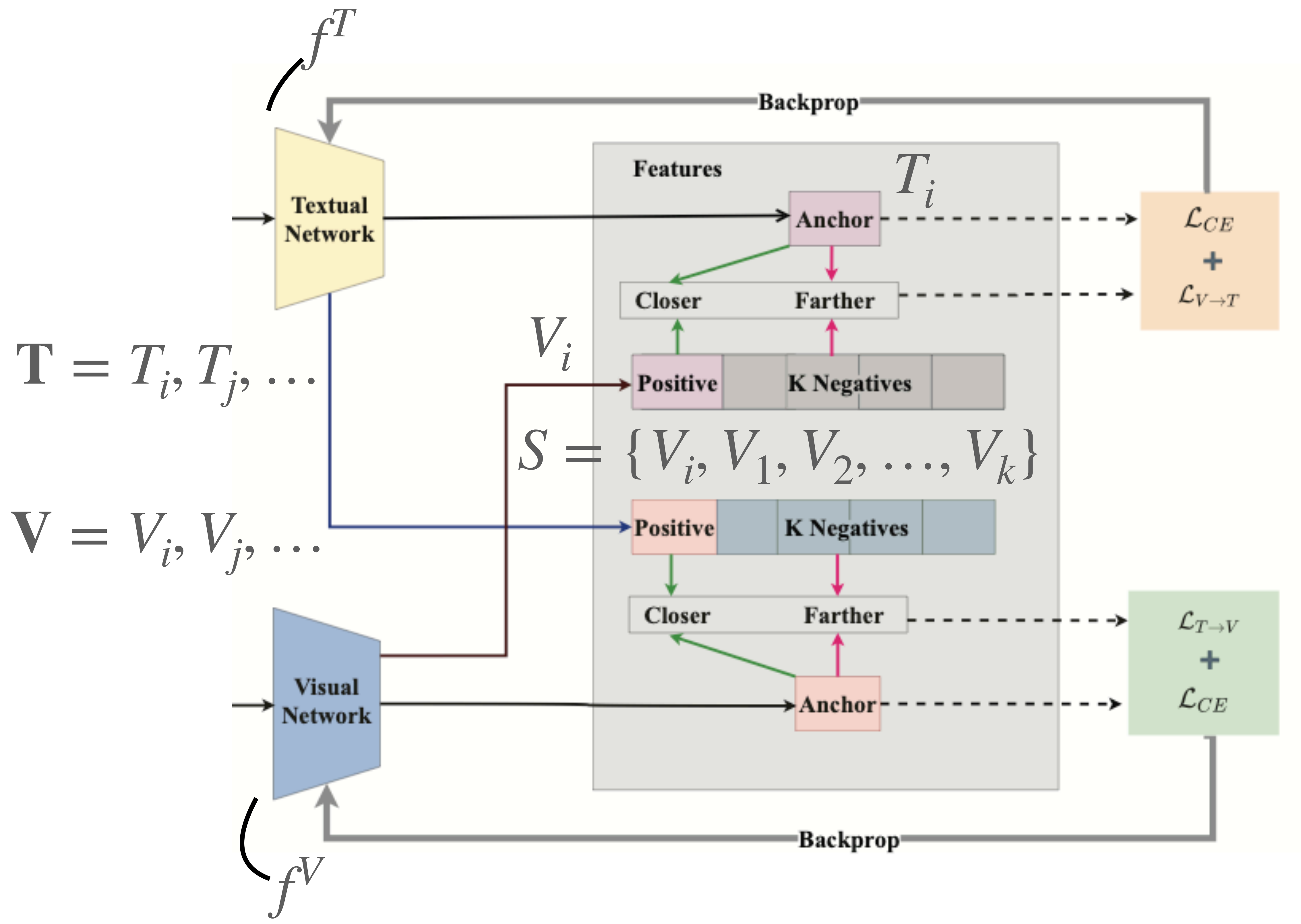
Cross-modal Knowledge Distillation



- Denote the negative distribution as p_n and positive distribution as p_m.
- Referring to NCE, formulate p_n as a uniform distribution over all atoms from $\mathbf{V}$.
- With N represent the dataset size, have following class-conditional probability:
- $p_n(V_j \mid D = 0, T_i) = \dfrac{1}{N}$

Methodology

Cross-modal Knowledge Distillation



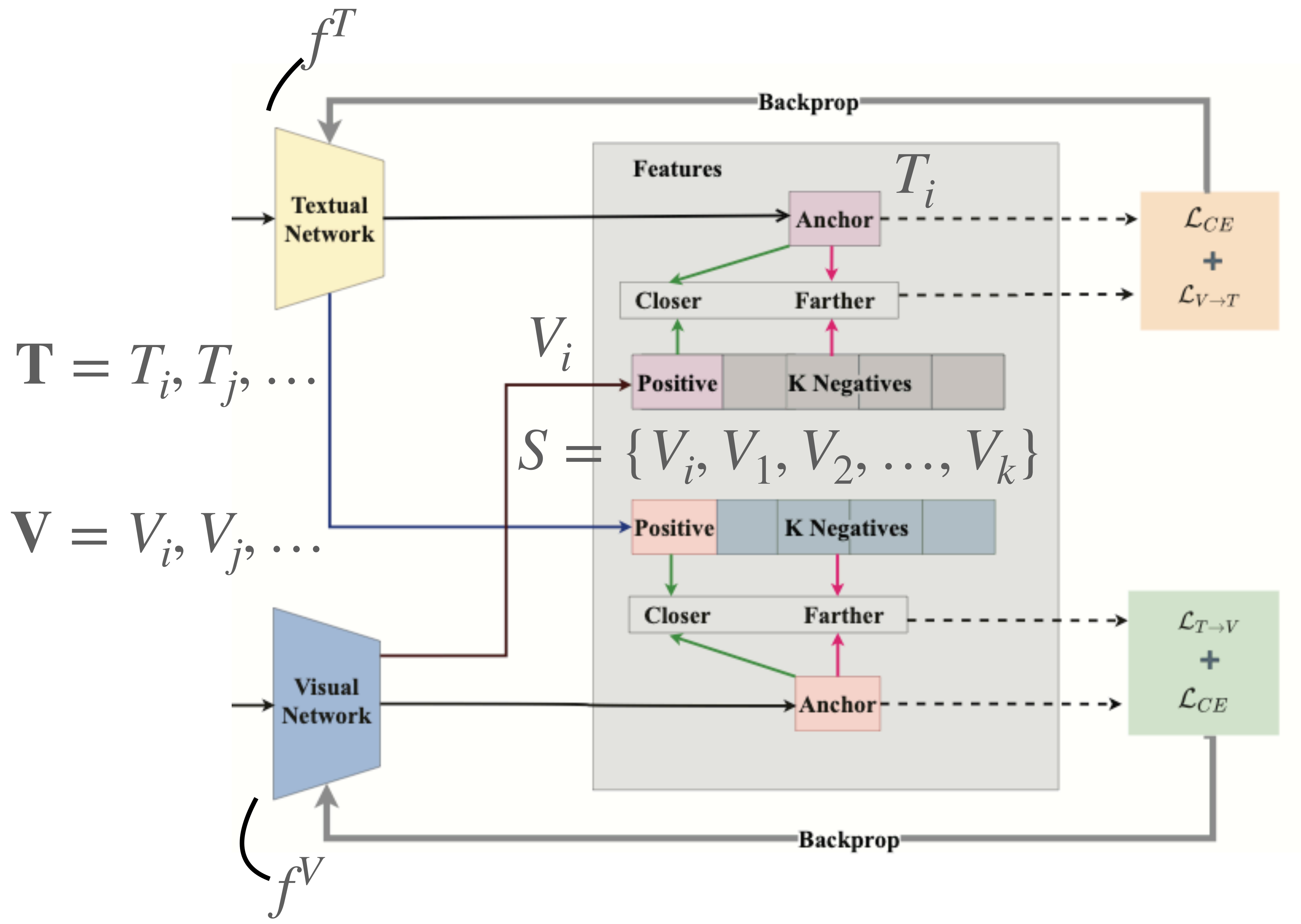
- Since $p_m(V_j \mid D = 1, T_i)$ is unknown, we model it by introducing a scoring function $\mathcal{H}(\,.\,)$ that is trained to achieve a high value for positive pairs and low for negative pairs.

$$p_m(V_j \mid D = 1, T_i) = \frac{\mathcal{H}(T_i, V_j)}{Z}$$

- $$= \frac{\exp\left(\frac{\phi_1(T_i) \cdot \phi_2(V_j)}{\|\phi_1(T_i)\| \cdot \|\phi_2(V_j)\|} \cdot \frac{1}{\tau}\right)}{Z}$$

Methodology

Cross-modal Knowledge Distillation



1×1 convolution layers
to transfer T_i, V_j to same dimension

- $$p_m(V_j \mid D = 1, T_i) = \frac{\mathscr{H}(T_i, V_j)}{Z} = \frac{\exp\left(\frac{\phi_1(T_i)\cdot\phi_2(V_j)}{\|\phi_1(T_i)\|\cdot\|\phi_2(V_j)\|}\cdot\frac{1}{\tau}\right)}{Z}$$

Normalizing constant

Temperature adjusts the concentration level

$$p(\mathbf{u}|C=1;\theta) = p_m(\mathbf{u};\theta) \quad p(\mathbf{u}|C=0) = p_n(\mathbf{u}). \quad (6)$$

Since we have equal probabilities for the two class labels, i.e. $P(C=1) = P(C=0) = 1/2$, we obtain the following posterior probabilities

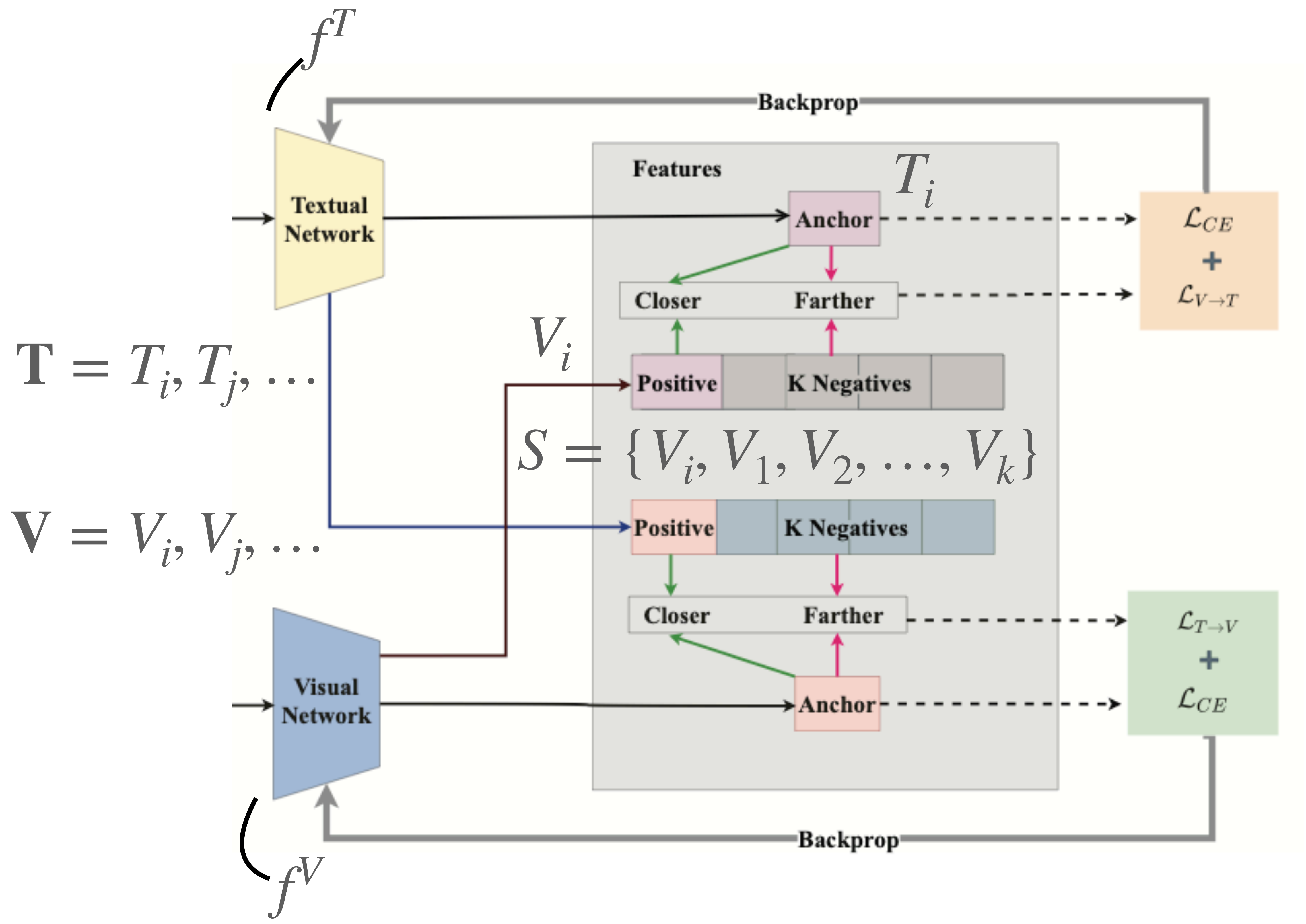
$$P(C=1|\mathbf{u};\theta) = \frac{p_m(\mathbf{u};\theta)}{p_m(\mathbf{u};\theta)+p_n(\mathbf{u})} \quad (7)$$

$$= h(\mathbf{u};\theta) \quad (8)$$

$$P(C=0|\mathbf{u};\theta) = 1-h(\mathbf{u};\theta). \quad (9)$$

Methodology

Cross-modal Knowledge Distillation



- The posterior probability for $D = 1$ is as follow:

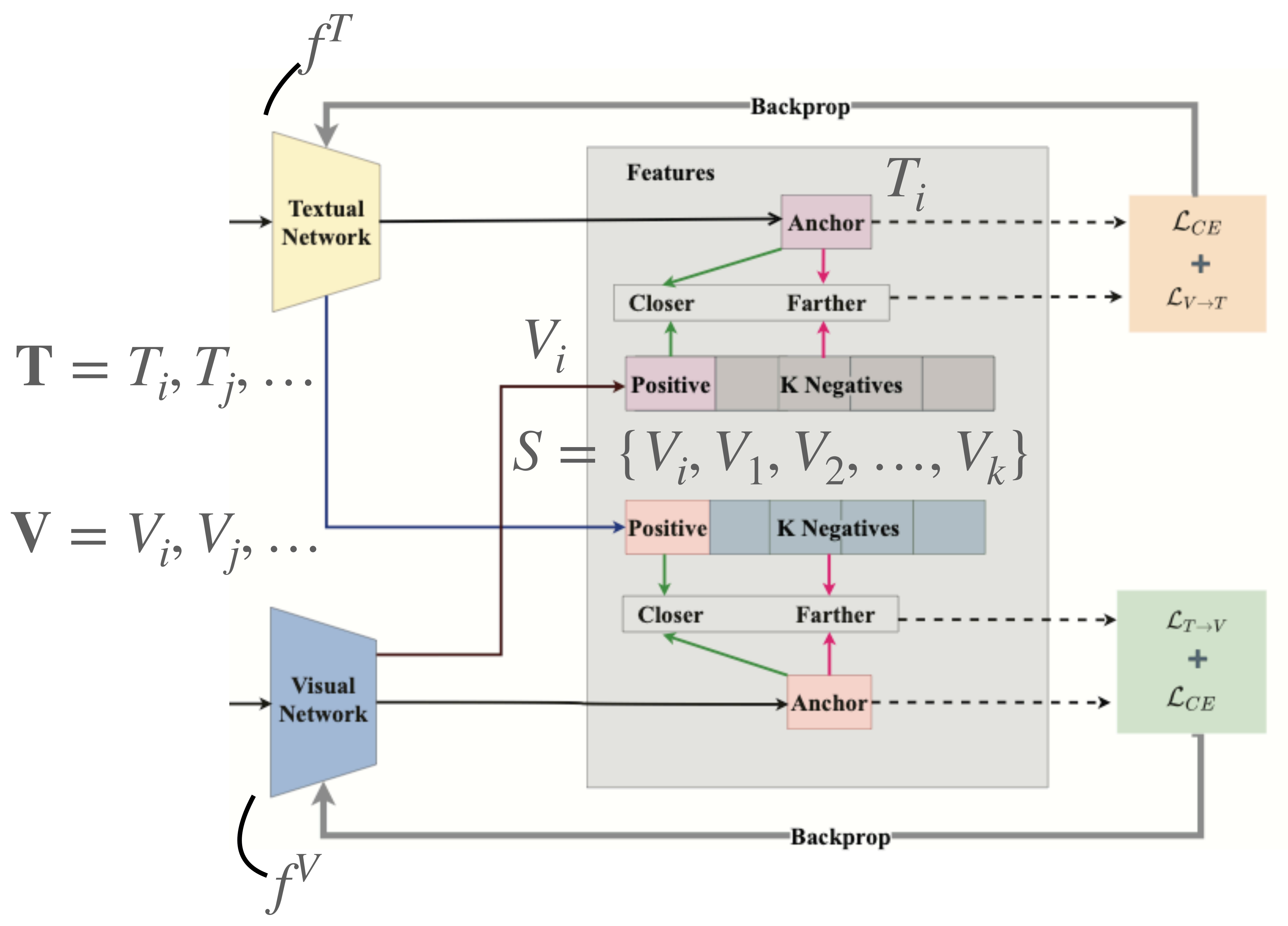
$$P(D=1 \mid V_j, T_i)$$

$$= \frac{p(D=1)p_m(V_j \mid D=1,T_i)}{p(D=1)p_m(V_j \mid D=1,T_i)+p(D=0)p_n(V_j \mid D=0,T_i)}$$

- $$= \frac{p_m(V_j \mid D=1,T_i)}{p_m(V_j \mid D=1,T_i)+\frac{k}{N}} = \frac{\mathscr{H}(T_i, V_j)}{\mathscr{H}(T_i, V_j)+\frac{k}{N}}$$

Methodology

Cross-modal Knowledge Distillation



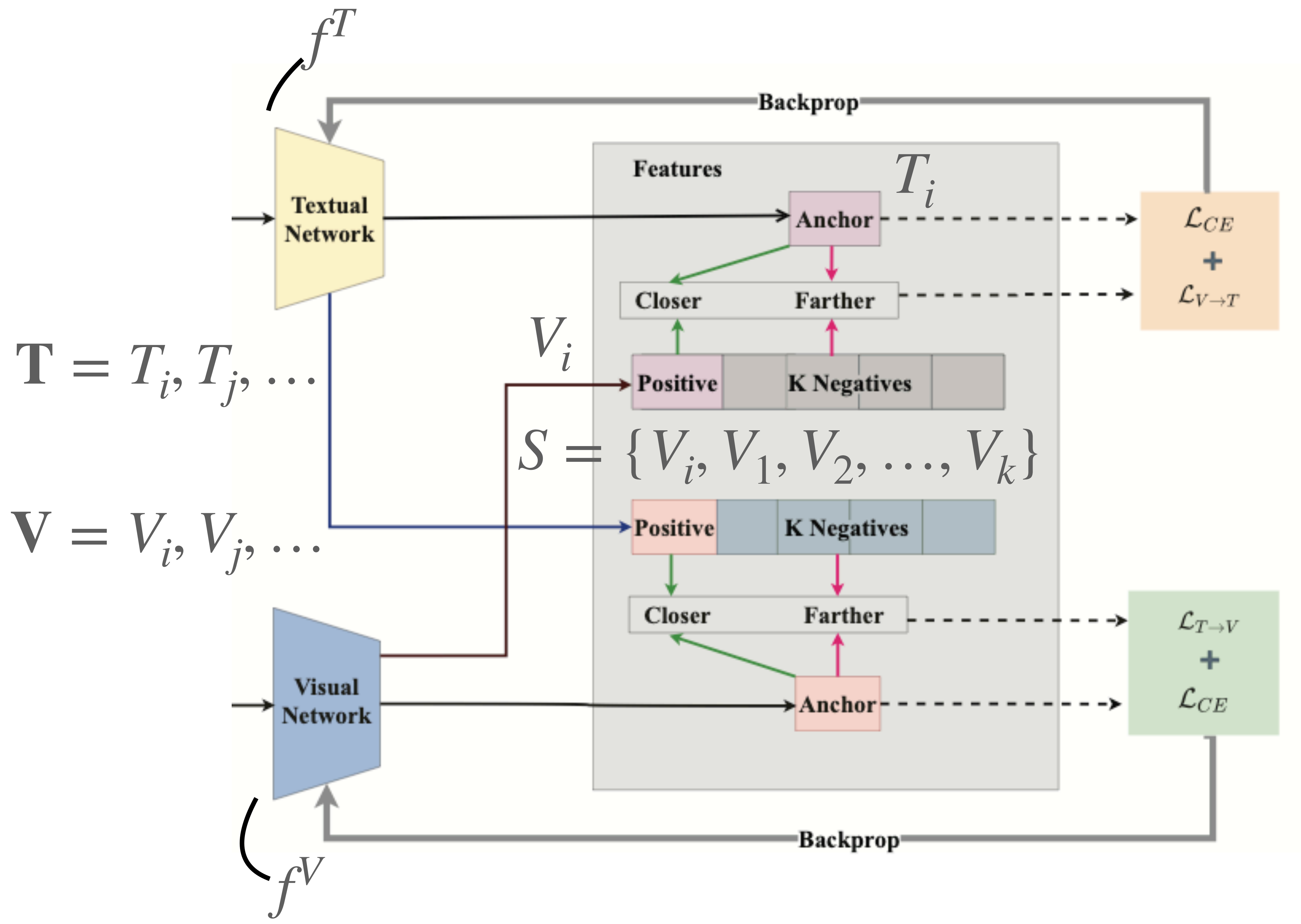
- The objective of partial cross-modal distillation for the textual network is formulated as follows:

$$\begin{aligned}
\mathscr{L}_{V\to T} &= -\underset{V_j \sim p_m(\cdot|T_i)}{\mathbb{E}} [\log(P(D=1 \mid V_j, T_i))] \\
&\quad -k \cdot \underset{V_j \sim p_n(\cdot|T_i)}{\mathbb{E}} [1 - \log(P(D=1 \mid V_j, T_i))] \\
&= -\underset{V_j \sim p_m(\cdot|T_i)}{\mathbb{E}} [\log(\frac{\mathscr{H}(T_i, V_j)}{\mathscr{H}(T_i, V_j) + \frac{k}{N}})] \\
&\quad -k \cdot \underset{V_j \sim p_n(\cdot|T_i)}{\mathbb{E}} [\log(1 - \frac{\mathscr{H}(T_i, V_j)}{\mathscr{H}(T_i, V_j) + \frac{k}{N}})]
\end{aligned}$$

-

Methodology

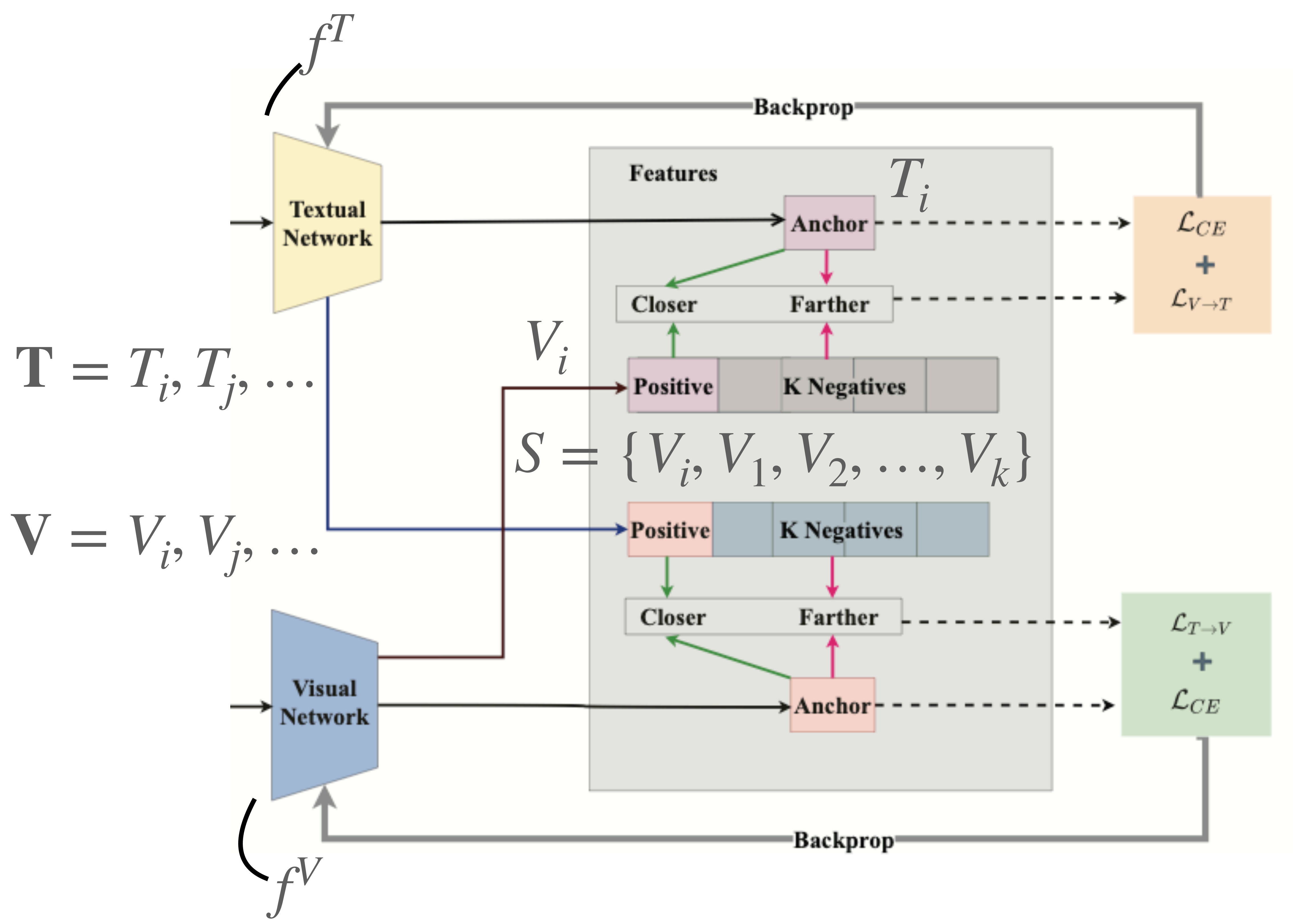
Cross-modal Knowledge Distillation



- Since train with a cohort of two networks, the total cross-modal distillation objective function is the summation of $\mathscr{L}_{V\to T}$, $\mathscr{L}_{T\to V}$ as follows:
- $\mathscr{L}_{distill} = \mathscr{L}_{V\to T} + \mathscr{L}_{T\to V}$

Methodology

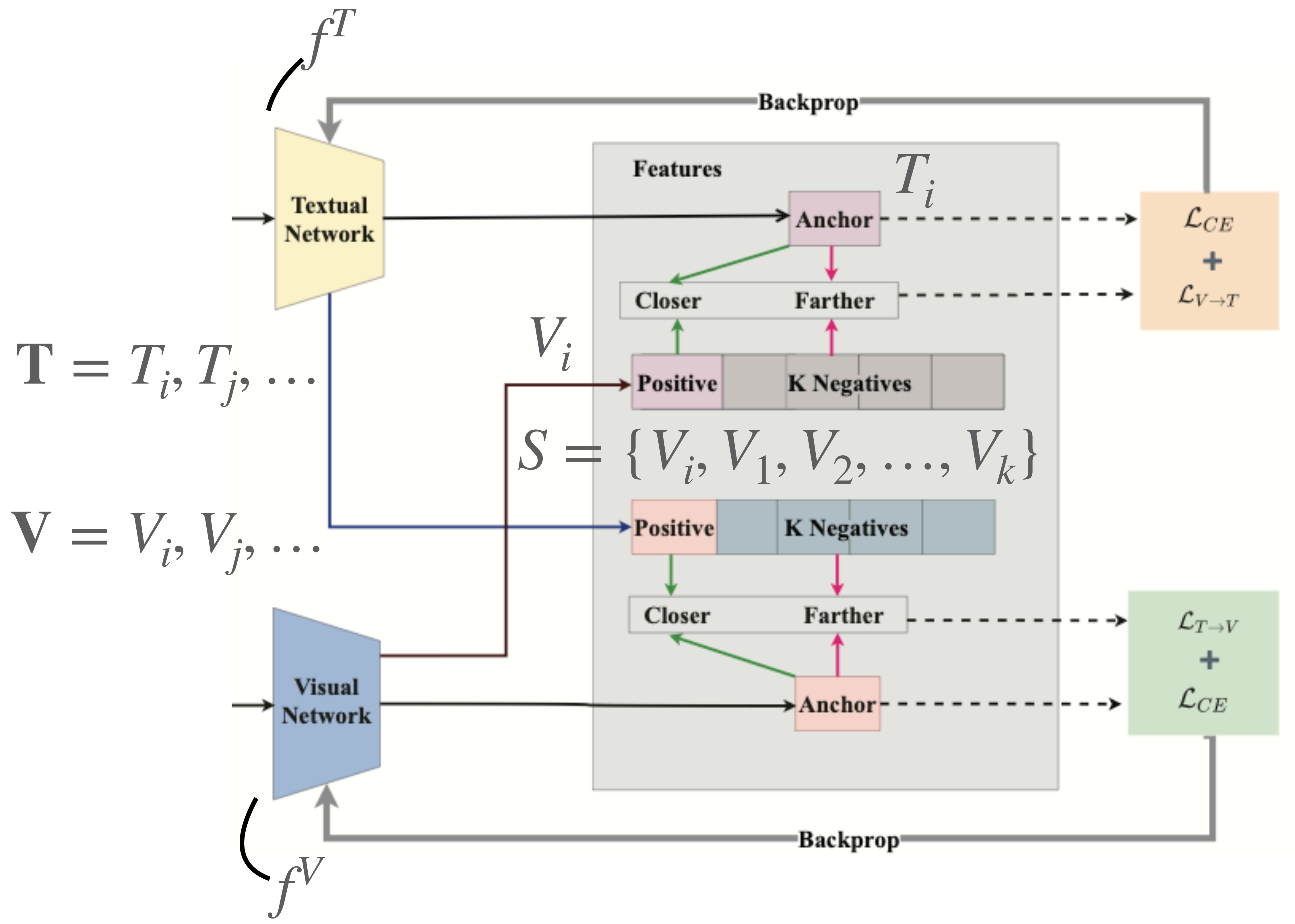
Cross-modal Knowledge Distillation



- Then the overall objective for two network f^T, f^V can be formulated as:
  - $\mathscr{L}_{obj_T} = \alpha \cdot \mathscr{L}_{distill} + \mathscr{L}^T_{CE}$
  - $\mathscr{L}_{obj_V} = \beta \cdot \mathscr{L}_{distill} + \mathscr{L}^V_{CE}$

Methodology

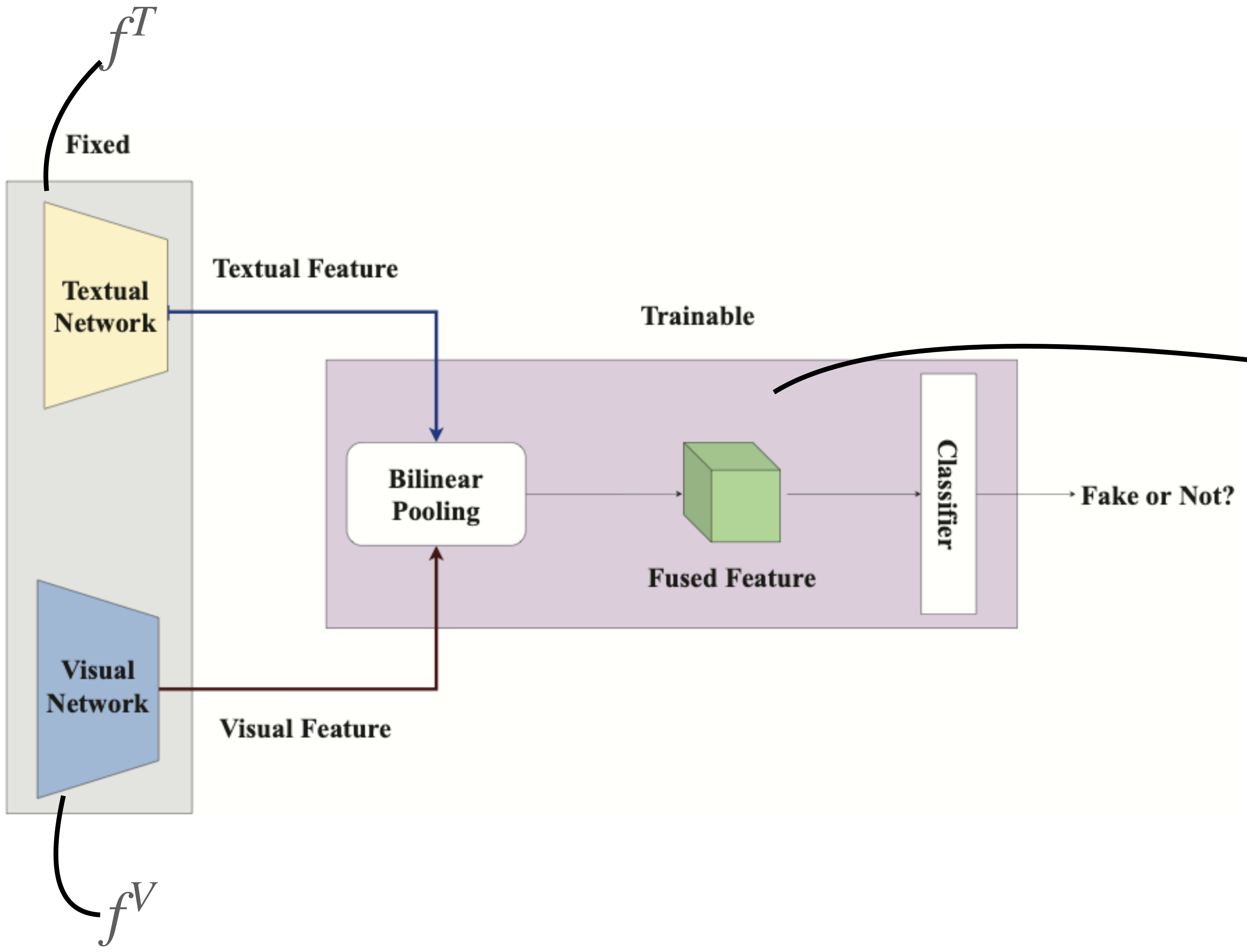
Mutual Learning Process



- The textual & visual networks perform fake news detection tasks separately.
  - Instead of sharing a concatenated representation.
- Their training process is closely intervened by each other.
- In each iteration, update the parameters of two networks according to their own predictions and representation correlations with the other peer.

Methodology

Fusion Mechanism



- Similar to BLOCK, the textual feature x^1 and the visual feature x^2 are projected to a new feature space by an associate tensor T.
- $r = T \times_1 x^1 \times_2 x^2$
- The final fused tensor r is feed into softmax function to identify fake news.

Experiments

Datasets & Results

Dataset	Method	Acc	Fake News			Real News		
			Prec	Rec	F1	Prec	Rec	F1
Weibo	att-RNN[11]	0.788	0.862	0.686	0.764	0.738	0.89	0.807
	EANN[14]	0.827	0.847	0.812	0.829	-	-	-
	MVAE[3]	0.824	0.854	0.769	0.809	0.802	0.875	0.837
	Spotfake[13]	0.892	0.902	**0.964**	**0.932**	0.847	0.656	0.739
	MVNN [1]	0.846	0.809	0.857	0.832	-	-	-
	CARMN [15]	0.869	0.935	0.796	0.860	0.820	0.944	0.878
	CMC	**0.908**	**0.940**	0.869	0.899	**0.876**	**0.945**	**0.907**
Politi	RoBERTa-MWSS [16]	0.82	-	-	-	0.82	-	-
	SAFE[5]	0.874	-	-	-	0.889	0.903	0.896
	Spotfake+[4]	0.846	-	-	-	-	-	-
	TM [17]	0.871	-	-	-	0.901	-	-
	LSTM-ATT [18]	0.832	-	-	-	0.836	0.832	0.829
	DistilBert [19]	-	**0.875**	0.636	0.737	0.647	0.88	0.746
	CMC	**0.894**	0.806	**0.862**	**0.833**	**0.944**	**0.92**	**0.932**
Gossip	RoBERTa-MWSS [16]	0.80	-	-	-	0.80	-	-
	SAFE[5]	0.838	-	-	-	0.857	0.937	0.895
	Spotfake+[4]	0.856	-	-	-	-	-	-
	TM [17]	0.842	-	-	-	0.896	-	-
	LSTM-ATT [18]	0.842	-	-	-	0.839	0.842	0.821
	DistilBert [19]	-	0.805	0.527	0.637	0.866	0.960	0.911
	CMC	**0.893**	**0.826**	**0.657**	**0.692**	**0.920**	**0.963**	**0.935**

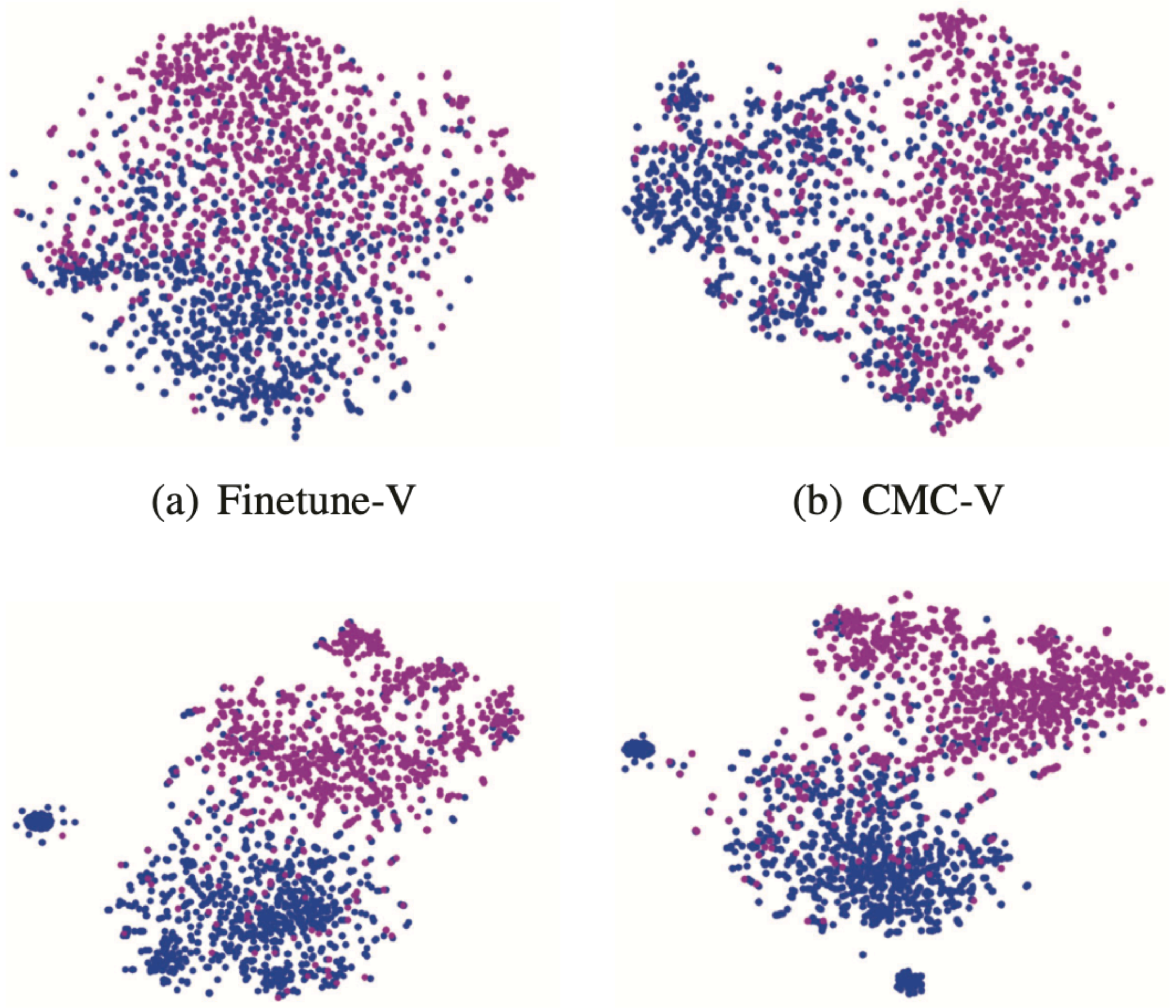
Statistic	**Training Set**		**Test Set**		**All**
	fake	real	fake	real	
Weibo	3749	3783	1000	996	9528
PolitiFact	135	246	29	75	485
GossipCop	2036	7974	545	2285	12840

Experiments

Ablation Study

Method	Modal	Acc	Fake News Prec	Fake News Rec	Fake News F1	Real News Prec	Real News Rec	Real News F1
Finetune-V	S	0.594	0.590	0.617	0.603	0.597	0.570	0.583
Finetune-T	S	0.898	0.905	0.867	0.898	0.870	0.906	0.899
CMC-V	S	0.689	0.666	0.764	0.711	0.722	0.614	0.664
CMC-T	S	0.904	0.936	0.869	0.898	0.874	0.941	0.900
CMC-shared	M	0.896	0.911	0.88	0.895	0.876	0.914	0.898
CMC	M	0.908	0.940	0.891	0.900	0.883	0.945	0.907

- Finetune-V & Finetune-T
  - Single-modal networks in CMC but are trained with a single cross-entropy loss
- CMC-V & CMC-T
  - Single-modal networks of CMC that are trained with a single cross-entropy loss and proposed cross-modal distillation
- CMC-shared
  - The variant of CMC that applies a shared representation between two single-modal networks.



(a) Finetune-V (b) CMC-V

(c) Finetune-T (d) CMC-T

Conclusion
of CMC

- Proposed a two-stage multi-modal fake news detection framework called CMC
    - To collaboratively train two single-modal networks
    - Transfers the cross-modal feature correlation by a novel distillation method
- The cross-model distillation loss is introduced to improve the capacity of single-modal networks
    - By the feature correlations from the other peer
- The fusion mechanism is trained to further improve the performance
    - By utilizing the discriminative information from different modalities

Comments
of CMC

- Focus on cross-modal feature correlation
- Design the distillation loss
- Fusion mechanism also different with other people
- Not showing the CMC / CMC-shared t-SNE result unfortunately