

AAAI2021

FL-MSRE: A Few-Shot Learning based Approach to Multimodal Social Relation Extraction

Hai Wan, Manrong Zhang, Jianfeng Du, Ziling Haung, Yufei Yang, Jeff Z. Pan

JAIST 情報科学系 修士1年
林 貴斗(Hayashi Takato)

Related Work – Few-Shot Learning

Few-Shot Learning :

少数のサンプルしかないクラスが含まれるデータでも，効率的に学習することができる学習手法

具体的なタスク例：

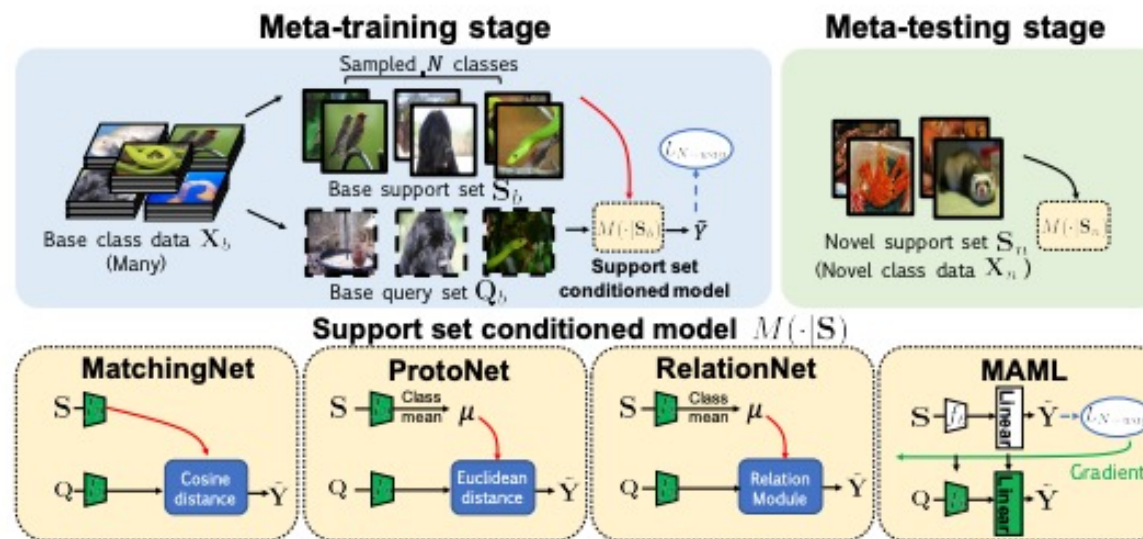
小鳥×100,000枚
イヌ×100,000枚
ヘビ×100,000枚

pre-trainingに使う

⋮
ネコ×100,000枚
牛×100,000枚

カニ×1~5枚
スカンク×1~5枚

こっちを分類したい(2値分類)



Wei-Yu Chen et al, "A closer Look at Few-shot Classification", ICLR2019

の画像があるときに，豊富なサンプルのあるクラスの画像を使って，カニやスカンクの画像を正確に分類したい

ただし，少数データをDNNにそのまま学習させると，過学習してしまうという問題がある。

少数データ→誤差が大きい→パラメータの更新幅が大きい→過学習（ハヤシの解釈）

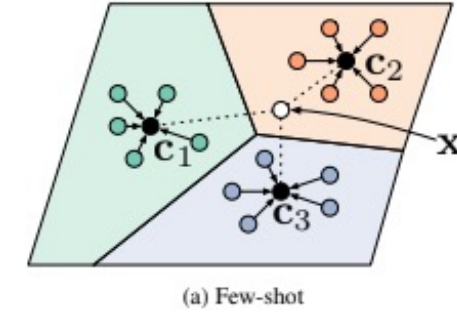
Related Work – Few-Shot Learning

Prototypical Networks : 同じクラスは近くに, 異なるクラスは遠くにマッピングする組み込み関数 f_ϕ をトレーニング

Snell et al , "Prototypical Networks for Few-shot Learning" , NeurIPS2017

Notation

- $\mathbf{x}(\in \mathbb{R}^D)$: D dimensional samples(vector)
- $S = \{(x_1, y_1), \dots, (x_N, y_N)\}$: N samples which are observed as input samples
- $y_i \in \{1, \dots, K\}$: label
- \mathbf{c}_k : M dimensional representation, or **prototype**
- $f_\phi(\cdot: \mathbb{R}^D \rightarrow \mathbb{R}^M)$: Embedded function with learnable parameters ϕ ex. DNN
- $d(A, B)$: distance function



$$\mathbf{c}_k = \frac{1}{|S_k|} \sum_{(\mathbf{x}_i, y_i) \in S_k} f_\phi(\mathbf{x}_i) \quad \cdot \cdot \cdot (1)$$

$$p_\phi(y = k | \mathbf{x}) = \frac{\exp(-d(f_\phi(\mathbf{x}), \mathbf{c}_k))}{\sum_{k'} \exp(-d(f_\phi(\mathbf{x}), \mathbf{c}_{k'}))} \quad \cdot \cdot \cdot (2)$$

Table 1: Few-shot classification accuracies on Omniglot. *Uses non-standard train/test splits.

Model	Dist.	Fine Tune	5-way Acc.		20-way Acc.	
			1-shot	5-shot	1-shot	5-shot
MATCHING NETWORKS [32]	Cosine	N	98.1%	98.9%	93.8%	98.5%
MATCHING NETWORKS [32]	Cosine	Y	97.9%	98.7%	93.5%	98.7%
NEURAL STATISTICIAN [7]	-	N	98.1%	99.5%	93.2%	98.1%
MAML [9]*	-	N	98.7%	99.9%	95.8%	98.9%
PROTOTYPICAL NETWORKS (OURS)	Euclid.	N	98.8%	99.7%	96.0%	98.9%

Objective function : $J(\phi) = -\log p_\phi(y = k | \mathbf{x})$

Related Work – Few-Shot Learning

Prototypical Networks :

Input: Training set $\mathcal{D} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)\}$, where each $y_i \in \{1, \dots, K\}$. \mathcal{D}_k denotes the subset of \mathcal{D} containing all elements (\mathbf{x}_i, y_i) such that $y_i = k$.

Output: The loss J for a randomly generated training episode.

$V \leftarrow \text{RANDOMSAMPLE}(\{1, \dots, K\}, N_C)$

▷ Select class indices for episode

for k in $\{1, \dots, N_C\}$ **do**

$S_k \leftarrow \text{RANDOMSAMPLE}(\mathcal{D}_{V_k}, N_S)$

▷ Select support examples

$Q_k \leftarrow \text{RANDOMSAMPLE}(\mathcal{D}_{V_k} \setminus S_k, N_Q)$

▷ Select query examples

$\mathbf{c}_k \leftarrow \frac{1}{N_C} \sum_{(\mathbf{x}_i, y_i) \in S_k} f_\phi(\mathbf{x}_i)$

▷ Compute prototype from support examples

end for

$J \leftarrow 0$

▷ Initialize loss

for k in $\{1, \dots, N_C\}$ **do**

for (\mathbf{x}, y) in Q_k **do**

$J \leftarrow J + \frac{1}{N_C N_Q} \left[d(f_\phi(\mathbf{x}), \mathbf{c}_k) + \log \sum_{k'} \exp(-d(f_\phi(\mathbf{x}), \mathbf{c}_{k'})) \right]$

▷ Update loss

end for

end for

N : number of examples in the **training set**

K : number of classes in the **training set**

$N_C (< K)$: number of classes per **episode**

N_S : number of **support** examples per class

N_Q : number of **query** examples per class

$\text{RandomSample}(S, N)$:

A set of N elements chosen uniformly at random from set S , without replacement.

TVシリーズとその原作から、二人の顔画像、二人について言及されているテキスト、二人の社会的関係性が含まれるデータセットを作成した。

顔画像とテキストを用いてマルチモーダルな社会的関係抽出(Social Relation Extraction:**SRE**)を提案した。また、社会的関係の不均衡さに対処するために、**Few-Shot Learning**の手法を取り入れることを提案した。

これらの手法を併用した新しいSREであるFE-MSRE(**F**ew-Shot **L**earning based Approach to **M**ultimodal **S**ocial **R**elation **E**xtraction)はテキスト単体のSREの分類精度を大きく上回った。

さらに、データセットに異なる画像から得られた顔画像を用いた場合でも、同じ画像から得られた顔画像を用いた場合と同様の性能を発揮することを明らかにした。

Introduction



仮説：画像とテキストがそれ単体では，認識できない部分を**補い合う**ことで社会的関係性の推定精度が向上するのではないかと

Case(b)：

「hug」というキーワードから親密な二人が写っていると推定．Barack ObamaとMaliaの顔画像から異性かつ年齢が離れていることがわかるので**父娘**の関係と推定

Case(c)：

「hug」というキーワードから親密な二人が写っていると推定．Barack ObamaとMichelle Obamasの顔画像から異性かつ年齢が近いことがわかるので**夫婦**の関係と推定

SREの研究は、自然言語処理分野や画像認識分野で一部行われているものの研究例は少ない。
特にマルチモーダルなSREについては**先行研究がほとんど存在しない**。

【Text】

- ・ micropostsから社会的関係性を予測

Du et al 2019, “*Extracting Deep Personae Social Relations in Microblog Posts*”, IEEE Access 8

【Image】

- ・ 二人の顔画像から社会的関係（親しみ、支配的、優しさ etc.）を予測

Zhang et al 2015, “*Learning Social Relation traits from Face Images*”, In ICV, 3631-3639

4つの中国の古典的小説とそのTVシリーズから二人の顔画像と（二人の社会的関係が推察できるような）テキスト，二人の社会的な関係性がアノテーションされたデータセットを構築

Sentence: テキスト
Head entity(h): 一人目の顔の座標と名前
Tail entity(t): 二人目の顔の座標と名前
g_h: 一人目が写っている画像
g_t: 二人目が写っている画像
r: 社会的関係性


Tuples (s, h, t, g_h, g_t) を作成

扈三娘来到筵前，宋江亲自与他陪话，说道：“我这兄弟王英，虽有武艺，不及贤妹。”
Hu Sanniang walked to the banquet. Song Jiang said to her, “My brother Wang Ying, although he has martial arts, he is still inferior to you.”

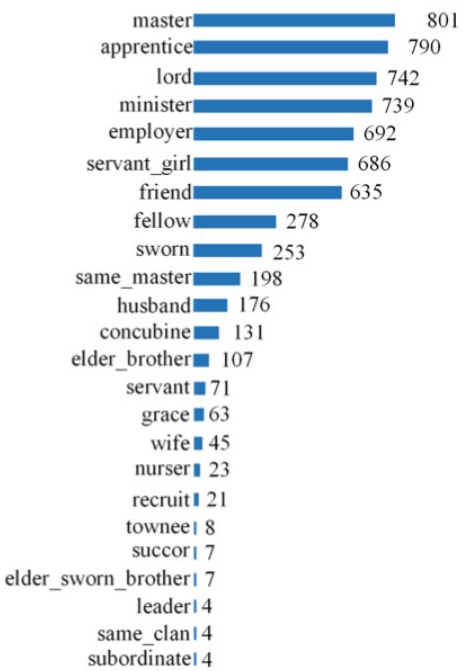


- Gound Truth: <A, wife, B>
- Proto(BERT): <A, sworn, B> ❌
- FL-MSRE: <A, wife, B> ✅

锦儿道：“正在五岳楼下来，撞见个诈玳不及的把娘子拦住了，不肯放！”林冲慌忙道：“却再来看望师兄，休怪，休怪。”林冲别了鲁智深。
Jiner said, “When we were coming down Wuyue Tower, we met a treacherous person who stopped your lady and refused to let her go!” Lin Chong hurriedly said, “I will visit brother next time, don’ t be angry.” Lin Chong said good bye to Lu Zhishen.



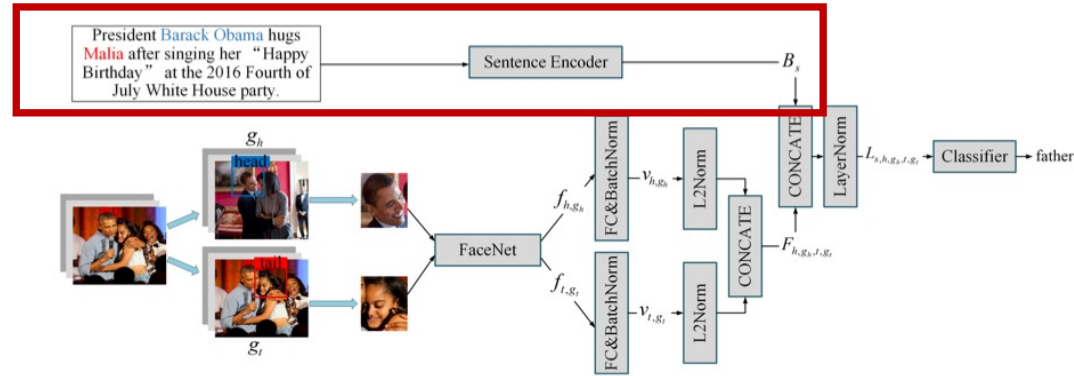
- Gound Truth: <C, sworn, D>
- Proto(BERT): <C, wife, D> ❌
- FL-MSRE: <C, sworn, D> ✅



Datasets	#rel	#char	#triple	#sen	#img
DRC-TF	9	47	59	1828	560
OM-TF	15	43	54	1489	1178
FC-TF	24	121	166	6485	3716

Proposed Approach FL-MSRE

- Sentence Encoder



1. テキストに含まれる単語をトークン化する

$$S = \{w_1, \dots, w_m\}$$

2. BERTで用いる特別なトークンを付与する

$$S' = \{[CLS], w_1, \dots, w_m, [SEP]\}$$

3. 事前学習済みBERTを用いて S' を $m+2$ 個の768次元ベクトルに変換する

$$V_{[CLS]}, V_{w_1}, \dots, V_{[SEP]}$$

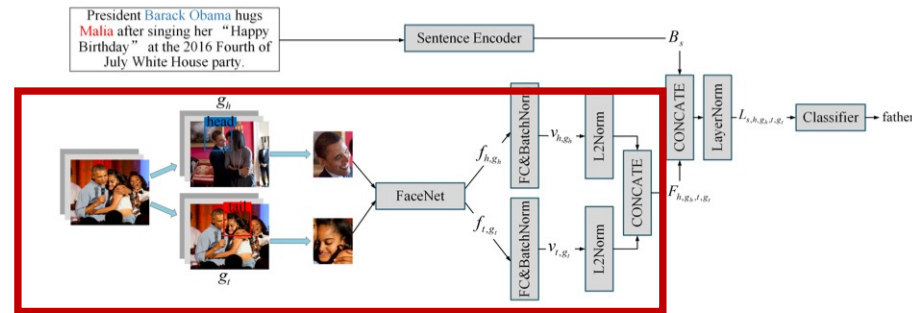
4. 文章の分散表現に有効である $V_{[CLS]}$ を用いて, 最終的なベクトルである B_s を出力する

$$B_s = \tanh(W_1 V_{[CLS]} + b_1)$$

ただし, $W_1 \in \mathbb{R}^{768 \times 768}$, $b_1 \in \mathbb{R}^{768}$ はパラメータ

Proposed Approach FL-MSRE

- Face Encoder



1. 一人目(h)と二人目(t)の顔位置の座標である b_h, b_t を用いて, それぞれの写真 g_h, g_t から顔画像だけを取り出す. 次にFaceNetを用いて顔の特徴を1792次元のベクトルに変換する.

$$f_{h,g_h} = \phi(\text{crop}(g_h, b_h))$$
$$f_{t,g_t} = \phi(\text{crop}(g_t, b_t))$$

2. f_{h,g_h} と f_{t,g_t} をそれぞれについて全結合し, バッチ正則化する

$$v_{h,g_h} = \text{BatchNorm}(W_2 f_{h,g_h} + b_2)$$
$$v_{t,g_t} = \text{BatchNorm}(W_3 f_{t,g_t} + b_3)$$

3. v_{h,g_h} と v_{t,g_t} をそれぞれについてL2正則化して, 結合することで最終ベクトルとする

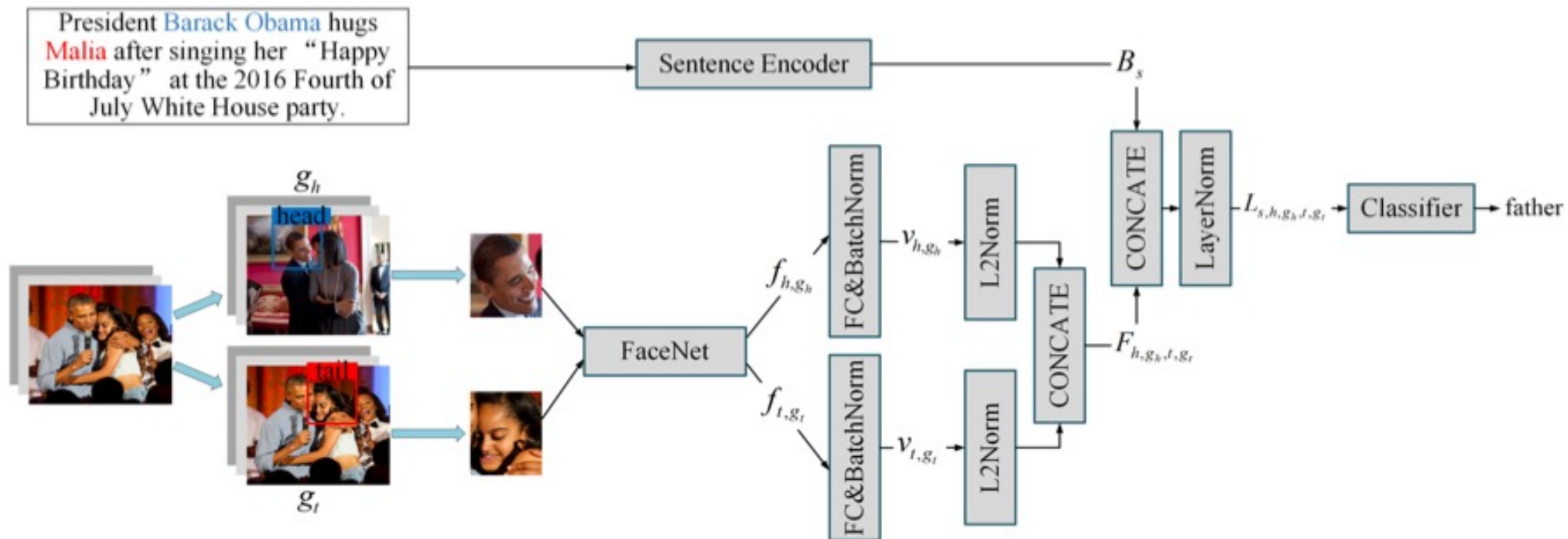
$$F_{h,t,g_h,g_t} = [\text{L2Norm}(v_{h,g_h}); \text{L2Norm}(v_{t,g_t})]$$

Proposed Approach FL-MSRE

- Cross-modality Encoder

1. Sentence EncoderとFace Encoderの出力を結合してL2正則化する.

$$L_{s,h,t,g_h,g_t} = \text{LayerNorm}[B_s; F_{h,t,g_h,g_t}]$$



Proposed Approach FL-MSRE

- Prototypical Network(N way K shot setting)

Notation

- $S = \{(s_{11}, h_{11}, t_{11}, g_{h,11}, g_{t,11}, r_1), \dots, (s_{1K}, h_{1K}, t_{1K}, g_{h,1K}, g_{t,1K}, r_1),$
 \vdots
 $(s_{N1}, h_{N1}, t_{N1}, g_{h,N1}, g_{t,N1}, r_N), \dots, (s_{NK}, h_{NK}, t_{NK}, g_{h,NK}, g_{t,NK}, r_N)\}$: support set
- $q = (s, h, t, g_h, g_t)$: query tuple
- $P_m(S)$: **prototype** of social relation m
- $d(A,B)$: distance function

$$P_m(S) = \frac{1}{K} \sum_{i=1}^K L_{s_{mi}, h_{mi}, g_{h,mi}, t_{mi}, g_{t,mi}} \quad \cdot \cdot \cdot (1)$$

$$\Pr(r_m \mid q) = \frac{\exp(-d(L_{s,h,g_h,t,g_t}, P_m(S)))}{\sum_{i=1}^N \exp(-d(L_{s,h,g_h,t,g_t}, P_i(S)))} \quad \cdot \cdot \cdot (2)$$

比較対象：SREタスクのためにfine-tuneしたBERTモデル

2つの顔画像収集法：

異なる画像から二人の顔画像を抽出→二人の年齢や性別といった**性質のみ**を考慮
同じ画像から二人の顔画像を抽出→二人が一緒にいる状況下での**表情なども**考慮

クラスの分割方法：(データセット名 training : validation : test)

FC-TF 14:5:5

DRC-TF 3:3:3

OM-TF 5:5:5

Experiments

• Result

すべての実験においてFL-MSREはBERTプロトタイプの精度を上回る。
特に、二人の性質(年齢や性別)と社会的関係性の関連が大きいOM-TFでは大きく改善

異なる画像から顔を抽出したモデルは同じ画像から抽出したモデルと同等の性能を発揮

異なるデータセットから学習した場合でも、一定の性能を発揮し、モデルの安定性を示した
BERTプロトタイプにとっては異なる文体で学習したモデルでテストするので厳しい
一方で、FL-MSREは二人の性質というデータセット間で不変の情報をを用いているため一定の精度を発揮できる

Methods	FC-TF				DRC-TF		OM-TF	
	5 way 1 shot	5 way 3 shot	3 way 1 shot	3 way 3 shot	3 way 1 shot	3 way 3 shot	3 way 1 shot	3 way 3 shot
Proto (BERT)	73.93±0.01	75.03±0.35	83.84±0.37	86.75±0.32	40.00±0.11	51.29±0.16	46.28±0.15	48.35±0.43
FL-MSRE (same)	79.05±0.10	79.59±0.14	87.57±0.41	90.24±0.32	58.44±0.21	71.96±0.49	54.16±0.22	61.67±0.67
FL-MSRE (different)	78.61±0.48	79.95±0.23	87.64±0.15	89.98±0.17	62.03±0.24	77.29±0.54	54.53±0.68	62.30±0.39

Methods	DRC-TF (trained on OM-TF)		OM-TF (trained on DRC-TF)	
	3 way 1 shot	3 way 3 shot	3 way 1 shot	3 way 3 shot
Proto (BERT)	38.38±0.16	40.74±0.18	37.27±0.69	44.09±0.69
FL-MSRE (same)	50.69±0.81	63.83±0.42	55.82±0.76	58.50±0.45
FL-MSRE (different)	50.34±0.15	64.59±0.56	52.76±0.91	55.43±1.25

SREの論文を全て読む

Prototypical Networkを実装する

Few-shot Learningの様々な手法についての論文を読む