Multinomial Adversarial Networks *for Multi-Domain Text Classification*

NAACL HLT
New Orleans, Louisiana | June 1-June 6, 2018

Xilun Chen and Claire Cardie
{xlchen, cardie}@cs.cornell.edu
Cornell CIS Computer Science | Cornell University

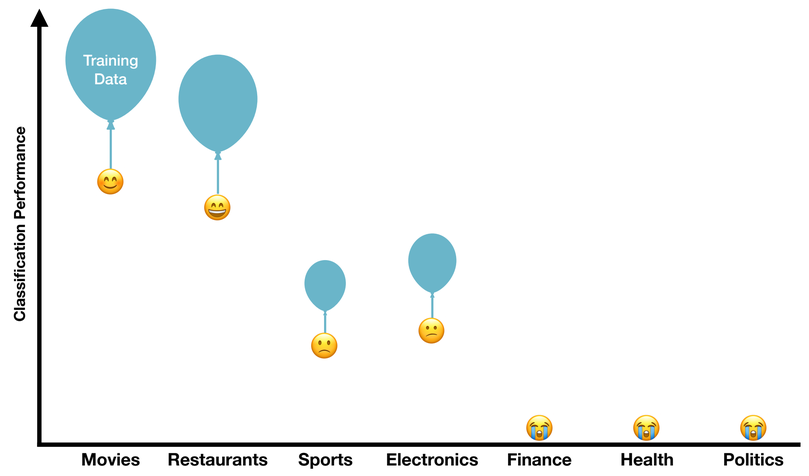
Multi-Domain Text Classification

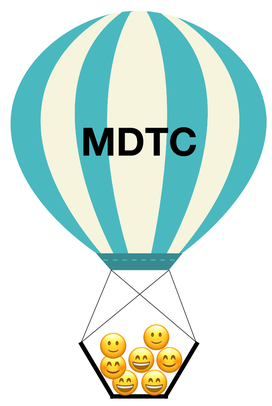
Text Classification is domain dependent

Automobile The car **runs fast** and handles well.

Electronics The battery of the camera **runs fast**.

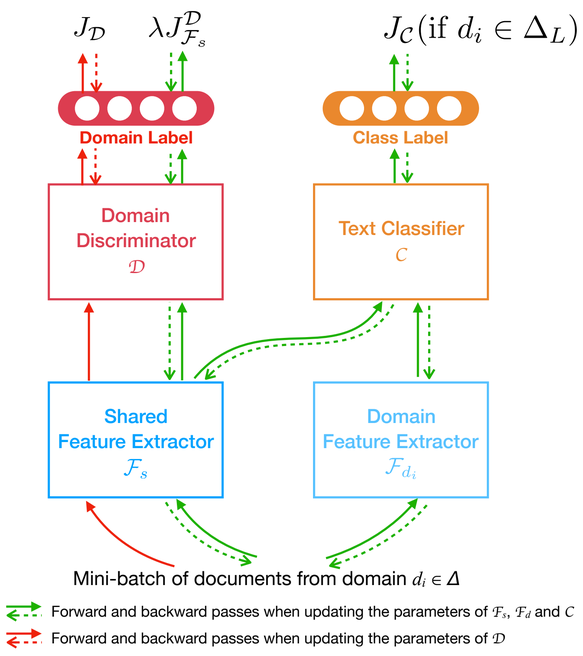
Domains are not born equal





MDTC utilizes all available training data to improve the **overall** performance across all (labeled and unlabeled) domains

MAN for MDTC



Algorithm 1 MAN Training

Require: labeled corpus $\mathbb{X}$; unlabeled corpus $\mathbb{U}$; Hyperparameter $\lambda > 0$, $k \in \mathbb{N}$

repeat
  ▷ $\mathcal{D}$ iterations
  for $diter = 1$ to k **do**
    $l_{\mathcal{D}} = 0$
    for all $d \in \Delta$ **do** ▷ For all N domains
      Sample a mini-batch $\boldsymbol{x} \sim \mathbb{U}_d$
      $\boldsymbol{f}_s = \mathcal{F}_s(\boldsymbol{x})$ ▷ Shared feature vector
      $l_{\mathcal{D}} \mathrel{+}= J_{\mathcal{D}}(\mathcal{D}(\boldsymbol{f}_s); d)$ ▷ Accumulate $\mathcal{D}$ loss
    Update $\mathcal{D}$ parameters using $\nabla l_{\mathcal{D}}$
  ▷ Main iteration
  $loss = 0$
  for all $d \in \Delta_L$ **do** ▷ For all labeled domains
    Sample a mini-batch $(\boldsymbol{x}, \boldsymbol{y}) \sim \mathbb{X}_d$
    $\boldsymbol{f}_s = \mathcal{F}_s(\boldsymbol{x})$
    $\boldsymbol{f}_d = \mathcal{F}_d(\boldsymbol{x})$ ▷ Domain feature vector
    $loss \mathrel{+}= J_{\mathcal{C}}(\mathcal{C}(\boldsymbol{f}_s, \boldsymbol{f}_d); \boldsymbol{y})$ ▷ Compute $\mathcal{C}$ loss
  for all $d \in \Delta$ **do** ▷ For all N domains
    Sample a mini-batch $\boldsymbol{x} \sim \mathbb{U}_d$
    $\boldsymbol{f}_s = \mathcal{F}_s(\boldsymbol{x})$
    $loss \mathrel{+}= \lambda \cdot J^{\mathcal{D}}_{\mathcal{F}_s}(\mathcal{D}(\boldsymbol{f}_s); d)$ ▷ Domain loss of $\mathcal{F}_s$
  Update $\mathcal{F}_s$, $\mathcal{F}_d$, $\mathcal{C}$ parameters using $\nabla loss$
until convergence

The non-cooperative coupling of $\mathcal{F}_s$ and $\mathcal{D}$ form a *Multinomial Adversarial Network*.

$\mathcal{D}$ attempts to identify the domain of a sample using the shared features.

$\mathcal{F}_s$ learns domain-invariant features by learning to stop leaking domain information to $\mathcal{D}$.

MAN Glossary

Δ_L The set of all *labeled* domains which have some annotated data.

Δ_U The set of all *unlabeled* domains which have *no* annotated data.

Δ The set of all domains: $\Delta = \Delta_L \cup \Delta_U$.

$\mathbb{X}_i$ The labeled data for domain $d_i \in \Delta_L$.

$\mathbb{U}_i$ The unlabeled corpus for domain $d_i \in \Delta$.

$\mathcal{F}_s$ The *shared feature extractor* that extracts domain-invariant features.

$\mathcal{F}_d$ The set of *domain feature extractors* that extracts domain-specific features.

$\mathcal{C}$ The *text classifier*.

$\mathcal{D}$ The adversarial *domain discriminator*.

$J_{\mathcal{C}}$ The cost $\mathcal{C}$ minimizes.

$J_{\mathcal{D}}$ The cost $\mathcal{D}$ minimizes.

$J^{\mathcal{D}}_{\mathcal{F}_s}$ The domain cost of $\mathcal{F}_s$ that is anticorrelated to $J_{\mathcal{D}}$.

$J_{\mathcal{F}_s}$ The cost of $\mathcal{F}_s$: $J_{\mathcal{F}_s} = J_{\mathcal{C}} + \lambda J^{\mathcal{D}}_{\mathcal{F}_s}$.

λ A hyperparameter balancing the classification loss and the domain loss.

k A hyperparameter synchronizing the training of $\mathcal{D}$ and the rest of MAN.

MAN-NLL The MAN variant with the Negative Log Likelihood loss.

MAN-L2 The MAN variant with the Least Square loss.

MAN Theory

- **Summary:** We show that MAN, as a versatile machine learning framework, directly minimizes the *f*-divergence among ***multiple*** distributions.
- ***f*-divergence** is a family of metrics measuring the difference between probability distributions. Many common divergences, such as KL-divergence, total variation divergence, are special cases of *f*-divergence.
- Nowozin et al. (2016) proved that standard adversarial nets are minimizers of various *f*-divergence metrics between *two* distributions, depending on the choice of loss function.
- MAN is hence a generalization of the impactful (binomial) adversarial networks to multiple distributions.

CAUTION MATH AHEAD

Let the distribution of the shared features $\mathbf{f}$ for instances in each domain $d_i \in \Delta$ be:

$$P_i(\mathbf{f}) \triangleq P(\mathbf{f} = \mathcal{F}_s(x) | x \in d_i) \tag{1}$$

CAUTION MATH AHEAD

We consider two MAN variants with the NLL and L2 loss, respectively:

$$J^{NLL}_{\mathcal{D}} = -\sum_{i=1}^{N} \mathop{\mathbb{E}}_{\mathbf{f} \sim P_i} [\log \mathcal{D}_i(\mathbf{f})] \tag{2}$$

$$J^{L2}_{\mathcal{D}} = \sum_{i=1}^{N} \mathop{\mathbb{E}}_{\mathbf{f} \sim P_i} \left[\sum_{j=1}^{N} (\mathcal{D}_j(\mathbf{f}) - \mathbb{1}_{\{i=j\}})^2 \right] \tag{3}$$

Lemma 1. *For any fixed $\mathcal{F}_s$, with either NLL or L2 loss, the optimum $\mathcal{D}^*$ is:*

$$\mathcal{D}^*_i(\mathbf{f}) = \frac{P_i(\mathbf{f})}{\sum_{j=1}^{N} P_j(\mathbf{f})} \tag{4}$$

Theorem 1. *Let $\overline{P} = \frac{\sum_{i=1}^{N} P_i}{N}$. When $\mathcal{D}$ is trained to its optimality, if $\mathcal{D}$ adopts the* NLL loss*:*

$$J^{\mathcal{D}}_{\mathcal{F}_s} = -\min_{\theta_{\mathcal{D}}} J_{\mathcal{D}} = -J_{\mathcal{D}^*} = -N \log N + N \cdot JSD(P_1, P_2, \ldots, P_N) = -N \log N + \sum_{i=1}^{N} KL(P_i \| \overline{P})$$

where $JSD(\cdot)$ is the generalized Jensen-Shannon Divergence (Lin, 1991), defined as the average Kullback-Leibler divergence of each P_i to the centroid $\overline{P}$ (Aslam and Paul, 2007).

Theorem 2. *If $\mathcal{D}$ uses the* L2 loss*:*

$$J^{\mathcal{D}}_{\mathcal{F}_s} = \sum_{i=1}^{N} \mathop{\mathbb{E}}_{\mathbf{f} \sim P_i} \left[\sum_{j=1}^{N} (\mathcal{D}^*_j(\mathbf{f}) - \frac{1}{N})^2 \right] = \frac{1}{N} \sum_{i=1}^{N} \chi^2_{Neyman}(P_i \| \overline{P})$$

where $\chi^2_{Neyman}(\cdot\|\cdot)$ is the Neyman χ^2 divergence (Nielson and Nock, 2014).

Consequently, by the non-negativity and joint convexity of the f-divergence:

Corollary 1. *The optimum of $J^{\mathcal{D}}_{\mathcal{F}_s}$ is $-N \log N$ when using NLL loss, and 0 for the L2 loss. The optimum value above is achieved if and only if $P_1 = P_2 = \cdots = P_N = \overline{P}$ for either loss.*

Experiments

	Book	DVD	Elec.	Kit.	Avg.
Domain-Specific Models Only					
LS	77.80	77.88	81.63	84.33	80.41
SVM	78.56	78.66	83.03	84.74	81.25
LR	79.73	80.14	84.54	86.10	82.63
MLP	81.70	81.65	85.45	85.95	83.69
Shared Model Only					
LS	78.40	79.76	84.67	85.73	82.14
SVM	79.16	80.97	85.15	86.06	82.83
LR	80.05	81.88	85.19	86.56	83.42
MLP	82.40	82.15	85.90	88.20	84.66
MAN-L2-MLP	82.05	83.45	86.45	88.85	85.20
MAN-NLL-MLP	81.85	83.10	85.75	89.10	84.95
Shared-Private Models					
RMTL[1]	81.33	82.18	85.49	87.02	84.01
MTLGraph[2]	79.66	81.84	83.69	87.06	83.06
CMSC-LS[3]	82.10	82.40	86.12	87.56	84.55
CMSC-SVM[3]	82.26	83.48	86.76	88.20	85.18
CMSC-LR[3]	81.81	83.73	86.67	88.23	85.11
SP-MLP	82.00	**84.05**	86.85	87.30	85.05
MAN-L2-SP-MLP	82.46 (±0.25)	83.98 (±0.17)	**87.22*** (±0.04)	88.53 (±0.19)	85.55* (±0.07)
MAN-NLL-SP-MLP	**82.98*** (±0.28)	84.03 (±0.16)	87.06 (±0.23)	**88.57*** (±0.15)	**85.66*** (±0.14)

[1] Evgeniou and Pontil (2004)
[2] Zhou et al. (2011)
[3] Wu and Huang (2015)

Table 1: MDTC results on the Amazon dataset. Models in bold are ours while the performance of the rest is taken from Wu and Huang (2015). Numbers in parentheses indicate standard errors, calculated based on 5 runs. Bold numbers indicate the highest performance in each domain, and $*$ shows statistical significance ($p < 0.05$) over CMSC under a one-sample T-Test.

MDTC Experiments

Amazon Dataset (Table 1)

✓ Widely adopted
– 2000 samples/domain
– 5-fold cross-validation
✗ 4 domains
✗ Preprocessed to bag-of-word features (no raw text, no word order information)

FDU-MTL Dataset (Table 3)

✗ Less reported results
– ~2000 samples/domain
– Pre-split train/dev/test sets
✓ 16 domains
✓ Original texts available

	books	elec.	dvd	kitchen	apparel	camera	health	music	toys	video	baby	magaz.	softw.	sports	IMDb	MR	Avg.
Domain-Specific Models Only																	
BiLSTM	81.0	78.5	80.5	81.2	86.0	86.0	78.7	77.2	84.7	83.7	83.5	91.5	85.7	84.0	85.0	74.7	82.6
CNN	85.3	87.8	76.3	84.5	86.3	89.0	87.5	81.5	87.0	82.3	82.5	86.8	87.5	85.3	83.3	75.5	84.3
Shared Model Only																	
FS-MTL	82.5	85.7	83.5	86.0	84.5	86.5	88.0	81.2	84.5	83.7	88.0	92.5	86.2	85.5	82.5	74.7	84.7
MAN-L2-CNN	**88.3**	88.3	87.8	88.5	85.3	90.5	**90.8**	85.3	89.5	89.0	89.5	91.3	88.3	89.5	**88.5**	73.8	87.7
MAN-NLL-CNN	88.0	87.8	87.3	88.5	86.3	90.8	89.8	84.8	89.3	89.3	87.8	91.8	90.0	**90.3**	87.3	73.5	87.6
Shared-Private Models																	
ASP-MTL	84.0	86.8	85.5	86.2	87.0	89.2	88.2	82.5	88.0	84.5	88.2	92.2	87.2	85.7	85.5	**76.7**	86.1
MAN-L2-SP-CNN	87.6* (0.2)	87.4 (1.0)	88.1* (0.4)	89.8* (0.4)	**87.6** (0.7)	**91.4*** (0.4)	89.8* (0.3)	**85.9*** (0.1)	90.0* (0.1)	89.5* (0.2)	90.0 (0.6)	92.5 (0.5)	90.4* (0.4)	89.0* (0.4)	86.6 (0.5)	76.1 (0.5)	88.2* (0.1)
MAN-NLL-SP-CNN	86.8* (0.4)	**88.8** (0.6)	**88.6*** (0.4)	**89.9*** (0.4)	**87.6** (0.4)	90.7 (0.4)	89.4 (0.3)	85.5* (0.1)	**90.4*** (0.2)	**89.6*** (0.3)	**90.2** (0.6)	**92.9** (0.4)	**90.9*** (0.7)	89.0* (0.2)	87.0* (0.1)	**76.7** (0.8)	**88.4*** (0.1)

Table 3: Results on the FDU-MTL dataset. Bolded models are ours, while the rest is from Liu et al. (2017). Highest performance is each domain is highlighted. For our full MAN models, standard errors are shown in parenthese and statistical significance ($p < 0.01$) over ASP-MTL is indicated by *.

Experiments on Unlabeled Domains

- Baselines: Multi-Source Domain Adaptation Methods
- Dataset: Amazon (4 domains)
- 3 labeled (source) domains
- 1 unlabeled (target) domain
- MAN achieves SToA performance
- MAN also has the potential to handle >1 unlabeled domains

Target Domain	Book	DVD	Elec.	Kit.	Avg.
MLP	76.55	75.88	84.60	85.45	80.46
mSDA[1]	76.98	78.61	81.98	84.26	80.46
DANN[2]	77.89	78.86	84.91	86.39	82.01
MDAN (H-MAX)[3]	78.45	77.97	84.83	85.80	81.76
MDAN (S-MAX)[3]	**78.63**	80.65	**85.34**	86.26	**82.72**
MAN-L2-SP-MLP	78.45	81.57	83.37	85.57	82.24
MAN-NLL-SP-MLP	77.78	**82.74**	83.75	**86.41**	82.67

[1] Chen et al. (2012)
[2] Ganin et al. (2016)
[3] Zhao et al. (2017)

Table 2: Results on unlabeled domains. Models in bold are our models while the rest is taken from Zhao et al. (2017). Highest domain performance is shown in bold.