MS-DIAL: Multi-Source Domain Alignment Layers for Unsupervised Domain Adaptation

Lucas Fernando Alvarenga e Silva

Instituto de Ciência e Tecnologia
Universidade Federal de São Paulo – UNIFESP
12247-014, São José dos Campos, SP – Brazil
Email: e.lucas@unifesp.br

Jurandy Almeida

Instituto de Ciência e Tecnologia
Universidade Federal de São Paulo – UNIFESP
12247-014, São José dos Campos, SP – Brazil
Email: jurandy.almeida@unifesp.br

***Abstract*—In general, deep neural networks trained on a given labeled dataset are expected to produce equivalent results when tested on a new unlabeled dataset. However, data are generally collected by different devices or under varying conditions and thus they often are not part of a same domain, yielding poor results. This is due to the domain shift between data distributions and has been the goal of a research area known as unsupervised domain adaptation. Many prior works have been designed to transfer knowledge between two domains: one source to one target. Since data may be taken from different sources and with different distributions, multi-source domain adaptation has received increasing attention. This paper presents the Multi-Source DomaIn Alignment Layers (MS-DIAL), which reduce the domain shift between multiple sources and a given target by embedding domain alignment layers in any given network. Except for the embedded layers, all the other network parameters are shared among all domains, saving processing time and memory usage. Experiments were performed on digit and object recognition tasks with five public datasets widely used to evaluate domain adaptation methods. Results show that the proposed method is promising and outperforms state-of-the-art approaches.**

I. Introdução

Nos últimos anos, o campo da visão computacional tem alcançado resultados surpreendentes em uma variedade de problemas desafiadores, em especial, na classificação de imagens em grandes bases de dados propostas para tarefas amplamente consideradas como difíceis, como a ImageNet [1]. Esses resultados têm sido obtidos a partir da utilização de métodos de aprendizado de maquina, em especial, graças aos avanços significativos introduzidos pela aprendizagem profunda (do inglês, *deep learning*) com as redes neurais convolucionais (do inglês, *convolutional neural networks* – CNNs).

Em geral, é esperado que modelos treinados em bases de dados anotadas (*i.e.*, conjunto de treinamento) produzam resultados equivalentes quando aplicados à novos dados não-anotados (*i.e.*, conjunto de teste). Tal premissa parte do pressuposto de que os dados anotados usados no treinamento e os dados não-anotados usados no teste pertencem ao mesmo domínio, isto é, apresentem uma mesma distribuição de probabilidade. Porém, na prática, os dados são geralmente coletados por dispositivos diferentes ou sob condições variadas e, portanto, não necessariamente fazem parte de um mesmo domínio, o que pode produzir resultados insatisfatórios. Isso acontece

Figura 1. Imagens do conjunto de dados Office-Home, em que as linhas representam, respectivamente, os domínios: arte, clipart, produto e mundo real; e nas colunas estão representadas algumas categorias do conjunto de dados, como: colher, pia, xícara, caneta e faca. Adaptado de Venkateswara *et al.* [2].

devido a mudança de domínio (do inglês, *domain shift*) que há entre as distribuições de dados e é objeto de pesquisa do campo denominado adaptação de domínio não-supervisionada (do inglês, *unsupervised domain adaptation* – UDA).

A grosso modo, as soluções existentes em UDA se enquadram em duas vertentes bem definidas: (*i*) as que exploram características invariantes entre domínios (do inglês, *domain invariant features*) ou (*ii*) as que reduzem a discrepância entre as distribuições de dados [3]. A maioria dos trabalhos anteriores considera a transferência de conhecimento entre dois domínios: um fonte (do inglês, *source*) e um alvo (do inglês, *target*). Todavia, na prática, os dados são normalmente provenientes de várias fontes e com distribuições distintas, como ilustrado na Figura 1, na qual imagens de uma mesma classe são coletadas de diversos sites da Internet e adquiridas sob condições distintas [4]. Nesse cenário, é comum agrupar os dados de vários domínios-fonte em um único conjunto e, em seguida, aplicar métodos projetados para lidar com a adaptação de um único domínio-fonte para um único domínio-alvo [5]. Entretanto, essa abordagem geralmente não produz resultados satisfatórios, uma vez que os domínios-fonte não necessariamente contribuem da mesma maneira para o processo de transferência de conhecimento para o domínio-alvo [6].

O problema de adaptação de domínio de várias fontes (do inglês, *multi-source domain adaptation* – MSDA) é mais

complexo e desafiador, já que pode haver um deslocamento entre as distribuições dos domínios-fonte, bem como eles podem fornecer informações complementares para o processo de transferência de conhecimento para o domínio-alvo. Além desses fatores, também podem ser encontradas classes diferentes entre os domínios-fonte (do inglês, *category shift*) [4].

Este trabalho contribui com uma nova proposta para o problema de MSDA, denominada MS-DIAL (do inglês, *multi-source domain alignment layers*), a qual reduz a discrepância entre as distribuições dos domínios-fonte e do domínio-alvo a partir da inserção de camadas de alinhamento de domínio em diversos níveis da rede. Nessa abordagem, o nível de alinhamento das distribuições é ajustado de forma automática pela rede por meio do uso de parâmetros aprendíveis em tempo de treinamento. Para uma melhor separação entre categorias do domínio-alvo, a entropia das predições obtidas para amostras do lote do domínio-alvo é usada como medida de erro para o otimizador, buscando assim ajustar os parâmetros da rede às características extraídas dos dados do domínio-alvo.

Experimentos foram realizados em cinco conjuntos de dados públicos usados para avaliar métodos de UDA: MNIST [7], MNIST-M [8], SVHN [9] e *Synthetic Digits* [10], que foram propostos para tarefas de reconhecimento de dígitos; e Office-Home [2], que aborda a tarefa de reconhecimento de objetos. Os resultados obtidos demonstram que o método proposto é eficaz, superando abordagens do estado da arte.

O restante deste trabalho está organizado da seguinte maneira. A Seção II discute trabalhos relacionados. A Seção III introduz o MS-DIAL e mostra como ele pode ser usado para lidar com tarefas de MSDA. A Seção IV apresenta o protocolo experimental e a comparação dos resultados do MS-DIAL com outros métodos. Por fim, conclusões e direções para trabalhos futuros são oferecidos na Seção V.

II. Trabalhos Relacionados

A adaptação de domínio não-supervisionada de única fonte para único alvo (do inglês, *single-source to single-target domain adaptation*) conta com um domínio-fonte anotado e um domínio-alvo não-anotado, e tem por objetivo adaptar um modelo treinado em dados anotados do domínio-fonte para reconhecer instâncias provenientes de dados não-anotados do domínio-alvo. É um problema desafiador e com diversas propostas de solução, algumas com modelos rasos e atualmente tem se voltando ao uso de redes neurais profundas. Os métodos rasos se baseiam na redução da discrepância entre domínios e buscam obter características invariantes, como a análise de componentes de transferência (do inglês, *transfer component analysis* – TCA) [11] e a incorporação de correspondência de distribuição (do inglês, *distribution-matching embedding* – DME) [12]. Em trabalhos recentes, redes neurais profundas, normalmente submetidas a um treinamento adversário, têm sido usadas em duas vertentes: (*i*) para mapear dados de ambos os domínios em uma distribuição comum ou (*ii*) para distinguir amostras provenientes de domínios fonte e alvo. Alguns exemplos são as redes neurais de domínio adversário (do inglês, *domain-adversarial neural networks* – DANN) [10] e a discrepância média máxima ponderada (do inglês, *weighted maximum mean discrepancy* – WMMD) [13]. Outras vertentes, como a inserção de camadas de alinhamento de domínio [3], [14], estão intimamente relacionadas a este trabalho.

Já a adaptação de domínio não-supervisionada de várias fontes para único alvo (do inglês, *multi-source to single-target domain adaptation*) é ainda mais desafiadora. É um problema emergente nos últimos anos e atualmente existem algumas propostas de solução, como a correspondência de momento para adaptação de domínio de várias fontes (do inglês, *moment matching for multi-source domain adaptation* – M3SDA) [15], a rede de cauda profunda (do inglês, *deep cocktail network* – DCTN) [4], a rede de agregação de domínio (do inglês, *domain aggregation network* – DARN) [6], a rede de correspondência de vários domínios (do inglês, *multiple domain matching network* – MDMN) [16], as redes adversárias de domínio de várias fontes (do inglês, *multisource domain adversarial networks* – MDAN) [17] e a adaptação de domínio de destilação de múltiplas fontes (do inglês, *multi-source distilling domain adaptation* – MDDA) [18]. Em geral, as abordagens existentes baseiam-se em redes neurais de múltiplos fluxos na qual a quantidade de classificadores e/ou extratores de características é ajustada proporcionalmente à quantidade de domínios. M3SDA [15] é uma rede que contém um único extrator de características comum a todos os domínios, porém, com um classificador para cada domínio-fonte, cujas saídas são agrupadas por média ponderada. De maneira similar, DCTN [4] e DARN [6] usam um único conjunto de pesos que é compartilhado pelos extratores de características de todos os domínios. Contudo, para realizar a adaptação de domínios, DCTN adota medidas de perplexidade e discriminadores de domínio, enquanto DARN emprega módulos de discrepância. MDDA [18] é uma rede adversária composta por um extrator de características e um classificador para cada domínio-fonte, cuja predição final é dada pela média ponderada das predições de todos os domínios, na qual os pesos são obtidos a partir de métricas de discriminação de domínio.

III. MS-DIAL: Multi-Source DomaIn Alignment Layers

Sejam os conjuntos de dados anotados $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_M$ referentes a M domínios-fonte que compartilham o mesmo conjunto de rótulos $\mathcal{Y}$ com um conjunto de dados não-anotados $\mathcal{T}$ referente ao único domínio-alvo. Suponha que cada domínio-fonte $\mathcal{S}_i = \{(\mathbf{x}_i^j, \mathbf{y}_i^j)\}_{j=1}^{N_i}$ corresponde a um conjunto de tuplas que associa dados observados $X_i = \{\mathbf{x}_i^j\}_{j=1}^{N_i}$ a seus respectivos rótulos $Y_i = \{\mathbf{y}_i^j\}_{j=1}^{N_i}$, os quais foram extraídos da distribuição-fonte $p_i(\mathbf{x}, \mathbf{y})$, em que N_i é o número de amostras em $\mathcal{S}_i$. Como os rótulos do domínio-alvo $\mathcal{T}$ não são conhecidos, assuma que ele seja formado por dados $X_T = \{\mathbf{x}_T^j\}_{j=1}^{N_T}$ extraídos da distribuição-alvo $p_T(\mathbf{x}, \mathbf{y}), \mathbf{y} \in \mathcal{Y}$, em que N_T é o número de amostras em $\mathcal{T}$. Assim, o problema de MSDA consiste em encontrar um conjunto de parâmetros $\theta \in \Theta$ para uma rede neural de forma que suas predições para o conjunto de rótulos $Y_T = \{\mathbf{y}_T^j\}_{j=1}^{N_T}$ ainda não conhecidos do domínio-alvo $\mathcal{T}$ sejam as melhores possíveis.

Em geral, trabalhos anteriores de MSDA usam topologias de rede com múltiplos fluxos, normalmente com um fluxo independente para cada domínio, algumas com um conjunto de parâmetros θ diferente para cada fluxo e, assim, cada domínio tem o seu próprio extrator de características e classificador; e outras com um conjunto de parâmetros θ compartilhado entre os fluxos, geralmente, pelos extratores de características de todos os domínios, mas cada um tendo o seu próprio classificador. Diferente dessas abordagens, na topologia de rede desenvolvida neste trabalho, o conjunto de parâmetros θ é compartilhado entre os fluxos de todos os domínios, tanto pelos extratores de características quanto pelos classificadores, exceto nas camadas de alinhamento de domínio com várias fontes (do inglês, *multi-source domain alignment layers* – MS-DIAL), como ilustrado na Figura 2. Durante a fase de treinamento, as amostras contidas nos mini-lotes são agrupadas de acordo com o domínio a qual pertencem e cada grupo de amostras segue um caminho diferente, sendo encaminhado a uma camada de normalização de lote associada ao seu respectivo domínio. Dessa forma, é possível utilizar uma única instância da topologia de rede e, portanto, um mesmo conjunto de parâmetros θ para todos os domínios, reduzindo assim o custo computacional e a utilização de memória.

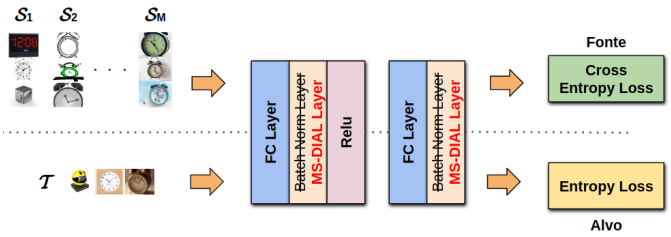

Figura 2. Exemplo de uma topologia de rede adaptada com camadas MS-DIAL para realizar a MSDA dos domínios-fonte $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_M$ para o domínio-alvo $\mathcal{T}$. Para isso, as camadas de normalização de lote foram substituídas por camadas MS-DIAL, mantendo, assim, todas as demais camadas compartilhadas entre todos os domínios $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_M, \mathcal{T}$.

A. Preditores de Fonte e Alvo

O ponto de partida para a abordagem proposta neste trabalho são as camadas de alinhamento de domínio (do inglês, *domain alignment layers* – DIAL) [3], que reduzem a discrepância entre as distribuições dos domínios fonte e alvo ao longo do fluxo de dados na topologia de rede, levando as diferentes distribuições a uma mesma distribuição de referência. Tais camadas foram projetadas para adaptação de domínio de única fonte para único alvo. Inicialmente, as amostras $x \subseteq \{X_S \cup X_T\}$ dos mini-lotes de entrada são divididas em dois grupos: (*i*) amostras do domínio-fonte $x_S \subseteq X_S$ e (*ii*) amostras do domínio-alvo $x_T \subseteq X_T$. Em seguida, cada grupo de amostras é encaminhado para uma camada de normalização de lote [19] associada ao seu respectivo domínio, as quais ajustam cada uma das distribuições dos domínios a uma distribuição de referência, porém, sem realizar transformações afins, como mostrado na Equação 1. Por fim, para controlar a sobreposição de todos os domínios, foram inseridos dois parâmetros aprendíveis pela rede em tempo de treinamento, denominados α e β, os quais tem por objetivo transformar linearmente as distribuições sobrepostas na distribuição referência, ou seja, deslocar e/ou escalar as distribuições de modo que maximize o acerto nas predições das amostras do domínio-alvo, como apresentado na Equação 2, em que $\oplus$ denota a operação de concatenação das saídas de camadas distintas.

$$BN(x) = \frac{x - E[x]}{\sqrt{Var[x] + \epsilon}} \tag{1}$$

$$DIAL(x) = \{BN_S(x_S) \oplus BN_T(x_T)\} \cdot \alpha + \beta \tag{2}$$

A grosso modo, este trabalho estende as camadas DIAL para realizar a adaptação de domínio de várias fontes para único alvo. Formalmente, as camadas MS-DIAL generalizam as transformações realizadas na Equação 2, aplicando-as a todos os domínios-fonte, como mostrado na Equação 3. Similar do DIAL, as amostras $x \subseteq \{X_1 \cup X_2 \cup \cdots \cup X_M \cup X_T\}$ dos mini-lotes de entrada são inicialmente agrupadas em $M + 1$ sub-lotes $x_1 \subseteq X_1, x_2 \subseteq X_2, \ldots, x_M \subseteq X_M, x_T \subseteq X_T$ de acordo com o domínio a qual pertencem e, em seguida, encaminhadas às camadas de normalização de lote $BN_1, BN_2, \ldots, BN_M, BN_T$, respectivamente, ajustando a distribuição de todos os domínios para uma mesma distribuição de referência e, por fim, a sobreposição das distribuições fonte e alvo é controlada por parâmetros α e β aprendíveis em tempo de treinamento.

A Figura 3 ilustra o fluxo de dados inerente às camadas MS-DIAL, desde o recebimento do mini-lote, a sua separação em vários domínios, a transformação para a distribuição referência e, por fim, a transformação afim para controle do alinhamento dos domínios. O caminho indicado em vermelho representa o fluxo de dados durante a fase de teste.

$$MS\text{-}DIAL(x) = \left\{ \left\{ \bigoplus_{i=1}^{M} BN_i(x_i) \right\} \oplus BN_T(x_T) \right\} \cdot \alpha + \beta \tag{3}$$

B. Treinamento e Inferência

Durante a fase de treinamento, os mini-lotes devem conter amostras $x = x_1 \oplus x_2 \oplus \cdots \oplus x_M \oplus x_T$ provenientes de todos os domínios-fonte, ou seja, $x_1 \subseteq \mathcal{S}_1, x_2 \subseteq \mathcal{S}_2, \ldots, x_M \subseteq \mathcal{S}_M$, as quais são acompanhadas de seus respectivos rótulos; e também do domínio-alvo, isto é, $x_T \subseteq \mathcal{T}$, para as quais os rótulos não são conhecidos. Ao final da passagem de um mini-lote pela rede, são obtidas predições $\{f_i^\theta(\mathbf{y}_i^k; \mathbf{x}_i^k)\}_{k=1}^{|x_i|}$ para amostras $x_1, x_2, \ldots, x_M$ dos domínios-fonte e também predições $\{f_T^\theta(\mathbf{y}; \mathbf{x}_T^k)\}_{k=1}^{|x_T|}$ para amostras x_T do domínio-alvo.

Similar ao DIAL, o valor do erro entregue ao otimizador é obtido a partir de uma função de perda $\mathcal{L}(\theta)$ composta por duas componentes, uma supervisionada $\mathcal{L}_\mathcal{S}(\theta)$ calculada a partir das predições obtidas para amostras dos domínios-fonte $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_M$ e outra não-supervisionada $\mathcal{L}_\mathcal{T}(\theta)$ calculada a partir das predições das amostras do domínio-alvo $\mathcal{T}$.

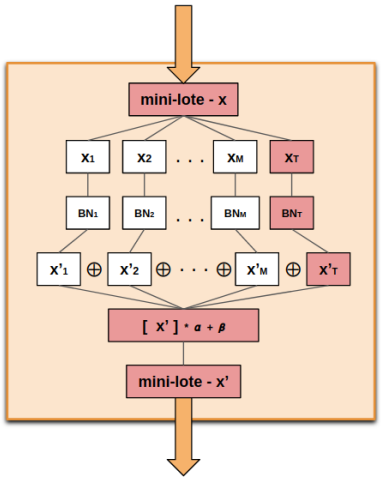

Figura 3. Fluxo de dados dos mini-lotes x nas camadas MS-DIAL durante a fase de treinamento, em que $BN_1, BN_2, \ldots, BN_M$ são as camadas de normalização de lote aplicadas às amostras $x_1, x_2, \ldots, x_M$ dos domínios-fonte $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_M$, respectivamente; BN_T é a camada de normalização de lote aplicada às amostras x_T do domínio-alvo $\mathcal{T}$; e α e β são parâmetros aprendíveis pela rede. O caminho destacado em vermelho refere-se ao fluxo de dados durante a fase de inferência.

A componente supervisionada $\mathcal{L}_\mathcal{S}$ é a entropia cruzada das amostras dos domínios-fonte, que é calculada pela Equação 4.

$$\mathcal{L}_\mathcal{S}(\theta) = -\sum_{i=1}^{M} \frac{1}{|x_i|} \sum_{k=1}^{|x_i|} \log f_i^\theta(\mathbf{y}_i^k; \mathbf{x}_i^k) \tag{4}$$

Já a componente não-supervisionada $\mathcal{L}_\mathcal{T}$ refere-se a entropia das amostras do domínio-alvo e é usada para forçar o modelo a decidir com mais confiança, sendo dada pela Equação 5.

$$\mathcal{L}_\mathcal{T}(\theta) = -\frac{1}{|x_T|} \sum_{k=1}^{|x_T|} \sum_{\mathbf{y} \in \mathcal{Y}} f_T^\theta(\mathbf{y}; \mathbf{x}_T^k) \log f_T^\theta(\mathbf{y}; \mathbf{x}_T^k) \tag{5}$$

A função de perda $\mathcal{L}(\theta)$ é a soma ponderada de $\mathcal{L}_\mathcal{S}(\theta)$ e $\mathcal{L}_\mathcal{T}(\theta)$, ou seja, $\mathcal{L}(\theta) = \mathcal{L}_\mathcal{S}(\theta) + \lambda \mathcal{L}_\mathcal{T}(\theta)$, em que λ é um hiperparâmetro associado ao peso da contribuição de $\mathcal{L}_\mathcal{T}(\theta)$. Nos experimentos, o hiperparâmetro λ foi fixado em 0,1.

Uma vez ajustados os parâmetros θ, os fluxos de dados associados aos domínios-fonte $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_M$ não são mais necessários. Dessa forma, as camadas MS-DIAL passam a operar como camadas padrões de normalização de lote quando usadas para realizar inferências, encaminhando os mini-lotes através dos caminhos associados somente ao domínio-alvo $\mathcal{T}$.

IV. Experimentos

Esta seção apresenta detalhes sobre o protocolo experimental adotado para avaliar o método proposto e também relata os resultados obtidos. A avaliação experimental foi conduzida em conjuntos de dados de pequeno e grande porte e o método proposto foi comparado com abordagens do estado da arte.

A. Conjuntos de Dados

O método proposto foi avaliado em duas tarefas distintas que envolvem cinco conjuntos de dados públicos amplamente usados para avaliar métodos de UDA: (*i*) no reconhecimento de dígitos dos conjuntos de dados MNIST [7], MNIST-M [8], SVHN [9] e *Synthetic Digits* [10]; e (*ii*) no reconhecimento de objetos usando o conjunto de dados Office-Home [2]. A seguir, são fornecidos detalhes de cada um desses conjuntos.

O conjunto de dados MNIST [7] é composto por imagens monocromáticas com resolução de 28x28 *pixels*, sendo 60000 imagens para treinamento e 10000 para teste. Essas imagens referem-se a dígitos manuscritos dos algarismos de 0 a 9, cada um correspondente a uma classe distinta.

O conjunto de dados MNIST-M [8] é composto por imagens coloridas de tamanho 32x32 *pixels*, sendo 59001 imagens para treinamento e 9001 para teste. Elas resultam da combinação das imagens do MNIST com padrões aleatórios extraídos de fotos coloridas do conjunto de dados BSDS500 [20], na qual os *pixels* que compõem os dígitos tem suas cores invertidas. MNIST-M, assim como MNIST, possui 10 classes que correspondem aos algarismos de 0 a 9. Embora para humanos a tarefa se torne um pouco mais difícil, a inserção de padrões aleatórios ao fundo e a cor não uniforme dos dígitos categorizam uma grande mudança de domínio.

O conjunto de dados SVHN (do inglês, *Street View House Number*) [9] é composto por imagens coloridas com resolução de 32x32 *pixels*, sendo 73257 imagens para treinamento e 26032 para teste. Tais imagens contém fotos de dígitos tiradas da numeração de casas e foram agrupadas em 10 classes correspondentes a dígitos no intervalo de 0 a 9. Apesar das semelhanças com MNIST e MNIST-M, o desbalanceamento no número de imagens por classe, as alterações severas de iluminação e a descentralização dos dígitos nas imagens representam mudanças significativas de domínio.

O conjunto de dados *Synthetic Digits* (Synth) [10] é composto por imagens coloridas com resolução de 32x32 *pixels*, sendo 479400 imagens para treinamento e 9553 para teste. Ele é composto por imagens sintéticas obtidas a partir de transformações de posição, orientação, borramento e coloração de dígitos de fontes do Windows™, cujos parâmetros foram manualmente ajustados para mimetizar amostras do conjunto SVHN. Tais imagens, assim como SVHN, possuem 10 classes que correspondem a dígitos no intervalo de 0 a 9 e, apesar de imitar o SVHN, seu desbalanceamento é muito maior, o que dificulta a transferência de conhecimento.

Office-Home [2] é um conjunto de dados de grande porte usado como referência para avaliar métodos de UDA. Ele é composto por 15500 imagens coletadas de vários sites e repositórios de imagens da Internet, apresentando resoluções que variam de 18x18 até 6500x4900 *pixels*. Essas imagens estão distribuídas em 65 classes de objetos e divididas em 4 domínios distintos: arte (2427), clipart (4365), produto (4439) e mundo real (4357). Os números entre parênteses indicam a quantidade de imagens em cada domínio.

As tarefas de reconhecimento de dígitos e de objetos nesses conjuntos de dados são bastante desafiadoras devido à grande

Tabela I
ACURÁCIA DE CLASSIFICAÇÃO (%) NOS CONJUNTOS DE DADOS DE DÍGITOS.

Métodos	Domínios				Média
	MNIST	MNIST-M	SVHN	Synth	
SRC	96,78 ± 0,08	60,80 ± 0,21	68,99 ± 0,69	84,09 ± 0,27	77,66 ± 0,14
DANN	96,41 ± 0,13	60,10 ± 0,27	70,19 ± 1,30	83,83 ± 0,25	77,63 ± 0,35
M3SDA	96,95 ± 0,06	65,03 ± 0,80	71,66 ± 1,16	80,12 ± 0,56	78,44 ± 0,36
MDAN	97,10 ± 0,10	64,09 ± 0,31	77,72 ± 0,60	85,52 ± 0,19	81,11 ± 0,21
MDMN	97,15 ± 0,09	64,34 ± 0,27	76,43 ± 0,48	85,80 ± 0,21	80,93 ± 0,16
DARN	**98,09 ± 0,03**	**67,06 ± 0,14**	81,58 ± 0,14	86,79 ± 0,09	83,38 ± 0,06
MS-DIAL	94.33 ± 0,06	61.24 ± 1,27	**85.61 ± 0,47**	**92.86 ± 0,20**	**83.51 ± 1,53**
TAR	99,02 ± 0,02	94,66 ± 0,10	87,40 ± 0,17	96,90 ± 0,09	94,49 ± 0,07

diferença entre os domínios, como ilustrado na Figura 4.

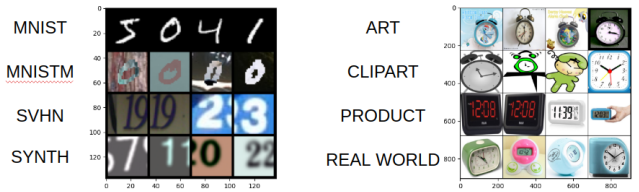

Figura 4. Exemplos de imagens do MNIST, MNIST-M, SVHN e Synth em (a) e dos domínios arte, clipart, produto e mundo real do Office-Home (b).

B. Protocolo Experimental

Os resultados do método proposto foram comparados com os relatados recentemente por Wen *et al.* [6] para cinco abordagens de referência: **DANN** [10], **M3SDA** [15], **MDAN** [17], **MDMN** [16], **DARN** [6]. Além disso, foram também considerados os resultados relatados por Wen *et al.* [6] para duas linhas de base: (*i*) **SRC**, que refere-se ao treinamento do modelo em um único conjunto composto por dados rotulados de todos domínios-fonte e, portanto, sem adaptação de domínio; e (*ii*) **TAR**, que refere-se ao treinamento do modelo em dados do domínio-alvo, porém, valendo-se do conhecimento prévio de seus rótulos verdadeiros, constituindo assim um limite superior para o desempenho de qualquer método de UDA.

Para se ter uma comparação justa, foi adotado o mesmo protocolo experimental usado por Wen *et al.* [6]. Em cada experimento, um domínio foi tomado como alvo e os demais foram usados como fonte. Esse processo foi repetido várias vezes, cada vez tomando um domínio diferente como alvo. Foi adotado a acurácia como métrica de classificação, descrita na Equação 6, que é calculada através da razão da quantidade de predições corretas, $VP+VN$, sendo VP e VN referentes a verdadeiros positivos e verdadeiros negativos, pela quantidade total de predições, $VP+VN+FP+FN$, onde FP e FN referem-se a falso positivo e falso negativo.

$$Acc = \frac{VP+VN}{VP+VN+FP+FN} \tag{6}$$

Para o reconhecimento de dígitos, os conjuntos de dados MNIST, MNIST-M, SVHN e Synth foram tratados como 4 domínios distintos. Para cada domínio, foram sub-amostrados aleatoriamente 20000 imagens para treinamento e 9000 para teste. Foram realizadas 20 repetições de cada experimento, sendo reportados a média e o erro padrão da acurácia do melhor modelo obtido em cada rodada. A topologia de rede adotada nesses experimentos foi a mesma usada por Peng *et al.* [15], substituindo-se camadas de normalização de lote padrão por camadas MS-DIAL. A rede foi treinada do zero por 120 épocas usando o algoritmo de otimização Adam [21] com tamanho de mini-lote de 64 (*i.e.*, 16 por domínio), decaimento de peso de 0,0005 e taxa de aprendizado inicial de 0,001 com decaimento programado por um fator de 10 nas épocas 50 e 90. Esse mesmo conjunto de parâmetros foi usado no trabalho de Roy *et al.* [22].

Para o reconhecimento de objetos, foram sub-amostrados aleatoriamente 2000 imagens de cada domínio para treinamento e as imagens restantes foram usadas para teste. Os resultados referem-se a média e o erro padrão das acurácias obtidas ao final de cada uma das 20 repetições que foram realizadas de cada experimento. Seguindo o trabalho de Roy *et al.* [22], foi adotada a rede ResNet-50 [23], substituindo-se camadas de normalização de lote padrão por camadas MS-DIAL. Primeiro, a rede foi inicializada com pesos pré-treinados na ImageNet [1] e a camada de saída foi substituída por uma camada totalmente conectada com 65 neurônios de saída e pesos inicializados aleatoriamente. Em seguida, a rede foi treinada por 60 épocas usando o algoritmo de otimização SGD (do inglês, *Stochastic Gradient Descent*) com tamanho de mini-lote de 80 (*i.e.*, 20 por domínio), fator de momentum de 0,9, decaimento de peso de 0,0005, taxa de aprendizado inicial de 0,01 para os parâmetros da camada de saída e de 0,001 para os demais parâmetros da rede. Decaimento programado foi usado para reduzir as taxas de aprendizado iniciais por um fator de 10 na época 54.

Os experimentos foram realizados em um servidor equipado com dois processadores Intel Xeon E5-2683v4 (16 núcleos de 2,1 GHz), 128 GBytes de memória DDR4 e 2 GPUs NVIDIA Tesla K80. O servidor executa o sistema operacional Linux CentOS 7.4 (*kernel* 3.10.0) e o sistema de arquivos ext4. Todos os códigos-fonte foram implementados em Python (versão 3.6.7) usando a biblioteca PyTorch (versão 1.2.0).

C. Resultados

Nas Tabelas I e II, são comparados os resultados obtidos pelo MS-DIAL e os relatados por Wen *et al.* [6] para tarefas de

Tabela II
ACURÁCIA DE CLASSIFICAÇÃO (%) NO CONJUNTO DE DADOS OFFICE-HOME.

Métodos	Domínios				Média
	Arte	Clipart	Produto	Mundo Real	
SRC	58,02 ± 0,47	57,29 ± 0,30	74,26 ± 0,22	77,98 ± 0,25	66,89 ± 0,16
DANN	57,39 ± 0,69	57,35 ± 0,35	73,78 ± 0,27	78,12 ± 0,21	66,66 ± 0,19
M3SDA	64,05 ± 0,61	62,79 ± 0,37	76,21 ± 0,30	78,63 ± 0,22	70,42 ± 0,18
MDAN	68,14 ± 0,58	67,04 ± 0,21	81,03 ± 0,22	82,79 ± 0,15	74,75 ± 0,18
MDMN	68,67 ± 0,55	67,75 ± 0,20	81,37 ± 0,18	83,32 ± 0,14	75,28 ± 0,15
DARN	70,00 ± 0,38	68,42 ± 0,14	**82,75 ± 0,21**	**83,88 ± 0,16**	76,26 ± 0,13
MS-DIAL	**82.85 ± 0,10**	**76.71 ± 0,10**	80.74 ± 0,09	82.70 ± 0,09	**80.75 ± 0,28**
TAR	71,19 ± 0,38	79,16 ± 0,16	90,66 ± 0,15	85,60 ± 0,14	81,65 ± 0,12

MSDA com os conjuntos de dados de dígitos e Office-Home, respectivamente. O melhor resultado obtido para as amostras de teste de cada domínio-alvo está destacado em negrito. Exceto pelo DARN, o MS-DIAL supera todas as demais abordagens comparadas em todas as tarefas de MSDA. Apesar do DARN alcançar uma acurácia média de classificação ligeiramente superior a do MS-DIAL para alguns domínios-alvos, o MS-DIAL tem um desempenho médio melhor que o DARN tanto para o reconhecimento de dígitos quanto de objetos.

V. CONCLUSÃO

Neste trabalho, foi apresentado o MS-DIAL, uma nova abordagem para o problema de MSDA. Nesse método, camadas de alinhamento de domínio são inseridas em diversos níveis da rede e, assim, o alinhamento entre as distribuições dos domínios-fonte e do domínio-alvo é realizada de forma automática a partir de parâmetros aprendíveis em tempo de treinamento. Dessa forma, os parâmetros das demais camadas da rede podem ser compartilhados entre todos os domínios, otimizando o tempo de processamento e uso de memória.

O MS-DIAL foi avaliado em tarefas de reconhecimento de dígitos e de objetos com cinco conjuntos de dados públicos amplamente usados para avaliar métodos de UDA. Os resultados obtidos com uso das camadas MS-DIAL foram promissores e superaram, em média, abordagens do estado da arte.

Em trabalhos futuros, pretende-se avaliar o MS-DIAL em outros conjuntos de dados. Além disso, pretende-se também investigar o uso do MS-DIAL em outras tarefas desafiadoras, como adaptação de domínio de conjunto aberto (do inglês, *open set domain adaptation* – OSDA) e generalização de domínio (do inglês, *domain generalization* – DG).

AGRADECIMENTOS

Este trabalho foi apoiado pela Fundação de Amparo à Pesquisa do Estado de São Paulo - FAPESP (processos 2017/25908-6 e 2019/10998-5) e pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq (processos 423228/2016-1, 313122/2017-2 e 167857/2019-3).

REFERÊNCIAS

[1] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and F.-F. Li, "Imagenet large scale visual recognition challenge," *IJCV*, vol. 115, no. 3, pp. 211–252, 2015.

[2] H. Venkateswara, J. Eusebio, S. Chakraborty, and S. Panchanathan, "Deep hashing network for unsupervised domain adaptation," in *CVPR*, 2017, pp. 5385–5394.

[3] F. M. Carlucci, L. Porzi, B. Caputo, E. Ricci, and S. R. Bulò, "Just DIAL: domain alignment layers for unsupervised domain adaptation," in *Int. Conf. Image Analysis and Processing*, 2017, pp. 357–369.

[4] R. Xu, Z. Chen, W. Zuo, J. Yan, and L. Lin, "Deep cocktail network: Multi-source unsupervised domain adaptation with category shift," in *CVPR*, 2018, pp. 3964–3973.

[5] S. Zhao, B. Li, X. Yue, P. Xu, and K. Keutzer, "MADAN: multi-source adversarial domain aggregation network for domain adaptation," *CoRR*, vol. abs/2003.00820, 2020.

[6] J. Wen, R. Greiner, and D. Schuurmans, "Domain aggregation networks for multi-source domain adaptation," in *ICML*, 2020, pp. 10 927–10 937.

[7] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[8] Y. Ganin and V. S. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *ICML*, 2015, pp. 1180–1189.

[9] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, "Reading digits in natural images with unsupervised feature learning," in *NIPS Work. Deep Learning and Unsupervised Feature Learning*, 2011.

[10] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. S. Lempitsky, "Domain-adversarial training of neural networks," *JMLR*, vol. 17, pp. 59:1–59:35, 2016.

[11] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Trans. Neural Networks*, vol. 22, no. 2, pp. 199–210, 2011.

[12] M. Baktashmotlagh, M. T. Harandi, and M. Salzmann, "Distribution-matching embedding for visual domain adaptation," *JMLR*, vol. 17, pp. 108:1–108:30, 2016.

[13] H. Yan, Y. Ding, P. Li, Q. Wang, Y. Xu, and W. Zuo, "Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation," in *CVPR*, 2017, pp. 945–954.

[14] F. M. Carlucci, L. Porzi, B. Caputo, E. Ricci, and S. R. Bulò, "Autodial: Automatic domain alignment layers," in *ICCV*, 2017, pp. 5077–5085.

[15] X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang, "Moment matching for multi-source domain adaptation," in *ICCV*, 2019, pp. 1406–1415.

[16] Y. Li, M. Murias, G. Dawson, and D. E. Carlson, "Extracting relationships by multi-domain matching," in *NeurIPS*, 2018, pp. 6799–6810.

[17] H. Zhao, S. Zhang, G. Wu, J. M. F. Moura, J. P. Costeira, and G. J. Gordon, "Adversarial multiple source domain adaptation," in *NeurIPS*, 2018, pp. 8568–8579.

[18] S. Zhao, G. Wang, S. Zhang, Y. Gu, Y. Li, Z. Song, P. Xu, R. Hu, H. Chai, and K. Keutzer, "Multi-source distilling domain adaptation," in *AAAI Conf. Artificial Intelligence*, 2020, pp. 12 975–12 983.

[19] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *ICML*, F. R. Bach and D. M. Blei, Eds., 2015, pp. 448–456.

[20] P. Arbelaez, M. Maire, C. C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *T-PAMI*, vol. 33, no. 5, pp. 898–916, 2011.

[21] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.

[22] S. Roy, A. Siarohin, E. Sangineto, S. R. Bulò, N. Sebe, and E. Ricci, "Unsupervised domain adaptation using feature-whitening and consensus loss," in *CVPR*, 2019, pp. 9471–9480.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.