

AutoDIAL: 自动域对齐层

法比奥·玛丽亚·卡鲁奇
Sapienza, 罗马, 意大利

洛伦佐·波尔齐
IRI CSIC-UPC, 西班牙巴塞罗那
Mapillary, 格拉茨, 奥地利

芭芭拉·卡普托
Sapienza, 罗马, 意大利

伊丽莎·里奇
FBK, 特伦托, 意大利
意大利佩鲁贾大学

塞缪尔·罗塔·布洛
FBK, 特伦托, 意大利
Mapillary, 格拉茨, 奥地利

抽象的

在对不同设置中获取的数据进行测试时, 在给定数据库上训练的分类器表现不佳。这在域适应中通过源域和目标域的分布之间的转移来解释。传统上, 通过在目标函数中引入适当的损失项、测量源和目标分布之间的差异来减少域偏移的尝试使它们对齐。在这里, 我们采取了不同的路线, 建议通过嵌入任何给定的特定网络来对齐学习的表示域对齐层, 旨在将源和目标特征分布与参考特征分布相匹配。与之前定义的作品相反先验 应该在哪些层进行自适应, 我们的方法能够自动地 学习深度网络不同层次所需的特征对齐程度。在无监督环境中对不同公共基准进行的彻底实验证实了我们方法的力量。

全部。已经提出了几种方法, 都考虑了手工制作的特征 [16, 12, 13, 22, 8] 和深度模型 [23, 35, 9, 25, 11, 21]。尤其是最近基于深度学习的工作取得了显着的表现。这些方法中的大多数试图通过学习对域转移不变的特征来减少源和目标分布之间的差异。传统上采用两种主要策略。一种是基于最大平均差异 (MMD) [23, 25] 的最小化: 通过最小化它们的平均嵌入之间的距离, 将学习到的源和目标表示的分布优化为尽可能相似。另一种策略 [35, 9] 依赖于域混淆损失, 引入以学习辅助分类器来预测样本是来自源域还是来自目标域。IE 通过强加辅助分类器表现出较差的性能, 可以获得域不变的特征。

最近, 研究人员也开始研究替代方向 [11, 3, 21, 4], 例如使用编码器 - 解码器网络联合学习源标签和重建无监督目标图像, 或减少域偏移的可能性设计特定的分布归一化层。特别是, 后一种想法在 [21] 中得到了利用, 其中提出了一种用于深度域适应的简单无参数方法, 称为自适应批量归一化 (AdaBN)。受流行的批量归一化 (BN) 技术 [17] 的启发, AdaBN 修改了 Inception-BN 网络, 并在测试时执行 BN 时通过对源域和目标域使用不同的均值/方差项来对齐学习的源和目标表示。这导致学习域不变特征而不需要额外的损失项 (例如 MMD, domainconfusion) 在优化函数和相关的额外参数。

一、简介

尽管深度学习在视觉识别方面取得了进步, 但跨不同视觉领域进行泛化的能力仍然遥不可及。训练 (源) 和测试 (目标) 数据独立且相同地从同一分布中抽取的假设在许多实际应用中并不成立。事实上, 已经表明, 即使使用强大的深度学习模型, 域转移问题也可以得到缓解, 但无法消除 [7]。

在过去的几年里, 研究界在解决领域转移方面付出了巨大的努力。在这种情况下, 无监督域适应的具体问题, IE 目标域中没有可用的标记数据, 值得特别注意。事实上, 在许多应用程序中, 注释数据是一项繁琐的操作, 或者可能无法在

受[21]的启发, 本文介绍了新颖的 域对齐层 (DA 层) (图 1), 它们嵌入在深层架构的不同级别以对齐

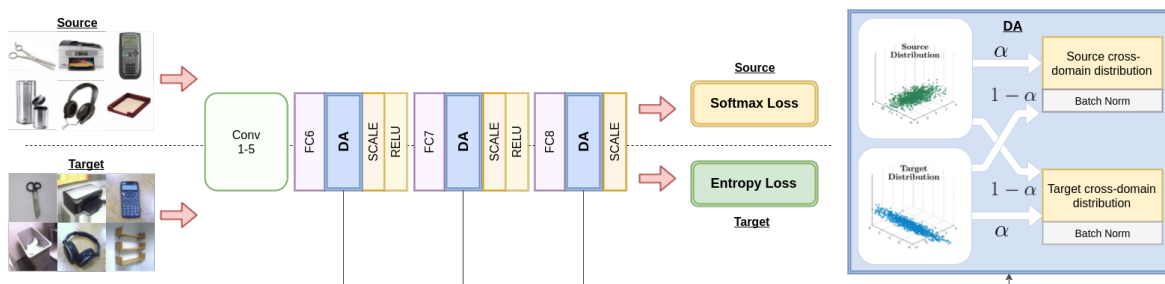


图 1. 应用于 AlexNet [19] 的 AutoDIAL。源图像和目标图像被馈送到网络。通过第一层后，它们进入我们的 DA 层，在那里源和目标分布对齐。DA 层学习新定义的统计信息 *源和目标跨域分布* 并根据计算的均值和方差对源和目标小批量进行归一化，这两个域不同（参见第 3.1 节）。每个分布受另一个影响的量，因此域对齐的程度，取决于一个参数， $\alpha \in [0.5, 1.0]$ ，这也是自动学习的。在流过整个网络后，源样本贡献了 Softmax 损失，而目标样本贡献了熵损失，这促进了最大程度分离未标记数据的分类模型。请注意，我们使用多个 DA 层来对齐不同级别的学习特征表示。

将学习到的源和目标特征分布转换为规范分布。不同于 [21] 和所有先前决定的深度域适应方法 *先验* 应该调整哪些层，我们赋予我们的 DA 层能力 *自动地* 了解应该在网络的不同级别上追求的对齐程度。据我们所知，这是试图追求这一目标的第一部作品。此外，我们认为在 [21] 中未标记的目标数据没有得到充分利用（参见第 3.1 节）。相反，我们利用来自目标域的信息来构建网络参数的先验分布，将学习到的解决方案偏向于能够很好地分离目标域中的类的模型（参见第 3.2 节和 [14]）。我们的 DA 层和考虑的先验分布在学习过程中协同工作：第一个对齐源和目标特征分布，第二个鼓励网络学习导致最大分离目标类的特征。我们将我们的算法称为 AutoDIAL——具有自动对齐参数的域对齐层。

贡献。这项工作的贡献是三方面的。首先，我们提出了一种无监督域适应的方法，基于引入 DA 层来明确解决域转移问题，该方法与在学习过程中利用无监督目标数据的熵损失协同作用。我们的解决方案同时对齐特征表示并学习应该在何处以及在何种程度上进行适应。其次，与之前优化域差异正则化项 [25, 35, 9, 23] 的工作相比，我们的 DA 层不需要任何额外的元参数。第三，我们对三个不同的基准进行了广泛的实验分析。我们发现我们的无监督域适应方法

优于最先进的方法，可以应用于不同的 CNN 架构，不断提高它们在域适应问题中的性能。

2. 相关工作

无监督域自适应侧重于标记数据仅在源域中可用的场景。传统方法通过考虑两种主要策略来解决减少源分布和目标分布之间差异的问题。第一个是基于实例重新加权 [16, 5, 37, 12, 38]。最初，源样本根据它们与目标数据的相似性被赋予不同的重要性。然后，重新加权的实例用于学习目标域的分类/回归模型。按照这个方案，黄等。[16] 介绍了核均值匹配，这是一种设置源样本权重的非参数方法，无需明确估计数据分布。铈等。[12] 提议自动发现地标数据点，*IE* 源实例的子集与目标数据更相似，并使用它们来创建域不变特征。楚等。[5] 在单个优化问题中形式化了样本选择和分类器学习这两个任务。虽然这些作品考虑了手工制作的特征，但最近类似的想法已应用于深度模型。例如，曾等。[[38] 描述了一种用于行人检测的无监督域自适应方法，使用深度自动编码器来加权源训练样本的重要性。

无监督域适应的第二种策略是基于特征对齐，*IE* 源数据和目标数据被投影到一个公共子空间中，以减少相关分布之间的距离。这种方法在过去几年引起了相当大的兴趣，并且已经提出了几种不同的方法，都考虑了浅层模型 [13、22、8] 和深层架构 [23、35、9、11、3]。关注最近的深度域适应方法，通常考虑两种不同的方案来对齐特征

真实表示：(i) 引入多种适应方案以减少最大平均差异 [23, 25, 32] 或 (ii) 在域对抗环境中学习深度特征，以最小化域混淆损失 [35, 9]。我们的方法属于采用深度学习进行领域适应的方法类别。然而，我们明显偏离了以前的工作，通过引入基于 DAlayers 的域对齐方法来减少源和目标分布之间的差异。与我们最接近的工作是 [21]，其中 Li 等。建议在域适应的上下文中使用 BN。我们的方法可以看作是 [21] 的概括，因为我们的 DA 层允许在深度网络的每个级别自动调整所需的适应程度。此外，我们还引入了网络参数的先验，以便在训练期间充分受益于目标样本。第 4 节中介绍的实验显示了我们想法的显著附加值。

3. 自动领域对齐层

让 X 是输入空间（例如图像）和 Y 是输出空间（例如图像类别）我们的学习任务。在无监督域适应中，我们有来源域和一个目标通过概率确定的域

能力分布 p_{θ}^{xy} 和 $p_{\theta}^{y|x}$ ，分别定义在 $X \times Y$ 。这两个分布通常不同且未知，但我们提供了一个源数据集 $\mathcal{S} = \{(X_{\mathcal{S}1}, y_{\mathcal{S}1}), \dots, (X_{\mathcal{S}n}, y_{\mathcal{S}n})\}$ 的身份证来自观察 \mathcal{S} 和一个未标记的目标数据集 $\mathcal{T} = \{X_{\mathcal{T}1}, \dots, X_{\mathcal{T}m}\}$ 的身份证来自边缘的观察 \mathcal{T} 。目标是估计配对预测器来自 \mathcal{S} 和 \mathcal{T} 可用于对来自目标域的样本点进行配对。这项任务特别具有挑战性，因为一方面我们缺乏对来自目标域的标签的直接观察，另一方面源域和目标域分布之间的差异阻止了在其上训练的预测器易于应用于目标域。

许多最先进的方法试图通过在特征或分类器级别执行某种形式的对齐来减少域差异。特别是，最近最成功的方法试图夫妻训练过程和域适应步骤深的神经架构 [9, 25, 23]，因为该解决方案可以在不同抽象级别进行对齐。我们提出的方法包含相同的哲学，同时背离了域对齐可以通过应用相同的源域和目标域的预测器。这是由不可能性定理 [2] 推动的，该定理直观地表明没有学习者依赖于协变量移位

假设， IE p_{θ}^{xy} 是 $X \times Y$ 上 $p_{\theta}^{y|x}$ ，并实现低差异源和目标未标记分布之间的关系 p_{θ}^{xy} 和 $p_{\theta}^{y|x}$ ，保证域适配成功训练之间没有进一步的相关性假设

和目标分布。出于这个原因，我们假设源和目标预测器通常是不同的职能。尽管如此，两个预测变量都依赖于一个共同的参数 θ 属于一个集合 Θ ，它明确地耦合了两个预测变量，而不直接参与源域和目标域的对齐。这与大多数最先进的方法形成对比，这些方法增加了用于训练预测器的损失函数，正则化项惩罚源和目标表示之间的差异（参见，例如 [9, 25, 23]）。我们采取的观点不同，在精神上与 AdaBN [21] 接近。它包括通过引入所谓的域对齐层（DA 层）。此外，我们通过赋予架构能力来回避决定哪些层应该对齐以及对齐到什么程度的问题。自动地调整每个域对齐层中应考虑的对齐程度。本节的其余部分致力于提供我们方法的详细信息。

3.1. 源和目标预测器

源和目标预测器被实现为两个几乎相同的深度神经网络，因为它们共享相同的结构和相同的权重（由参数给出 θ ）。然而，这两个网络还包含许多特殊层，即 DA 层，它们实现了特定于域的操作。事实上，这些层的作用是应用数据转换，将观察到的输入分布与参考分布对齐。由于通常源预测器和目标预测器的输入分布不同，而参考分布保持不变，因此两个预测器在相应的 DA 层中经历了不同的变换。因此，源预测器和目标预测器实际上实现了不同的功能，这对于第 2 节中给出的原因很重要。3.

我们的 DA 层的实际实现受到 AdaBN [21] 的启发，其中批标准化层用于通过匹配一阶和二阶矩来独立地将源和目标分布与标准正态分布对齐。他们提出的方法包括在具有 BN 层的源网络上训练，从而获得源预测器，并导出目标预测器作为后处理步骤，仅使用目标样本重新估计 BN 统计数据。因此，源和目标预测器共享相同的网络参数但具有不同的 BN 统计数据，从而使两个预测器具有不同的功能。

我们提出的方法坚持使用 BN 层对齐域的相同想法，但我们引入了根本性的变化。AdaBN 的一个限制是目标样本对网络参数没有影响，因为它们在训练过程中没有被观察到。我们的方法通过耦合网络参数克服了这个限制

在训练时到目标和源样本。这是通过两种方式实现的：首先，我们基于目标样本引入网络参数的先验分布；其次，我们通过向每个特定领域 DA 层的输入分布引入参数化的跨域偏差，赋予架构学习适应程度的能力。本小节的其余部分专门用于描述新层，而我们将在下一小节中介绍先验分布。

DA 层。如前所述，我们的 DA 层源自 Batch Normalization，但与 BN 不同，BN 根据来自 mini-batch 的输入分布计算一阶和二阶矩，我们让后者的统计数据被来自另一个域，从而引入跨域偏差。由于源和目标预测器共享相同的网络拓扑，一个预测器中的每个 DA 层在另一个预测器中都有一个匹配的 DA 层。让 $X_{秒}$ 和 $X_{吨}$ 对于给定的特征通道和空间位置，分别表示源和目标预测器中匹配 DA 层的输入。认为 $q_{秒}$ 和 $q_{吨}$ 是分布 $X_{秒}$ 和 $X_{吨}$ ，分别

积极地，让 $q_{英石} = \alpha q_{秒} + (1 - \alpha) q_{吨}$ 并且对称地， $q_{吨} = \alpha q_{秒} + (1 - \alpha) q_{英石}$ 是由一个因子混合的跨域分布 $\alpha \in [0, 1]$ 。然后，源网络和目标网络中 DA 层的输出分别由下式给出

$$DA(X_{秒}; \alpha) = \sqrt{\frac{X_{秒} - \text{微米}_{st,\alpha}}{\varepsilon + \sigma_{st,\alpha}^2}}, \quad DA(X_{吨}; \alpha) = \sqrt{\frac{X_{吨} - \text{微米}_{ts,\alpha}}{\varepsilon + \sigma_{ts,\alpha}^2}} (1)$$

在哪里 $\varepsilon > 0$ 是一个很小的数字，以避免在零方差的情况下出现数值问题， $\text{微米}_{st,\alpha} = \mathbb{E}_{X \sim q_{英石}}[X]$ ， $\sigma_{st,\alpha}^2 = \mathbb{E}_{X \sim q_{英石}}[X^2] - (\text{微米}_{st,\alpha})^2$ ，同样 $\text{微米}_{ts,\alpha}$ 和 $\sigma_{ts,\alpha}^2$ 是平均的和方差 $X \sim q_{ts}$ 。类似于 BN，我们估计统计数据基于小批量并通过统计类似地导出梯度（参见补充材料）。

引入混合因子的基本原理 α 是我们可以从具有类似于 AdaBN 的两个域的独立对齐转移，当 $\alpha = 1$ 、进行耦合归一化，当 $\alpha = 0.5$ 。在前一种情况下，DA 层在源预测器和目标预测器中计算两个不同的函数，等效于考虑完整的域对齐程度。后一种情况，

相反，产生相同的功能，因为 $q_{英石} = 0.5 q_{秒} + 0.5 q_{吨}$ 因此平等地转换两个域，这不会产生域对齐。由于混合参数 α 不是先验固定的，而是在训练阶段学习的，我们得到的结果是，网络可以决定在应用 DA 层的架构的每个级别上域对齐应该有多强。有关用于实现两个域预测器的实际 CNN 架构的更多详细信息，请参见第 4.1 节。

3.2. 训练

在训练阶段，我们估计参数 θ ，它使用源数据集提供的观测值保存源和目标预测器共享的神经网络权重，包括与 DA 层相关的混合因子 α 和目标数据集 吨 。由于我们坚持判别模型，因此无法使用未标记的目标数据集来表达数据可能性。但是，我们可以利用 吨 构造参数的先验分布 θ 。因此，我们形成了一个后验分布 θ 鉴于观察 秒 和 吨 作为

$$\pi(\theta/S, T) \propto \pi(\text{是秒}/X_{秒}, T, \theta) \pi(\theta/T, X_{秒}), \quad (2)$$

在哪里 $\text{是秒} = \{\text{是秒}_1, \dots, \text{是秒}_n\}$ 和 $X_{秒} = \{X_{秒_1}, \dots, X_{秒_n}\}$ 搜集中观测值的标签和数据点 秒 ，分别。后验分布在 Θ 上最大化以获得最大后验估计 θ^* 源和目标预测器中使用的参数：

$$\theta^* \in \underset{\theta \in \Theta}{\text{最大参数}} \pi(\theta/S, T). \quad (3)$$

期限 $\pi(\text{是秒}/X_{秒}, T, \theta)$ 在 (2) 中表示可能性 θ 对于源数据集，而 $\pi(\theta/T, X_{秒})$ 是取决于目标数据集的先验项，它在经典学习理论意义上充当正则化器。由于我们在 DA 层中的跨域统计数据，这两个术语实际上都取决于两个域¹

$\alpha < 1$ 并从源的样本中估计 α 和目标域。

由于数据样本，似然在样本点上分解为以下乘积 ² 身份假设：

$$\pi(\text{是秒}/X_{秒}, T, \theta) = \prod_{i=1}^n F_{\theta}(\text{是秒}_i; X_{秒-i}), \quad (4)$$

在哪里 $F_{\theta}(\text{是秒}_i; X_{秒-i})$ 是样本点的概率 $X_{秒}$ 拿标签 是秒_i 根据源预测器（我们省略了依赖 吨 和 $X_{秒}$ 为符号方便）。

在深入研究前一项的细节之前，我们想评论一下概率模型中缺少一个显式组件来尝试对齐源分布和目标分布。这是因为在我们的模型中，域对齐步骤由每个预测器独立地通过域对齐层接管，如前一小节所示。

预先分配。参数的先验分布 θ 源和目标预测器共享是根据观察到的目标数据分布构建的。这种选择的动机是从理论上从未标记的数据点中提取更多信息，因为它们表现出低级别的类重叠 [27]。因此，基于预测变量的先验偏差是合理的

当相同的预测器应用于目标样本时观察到的标签不确定性程度。这种意义上的不确定性可以通过假设来衡量 θ 根据经验熵是 $h(\theta)$ 如下

$$H(\theta/T, X^{秒}) = - \frac{1}{n} \sum_{t=1}^n \sum_{y \in Y} F_{\theta}(是; X_{t-1}^{秒}) \log F_{\theta}(是; X_{t-1}^{秒}), \quad (5)$$

在哪里 $F_{\theta}(是; X_{t-1}^{秒})$ 表示样本点的概率 $X_{t-1}^{秒}$ 拿标签 是根据目标预测器 (我们再次省略了依赖 θ 和 $X^{秒}$)。

现在可以推导出先验分布 $\pi(\theta/T, X^{秒})$ 在标签不确定性度量方面 $H(\theta/T, X^{秒})$ 通过要求先验分布 π 离子在约束下最大化熵

$H(\theta/T, X^{秒}) \pi(\theta/T, X^{秒}) d\theta = \varepsilon$, 其中常数 $\varepsilon > 0$ 指定标签不确定性平均应该有多小。这产生了一个带解的凹变分优化问题:

$$\pi(\theta/T, X^{秒}) \propto \exp(-\lambda h(\theta/T, X^{秒})), \quad (6)$$

在哪里 λ 是对应于的拉格朗日乘数 ε . 由此产生的先验分布满足偏好模型的期望属性, 这些模型表现出很好的分离类 (IE 具有较低的值 $H(\theta/T, X^{秒})$), 从而能够在有区别的环境中利用未标记的目标观察的信息内容 [14].

这种先验分布在其他工作中也被采用 [25] 以利用来自目标分布的更多信息, 但之前从未与显式域对齐方法结合使用 (IE 不是基于额外的正则化术语, 例如 MMD 和域混淆), 就像我们提议的那样。

推理。一旦我们估计了最佳网络参数 θ^* 通过求解 (3), 我们可以消除目标预测器对 θ 和 $X^{秒}$. 其实修复后 θ^* 每个 DA 层的输入分布也变得固定, 因此我们可以一次计算和存储所需的统计数据, 类似于标准 BN。

3.3. 实施说明

DA 层可以作为标准批量归一化的最直接的修改来实现。我们建议读者参考补充材料以获得完整的推导。特别是在我们的实现中, 我们将每对 DA 层视为一个网络层, 它同时计算等式 (1) 中的两个归一化函数并学习 α 范围。在训练期间, 每个批次包含固定数量的源样本, 然后是固定数量的目标样本, 这使得我们的 DA 层可以简单地区分两者。与标准 BN 类似, 我们保持单独的移动平均线

源和目标统计信息。请注意, 如前所述, $\alpha \in [0.5, 1]$ 。我们通过在网络的每次前向传递中将其值裁剪在允许的范围内来强制执行此操作。

通过用负对数的等效最小化替换 (3) 中的优化问题 $\pi(\theta/S, T)$ 并结合 (2)、(4)、(5) 和 (6) 我们得到一个损失函数 $\mathcal{L}(\theta) = \mathcal{L}^{秒}(\theta) + \lambda \mathcal{L}^{吨}(\theta)$, 在哪里:

$$\begin{aligned} \mathcal{L}^{秒}(\theta) &= - \frac{1}{n} \sum_{t=1}^n \log F_{\theta}(是; X_{t-1}^{秒}), \\ \mathcal{L}^{吨}(\theta) &= - \frac{1}{n} \sum_{t=1}^n \sum_{y \in Y} F_{\theta}(是; X_{t-1}^{吨}) \log F_{\theta}(是; X_{t-1}^{吨}). \end{aligned}$$

期限 $\mathcal{L}^{秒}(\theta)$ 是应用于源样本的标准对数损失, 而 $\mathcal{L}^{吨}(\theta)$ 是应用于目标的熵损失

样品。第二项可以通过喂食来实现

$F_{\theta}(是; X_{t-1}^{吨})$ 到交叉熵损失层的两个输入, 其中由选择的深度学习工具包支持。在我们的实现中, 基于 Caffe [18], 我们通过稍微修改现有的最大损失层。¹

4. 实验

在本节中, 我们广泛评估了我们的方法, 并将其与最先进的无监督域适应方法进行了比较。我们还提供了对拟议框架的详细分析, 从经验上证明了我们贡献的影响。请注意, 以下所有结果均报告为五次训练/测试运行的平均值。

4.1. 实验装置

数据集。我们在三个公开可用的数据集上评估了所提出的方法。

这办公室 31[29] 数据集是测试领域适应方法的标准基准。它包含来自三个不同领域的 31 个类别的 4652 张图片: 亚马逊 (A)、DSRL (D) 和网络摄像头 (W)。亚马逊图片收集自亚马逊网站, 网络摄像头和 DSLR 图像是在办公室环境中手动收集的。在我们的实验中, 我们考虑了这些域的所有可能的源/目标组合, 并采用了完整协议设置 [12], IE 我们在整个标记的源和未标记的目标数据上进行训练, 并在带注释的目标样本上进行测试。

这办公室-加州理工 [13] 数据集是通过选择子集获得的 10 Office31 和 Caltech256[15] 数据集中的常见类别。它包含了 2533 其中大约一半的图像属于 Caltech256。Amazon (A)、DSLR (D)、Webcam (W) 和 Caltech256 (C) 中的每一个都被视为单独的域。在我们的实验中, 我们只考虑包含 C 的源/目标组合作为源域或目标域。

¹源代码可在 <https://github.com/ducksoup/autodial>

为了进一步对大规模数据集进行分析，我们还考虑了最近的跨数据集测试台在 [34] 中介绍，特别是加州理工学院-ImageNet 环境。该数据集是通过收集对应于 40 Caltech256 (C) 和 Imagenet (I) [6] 数据集之间共享的类。为了便于与之前的作品 [35, 33, 31] 进行比较，我们在两种不同的设置中进行了实验。[33, 35] 中采用的第一个设置考虑了通过在所有 40 个类别中选择来自 ImageNet 的 5534 张图像和来自 Caltech256 的 4366 张图像获得的 5 个分割。[31] 中采用的第二个设置为 Caltech256 使用了 3847 张图像，为 ImageNet 使用了 4000 张图像。

网络和培训。我们将所提出的方法应用于两个最先进的 CNN，即 AlexNet [19] 和 Inception-BN [17]。我们使用带有动量的小批量随机梯度下降训练我们的网络，如在 Caffe 库中实现的那样，使用以下元参数：权重衰减 5×10^{-4} ，势头 0.9、初始学习率 10^{-3} 。我们通过将所有图像缩放到 256×256 pxls，随机裁剪 227×227 pxls（用于 AlexNet）或 224×224 pxls (Inception-BN) 补丁和执行随机翻转。在所有实验中，我们选择参数 λ 通过根据 [25] 中的协议对源集进行交叉验证。

AlexNet [19] 是一个众所周知的架构，具有五个卷积层和三个全连接层，表示为 fc6, fc7 和 fc8。的输出 FC6 和 FC7 通常在领域适应文献中用作传统机器学习方法的预训练特征表示 [7, 31]。在我们的实验中，我们通过向每个全连接层附加一个 DA 层来修改 AlexNet。与最初的 AlexNet 不同，我们不对输出执行 dropout FC6 和 fc7。我们从在 ILSVRC-2012 数据集上训练的公开可用模型初始化网络参数，我们冻结所有卷积层，并提高学习率 fc8 通过一个因素 10。在训练期间，每个 mini-batch 包含一些与相应数据集大小成比例的源和目标样本，而批大小保持固定为 256。我们总共训练了 60 个 epoch（其中“epoch”是指完整的传递源集），将学习率降低一个因子 10^{-5} 个时代之后。

Inception-BN [17] 是通过连接“inception”块获得的非常深的架构。每个块由多个具有批量归一化和池化层的并行卷积组成。为了将所提出的方法应用于 Inception-BN，我们将每个批量归一化层替换为 DA 层。与 AlexNet 类似，我们从基于 ILSVRC-2012 数据集训练的公开可用模型初始化网络参数，并冻结前三个初始块。这 α 参数也固定为一个值 0.5 在

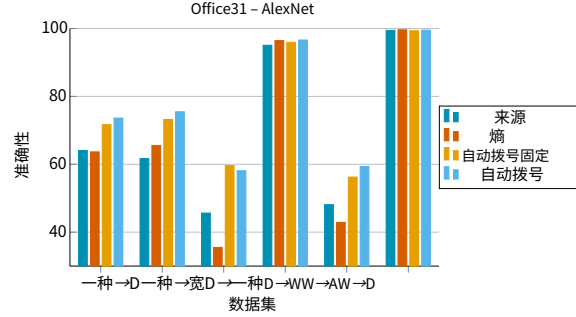


图 2. 考虑基于 AlexNet 的不同架构时 Office31 数据集的准确性。

前三个块的 DA 层，相当于保留了原始的批归一化层。由于 GPU 内存限制，我们使用比 AlexNet 小得多的批大小，并将每批中源和目标样本的数量分别固定为 32 和 16。在 Office-31 实验中，我们训练了 1200 次迭代，从而减少了 1000 次迭代后学习率降低 10 倍，而在跨数据集测试台实验中，我们训练了 2000 次迭代，在 1500 次之后降低了学习率。

4.2. 建议方法的分析

我们对所提出的方法进行了深入分析，评估了我们三个主要贡献的影响：i) 通过将源和目标分布与参考分布相匹配来对齐特征；ii) 学习适应系数 α ；iii) 应用基于熵的正则化项。作为第一组实验，我们对 Office31 数据集进行了消融研究，并在图 2 中报告了结果。在这里，我们比较了 AlexNet 网络的四种变体的性能：仅在源数据上训练（来源）；加上熵损失（Entropy）；与 DA layers 和 α 固定为 1（自动拨号固定）；使用 DA 层并学习 α （自动拨号）。学习的优势在这里 α 很明显，因为 AutoDIAL 在除一项实验设置之外的所有设置中都优于 AutoDIAL-fixed。有趣的是，熵项本身的添加似乎对最终精度有不同的影响：在 D→A 和 W→A 特别是，与 Source 相比，Entropy 的性能急剧下降。这并不奇怪，因为这两个设置对应于标记源样本数量非常有限且域偏移更严重的情况。然而，将 DA 层与熵损失结合使用总是会带来可观的性能提升。这些结果证实了我们贡献的有效性：当源和目标数据表示对齐时，熵正则化项特别有用。

在图 3 中，我们绘制了 α 由 AutoDIAL - AlexNet 和 AutoDIAL - Inception-BN 中的 DA layers 在 Office31 数据集上学习。在这两个网络中，我们观察到较低层倾向于学习更接近于 1、 $1/E$ 与层相比需要更高程度的适应

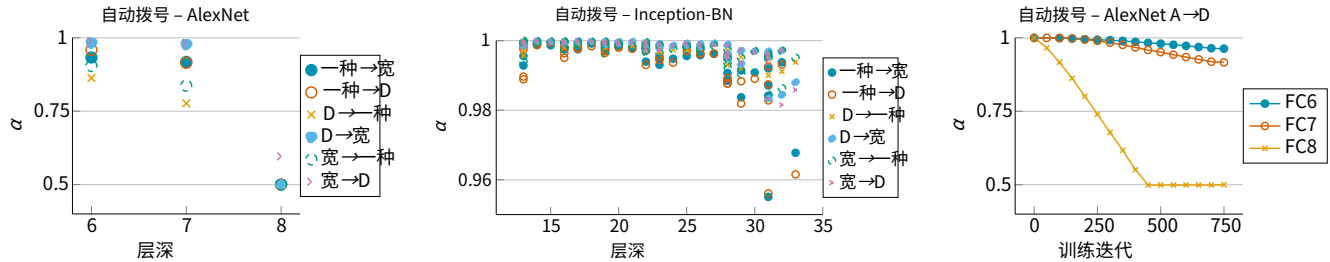


图 3. α 在 Office31 数据集上学习的参数，绘制为层深度（左和中）和训练迭代（右）的函数。

更接近分类器。然而，与 AutoDIAL - Inception-BN 相比，这种行为在 AutoDIAL - AlexNet 中似乎更为明显。我们的结果与文献 [1] 中的最新发现一致，根据该结果，网络中的较低层比最后一层更容易受到域转移的影响。在训练期间， α 能够在几次迭代中收敛到它们的最终值（图 3，右）。

4.3. 与最先进方法的比较

在本节中，我们将我们的方法与最先进的深度域适应方法进行比较。我们首先考虑 Office-31 数据集。我们的评估结果分别总结在表 1 和表 2 中，将提议的 DA 层嵌入 AlexNet 和 Inception-BN 网络中，如第 4.1 节所述。作为基线，我们考虑：深度适应网络 (DAN) [23]、深度域混淆 (DDC) [36]、ReverseGrad 网络 [9]、残差转移网络 (RTN) [25]、联合适应网络 (JAN) [24]、深度重建分类网络 (DRCN) [11] 和 AdaBN [21] 有和没有 CORAL 特征对齐 [31]。与基线方法相关的结果来自原始论文。作为参考，

[25] 中最近的残差传输网络 (RTN)。从表中可以清楚地看出，我们的方法和 RTN 具有大致相同的性能（平均 90.6% 对 90.4%），而它们的性能明显优于其他基线。

最后，我们在 [34] 的跨数据集测试平台的 Caltech-ImageNet 子集上进行了一些实验。如上所述，为了便于与之前也考虑过该数据集的工作进行比较，我们在两种不同的设置中进行了实验。作为基线，我们考虑测地流核 (GFK) [13]、子空间对齐 (SA) [8]、CORAL [31]、传输分量分析 (TCA) [28]、同时深度传输 (SDT) [35] 和 [33] 中的最新方法。表 4 和表 5 显示了我们的结果。所提出的方法显着优于以前的方法，并在该数据集上设置了最新的技术水平。我们方法的更高性能不仅归功于 Inception-BN 的使用，还归功于我们贡献的有效性。事实上，提议的对齐策略，/EAdaBN [21]。

5. 结论

我们提出了 AutoDIAL，这是一种用于无监督、深度域适应的新颖框架。我们贡献的核心是引入了新的域对齐层，它通过将源和目标分布与参考分布相匹配来减少域偏移。我们的 DA 层被赋予了一组对齐参数，这些参数也由网络学习，这使 CNN 不仅可以对齐源和目标特征表示，还可以在每一层自动决定所需的适应程度。我们的框架通过计算 DA 层中的统计数据并引入熵损失来利用目标数据，这促进了对未标记样本具有高置信度的分类模型。我们的实验结果表明，我们的方法优于最先进的领域适应方法。

在基于 AlexNet 架构的深度方法中，AutoDIAL - AlexNet 表现出最佳的平均性能，清楚地展示了所提出的适应策略的好处。在 Inception-BN 网络的实验中发现了类似的结果，我们的方法也优于所有基线。将 AutoDIAL 与 AdaBN 方法 [21] 进行比较是很有趣的，因为这种方法也是从与我们类似的直觉发展而来的。我们的结果清楚地证明了我们贡献的附加值：对齐参数的引入 α ，加上熵正则化项的采用，显着提高了性能。

在我们的第二组实验中，我们分析了几种方法在 Office-Caltech 数据集上的性能。结果如表 3 所示。我们将注意力限制在基于深度架构的方法上，为了公平比较，我们考虑了所有基于 AlexNet 的方法。在这里，我们报告了使用 DDC [36]、DAN [23] 和

虽然本文侧重于无监督域适应的挑战性问题，但我们的方法也可以在半监督环境中加以利用。未来的工作将致力于分析 AutoDIAL 在

方法	来源 目标	亚马逊 单反	亚马逊 网络摄像头	单反 亚马逊	单反 网络摄像头	摄像头 亚马逊	摄像头 单反	平均数
AlexNet – 来源 [19]		63.8	61.6	51.1	95.4	49.8	99.0	70.1
DDC [36]		64.4	61.8	52.1	95.0	52.2	98.5	70.6
丹 [23]		67.0	68.5	54.0	96.0	53.1	99.0	72.9
反向梯度 [9]		67.1	72.6	54.5	96.4	52.7	99.2	72.7
DRCN [11]		66.8	68.7	56.0	96.4	54.9	99.0	73.6
RTN [25]		71.0	73.3	50.5	96.8	51.0	99.6	73.7
一月 [24]		71.8	74.9	58.3	96.6	55.0	99.5	76.0
自动拨号 – AlexNet		73.6	75.5	58.1	96.6	59.4	99.5	77.1

表 1. Office31/完整采样协议上基于 AlexNet 的方法。

方法	来源 目标	亚马逊 单反	亚马逊 网络摄像头	单反 亚马逊	单反 网络摄像头	摄像头 亚马逊	摄像头 单反	平均数
Inception-BN – 来源 [17]		70.5	70.3	60.1	94.3	57.9	100.0	75.5
AdaBN [21]		73.1	74.2	59.8	95.7	57.4	99.8	76.7
AdaBN + 珊瑚 [21]		72.7	75.4	59.0	96.2	60.5	99.6	77.2
DDC [36]		73.2	72.5	61.6	95.5	61.6	98.1	77.1
丹 [23]		74.4	76.0	61.5	95.9	60.3	98.6	77.8
一月 [24]		77.5	78.1	68.4	96.4	65.0	99.3	80.8
自动拨号 – Inception-BN		82.3	84.2	64.6	97.9	64.2	99.9	82.2

表 2. Office31/完整采样协议上基于初始的方法。

方法	来源 目标	亚马逊 网络摄像头 加州理工学院	单反 加州理工学院	加州理工学院 亚马逊网络摄像头	加州理工学院 单反	加州理工学院 网络摄像头	平均数
AlexNet – 来源 [19]		83.8	76.1	80.8	91.1	83.1	84.0
DDC [36]		85.0	78.0	81.1	91.9	85.4	85.0
丹 [23]		85.1	84.3	82.4	92.0	90.6	87.5
RTN [25]		88.1	86.6	84.6	93.7	96.9	90.6
自动拨号 – AlexNet		87.4	86.8	86.9	94.3	96.3	90.3

表 3. 使用完整协议的 Office-Caltech 数据集的结果。

方法	来源 目标	加州理工学院 影像网	影像网 加州理工学院
标准差 [35]		—	73.6
托马西 等。 [33]		—	75.4
Inception-BN – 来源 [17]		82.1	88.4
AdaBN [21]		82.2	87.3
自动拨号 – Inception-BN		85.2	90.5

表 4. 使用 [34] 中的实验设置在跨数据集测试台上的结果。

这种情况。此外，我们计划扩展提议的框架以处理多个源域。

6. 致谢

这项工作的部分创建者是：项目 CHIST-ERA ALOOF，项目 ERC #637076 RoboExNovo（FMC，

方法	来源 目标	加州理工学院 影像网	影像网 加州理工学院
萨 [8]		43.7	52.0
GFK [13]		52.0	58.5
三氯乙烯 [28]		48.6	54.0
珊瑚 [31]		66.2	74.7
Inception-BN – 来源 [17]		82.1	88.4
AdaBN [21]		81.9	86.5
自动拨号 – Inception-BN		84.2	89.8

表 5. 使用 [31] 中的实验设置在跨数据集测试台上的结果。

BC) 和 DIGIMAP 项目，由奥地利研究促进局资助 #860375。

附录

本文件为我们的 ICCV 2017 提交提供了以下额外贡献：

- 在 A 部分，我们为 DA 层中的批量统计以及层的反向传播方程提供了明确的公式；
- 在 B 部分，我们提供了 SVHN – MNIST 基准测试的结果。
- 在 C 部分，我们提供了一些通过 AutoDIAL – Inception-BN 在 Office 31 数据集上学习的特征分布示例。

附录 A. DA 层公式

我们重写方程。(1) 主要论文使样本索引明确：

$$\begin{aligned} \text{是秒} &= \text{DA}(X_{\text{秒}^{\text{世}}}; \alpha) = \sqrt{\frac{X_{\text{秒}^{\text{世}}} - \text{微米}_{st,\alpha}}{\varepsilon + \sigma_{st,\alpha}}} \\ \text{是吨} &= \text{DA}(X_{\text{吨}^{\text{世}}}; \alpha) = \sqrt{\frac{X_{\text{吨}^{\text{世}}} - \text{微米}_{ts,\alpha}}{\varepsilon + \sigma_{ts,\alpha}}} \end{aligned} \quad (7)$$

使用这种表示法，批次统计信息是

$$\begin{aligned} \text{微米}_{st,\alpha} &= \frac{\alpha \sum_{\text{世}=1}^{n_{\text{秒}}} X_{\text{秒}^{\text{世}}} + \frac{1-\alpha}{n_{\text{吨}}} \sum_{\text{世}=1}^{n_{\text{吨}}} X_{\text{吨}^{\text{世}}}}{n_{\text{秒}} - \text{世} + 1} \\ \text{微米}_{ts,\alpha} &= \frac{\frac{1-\alpha}{n_{\text{秒}}} \sum_{\text{世}=1}^{n_{\text{秒}}} X_{\text{秒}^{\text{世}}} + \alpha \sum_{\text{世}=1}^{n_{\text{吨}}} X_{\text{吨}^{\text{世}}}}{n_{\text{吨}} - \text{世} + 1} \\ \sigma_{st,\alpha} &= \frac{\alpha \sum_{\text{世}=1}^{n_{\text{秒}}} (X_{\text{秒}^{\text{世}}} - \text{微米}_{st,\alpha})^2 + \frac{1-\alpha}{n_{\text{吨}}} \sum_{\text{世}=1}^{n_{\text{吨}}} (X_{\text{吨}^{\text{世}}} - \text{微米}_{st,\alpha})^2}{n_{\text{秒}} - \text{世} + 1} \\ \sigma_{ts,\alpha} &= \frac{\frac{1-\alpha}{n_{\text{秒}}} \sum_{\text{世}=1}^{n_{\text{秒}}} (X_{\text{秒}^{\text{世}}} - \text{微米}_{ts,\alpha})^2 + \alpha \sum_{\text{世}=1}^{n_{\text{吨}}} (X_{\text{吨}^{\text{世}}} - \text{微米}_{ts,\alpha})^2}{n_{\text{吨}} - \text{世} + 1} \end{aligned} \quad (8)$$

在哪里 $n_{\text{秒}}$ 和 $n_{\text{吨}}$ 分别是批次中源样本和目标样本的数量。
输入的统计数据是

$$\begin{aligned} \frac{\partial \mu_{st,\alpha}}{\partial \chi_{\text{秒}}} &= \frac{\alpha}{n_{\text{秒}}}, & \frac{\partial \mu_{st,\alpha}}{\partial \chi_{\text{吨}}} &= \frac{1-\alpha}{n_{\text{吨}}}, \\ \frac{\partial \mu_{ts,\alpha}}{\partial \chi_{\text{秒}}} &= \frac{1-\alpha}{n_{\text{秒}}}, & \frac{\partial \mu_{ts,\alpha}}{\partial \chi_{\text{吨}}} &= \frac{\alpha}{n_{\text{吨}}} \end{aligned} \quad (9)$$

$$\begin{aligned} \frac{\partial \sigma_{st,\alpha}}{\partial \chi_{\text{秒}}} &= 2 \frac{\alpha}{n_{\text{秒}}} (X_{\text{秒}^{\text{世}}} - \text{微米}_{st,\alpha}), \\ \frac{\partial \sigma_{st,\alpha}}{\partial \chi_{\text{吨}}} &= 2 \frac{1-\alpha}{n_{\text{吨}}} (X_{\text{吨}^{\text{世}}} - \text{微米}_{st,\alpha}), \\ \frac{\partial \sigma_{ts,\alpha}}{\partial \chi_{\text{秒}}} &= 2 \frac{1-\alpha}{n_{\text{秒}}} (X_{\text{秒}^{\text{世}}} - \text{微米}_{ts,\alpha}), \\ \frac{\partial \sigma_{ts,\alpha}}{\partial \chi_{\text{吨}}} &= 2 \frac{\alpha}{n_{\text{吨}}} (X_{\text{吨}^{\text{世}}} - \text{微米}_{ts,\alpha}). \end{aligned} \quad (10)$$

损失的偏导数 升输入是

$$\begin{aligned} \frac{\partial L}{\partial \chi_{\text{秒}}} &= \sqrt{\frac{1}{\varepsilon + \sigma_{st,\alpha}}} \left[\frac{\partial L}{\partial y_{\text{秒}^{\text{世}}}} - \frac{\alpha}{n_{\text{秒}}} \sum_{\text{世}=1}^{n_{\text{秒}}} \frac{\partial L}{\partial y_{\text{秒}^{\text{世}}}} - \frac{1-\alpha}{n_{\text{吨}}} \sum_{\text{世}=1}^{n_{\text{吨}}} \frac{\partial L}{\partial y_{\text{秒}^{\text{世}}}} \right] \\ &\quad - \sqrt{\frac{1}{\varepsilon + \sigma_{ts,\alpha}}} \frac{1-\alpha}{n_{\text{秒}}} \sum_{\text{世}=1}^{n_{\text{秒}}} \frac{\partial L}{\partial y_{\text{吨}^{\text{世}}}} + \frac{\alpha}{n_{\text{吨}}} \sum_{\text{世}=1}^{n_{\text{吨}}} \frac{\partial L}{\partial y_{\text{吨}^{\text{世}}}}, \\ \frac{\partial L}{\partial \chi_{\text{吨}}} &= \sqrt{\frac{1}{\varepsilon + \sigma_{ts,\alpha}}} \left[\frac{\partial L}{\partial y_{\text{吨}^{\text{世}}}} - \frac{\alpha}{n_{\text{吨}}} \sum_{\text{世}=1}^{n_{\text{吨}}} \frac{\partial L}{\partial y_{\text{吨}^{\text{世}}}} - \frac{1-\alpha}{n_{\text{秒}}} \sum_{\text{世}=1}^{n_{\text{秒}}} \frac{\partial L}{\partial y_{\text{吨}^{\text{世}}}} \right] \\ &\quad - \sqrt{\frac{1}{\varepsilon + \sigma_{st,\alpha}}} \frac{1-\alpha}{n_{\text{吨}}} \sum_{\text{世}=1}^{n_{\text{吨}}} \frac{\partial L}{\partial y_{\text{秒}^{\text{世}}}} + \frac{\alpha}{n_{\text{秒}}} \sum_{\text{世}=1}^{n_{\text{秒}}} \frac{\partial L}{\partial y_{\text{秒}^{\text{世}}}}, \end{aligned} \quad (11)$$

在哪里 是英石 和 是ts 是“交叉归一化”的输出

$$\text{是英石} = \frac{X_{\text{秒}^{\text{世}}}}{\varepsilon + \sigma_{ts,\alpha}}, \quad \text{是ts} = \frac{X_{\text{吨}^{\text{世}}}}{\varepsilon + \sigma_{st,\alpha}} \quad (12)$$

使用这些定义，我们还可以计算偏导数 升写域混合参数 α 作为

$$\begin{aligned} \frac{\partial L}{\partial \alpha} &= \left(\frac{1}{n_{\text{秒}} - \text{世} + 1} \sum_{\text{世}=1}^{n_{\text{秒}}} \frac{\partial L}{\partial y_{\text{秒}^{\text{世}}}} - \frac{1}{n_{\text{吨}} - \text{世} + 1} \sum_{\text{世}=1}^{n_{\text{吨}}} \frac{\partial L}{\partial y_{\text{秒}^{\text{世}}}} \right) \\ &\quad + \left(\frac{1}{n_{\text{秒}} - \text{世} + 1} \sum_{\text{世}=1}^{n_{\text{秒}}} \frac{\partial L}{\partial y_{\text{吨}^{\text{世}}}} - \frac{1}{n_{\text{吨}} - \text{世} + 1} \sum_{\text{世}=1}^{n_{\text{吨}}} \frac{\partial L}{\partial y_{\text{吨}^{\text{世}}}} \right) \\ &\quad + \left(\frac{1}{n_{\text{秒}} - \text{世} + 1} \sum_{\text{世}=1}^{n_{\text{秒}}} \frac{\partial L}{\partial y_{\text{秒}^{\text{世}}}} - \frac{1}{n_{\text{吨}} - \text{世} + 1} \sum_{\text{世}=1}^{n_{\text{吨}}} \frac{\partial L}{\partial y_{\text{秒}^{\text{世}}}} \right) \\ &\quad + \left(\frac{1}{n_{\text{秒}} - \text{世} + 1} \sum_{\text{世}=1}^{n_{\text{秒}}} \frac{\partial L}{\partial y_{\text{吨}^{\text{世}}}} - \frac{1}{n_{\text{吨}} - \text{世} + 1} \sum_{\text{世}=1}^{n_{\text{吨}}} \frac{\partial L}{\partial y_{\text{吨}^{\text{世}}}} \right) \end{aligned} \quad (13)$$

请注意，对于标准的批量归一化，梯度不依赖于层的输入，允许在选择的深度学习框架允许的情况下进行就地计算。

附录 B. SVHN – MNIST 的结果基准

在本节中，我们将在 SVHN [26] 中获得的结果报告给 MNIST [20] 转移基准。我们遵循-

方法	准确性
珊瑚 [21]	63.1
丹 [23]	71.1
丹 [10]	73.9
DRCN [11]	82.0
DSN [3]	82.7
ATN [30]	86.2
自动拨号	90.3

表 6. SVHN 到 MNIST 基准的结果。

降低[10]中的实验协议，使用所有SVHN图像作为源域，所有MNIST图像作为目标域，并与以下基线进行比较：CORAL [31]；深度适应网络 (DAN) [23]；[10] 中的领域对抗神经网络 (DANN)；[11]中的深度重建分类网络 (DRCN)；[3]中的域分离网络 (DSN)；[30] 中的非对称三重训练网络 (ATN)。

与所有基线一样，我们采用 [9] 中的网络架构，在每层之后添加带有参数的 DA 层。训练是从头开始进行的，使用与 AlexNet 相同的元参数（参见主要论文），但有以下例外：初始学习率 $\eta_0 = 0.01$ ；25 个时代；学习率计划定义为 $\eta_{\text{磷}} = \eta_0 / (1 + \gamma p)^\beta$ ，在哪里 $\gamma = 10$ 、 $\beta = 0.75$ 和 磷学习进度是否从 0 到 1。

如表 6 所示，我们在此基准测试中设置了最新的技术水平。值得注意的是，AutoDIAL 也优于 ATN 和 DSN 等方法，这些方法通过添加大量可学习参数来扩展原始网络的容量，而在每个 DA 层中仅使用一个额外的可学习参数。这 α AutoDIAL 在这个数据集上学习的参数绘制在图 4 中。与 Office-31 数据集上的 AlexNet 和 Inception-BN 的情况类似，网络学习到更高的值 α 在网络的底部和较低的值 α 前几名。然而，在这种情况下，我们观察到从 1 到 0.5、有趣的是，这对应于网络中从卷积层到全连接层的转变。

附录 C. 特征分布

在本节中，我们研究来自 AutoDIAL 不同层的一组随机采样特征的分布

– Inception-BN，在 Office 31 数据集的 Amazon-DSLR 任务上学习。在图 5 中，我们比较了这些特征的直方图，在整个源和目标集上计算并取后 DA 层。这些图清楚地显示了我们的 DA 层的对齐效果，因为大多数直方图非常匹配。同样有趣的是，对齐效果似乎与

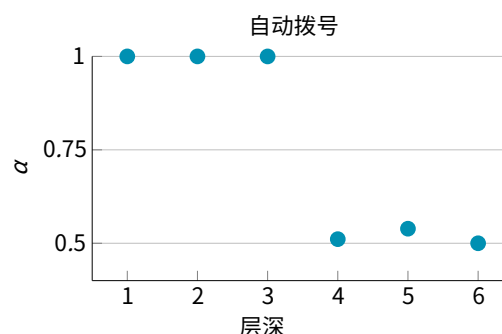


图 4。 α 在 SVHN – MNIST 数据集上学习的参数，绘制为层深度的函数。

分布可能采用的特定形状。

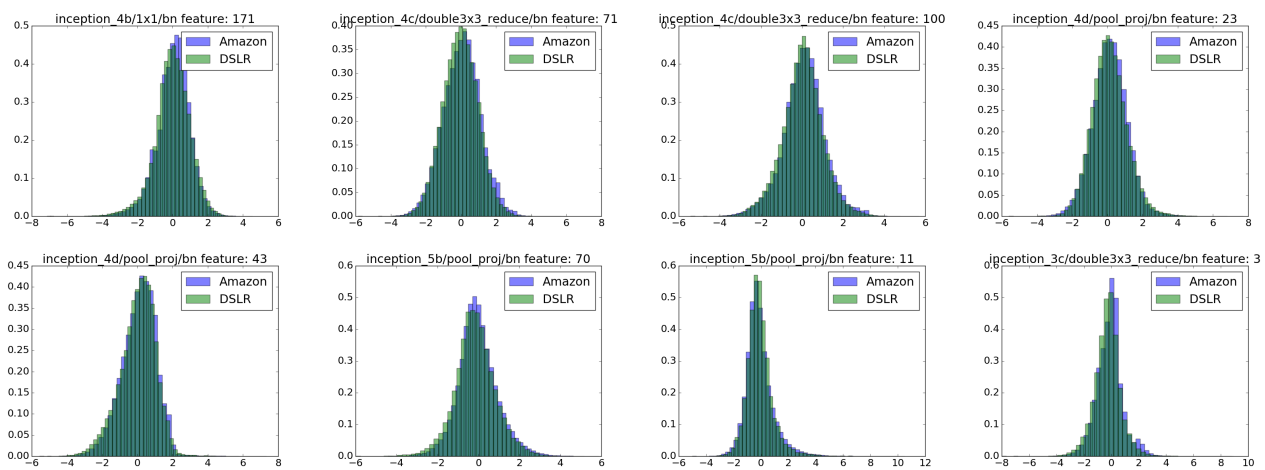


图 5. 来自 AutoDIAL 不同层的随机采样源/目标特征的分布——Inception-BN 在 Office 31 数据集的 Amazon-DSLR 任务中学习（在屏幕上查看效果最佳）。

参考

- [1] R. Aljundi 和 T. Tuytelaars. 通过卷积滤波器重建的轻量级无监督域自适应。在 *ECCV TASK-CV 工作坊*, 2016 年。
- [2] S. Ben-David, T. Lu, T. Luu 和 D. Pl. 域适应的不可能定理。在 *统计局*, 2010 年。
- [3] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan 和 D. 尔罕. 域分离网络。在 *国家知识产权局*, 2016 年。
- [4] FM Carlucci, L. Porzi, B. Caputo, E. Ricci 和 SR Bulò. 只需拨打: 用于无监督域适应的域对齐层。在 *国际商会*, 2017。
- [5] W.-S. Chu, F. De la Torre 和 JF Cohn. 用于个性化面部动作单元检测的选择性转移机。在 *CVPR*, 2013。
- [6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li 和 L. Fei-Fei. Imagenet: 大规模分层图像数据库。在 *CVPR*, 2009。
- [7] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng 和 T. Darrell. Decaf: 用于通用视觉识别的深度卷积激活功能。在 *ICML*, 2014 年。
- [8] B. Fernando, A. Habrard, M. Sebban 和 T. Tuytelaars. 使用子空间对齐的无监督视觉域适应。在 *ICCV*, 2013。
- [9] Y. Ganin 和 V. Lempitsky. 通过反向传播进行无监督域适应。 *ICML*, 2015 年。
- [10] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand 和 V. Lempitsky. 神经网络的领域对抗训练。 *机器学习研究杂志*, 17(59):1–35, 2016。
- [11] M. Ghifary, WB Kleijn, M. Zhang, D. Balduzzi 和 W · 李 用于无监督域适应的深度重建分类网络。在 *ECCV*, 2016 年。
- [12] B. Gong, K. Grauman 和 F. Sha. 将点与地标连接起来: 用于无监督域适应的有区别地学习域不变特征。在 *ICML*, 2013。
- [13] B. Gong, Y. Shi, F. Sha 和 K. Grauman. 用于无监督域适应的测地线流内核。在 *CVPR*, 2012。
- [14] Y. Grandvalet 和 Y. Bengio. 通过熵最小化的半监督学习。在 *国家知识产权局*, 2004。
- [15] G. Griffin, A. Holub 和 P. Perona. Caltech-256 对象类别数据集。2007 年。
- [16] J. Huang, A. Gretton, KM Borgwardt, B. Schölkopf 和阿杰 · 斯莫拉. 通过未标记数据校正样本选择偏差。在 *国家知识产权局*, 2006。
- [17] S. Ioffe 和 C. Szegedy. 批量归一化: 通过减少内部协变量偏移来加速深度网络训练。 *arXiv 预印本 arXiv:1502.03167*, 2015 年。
- [18] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama 和 T. Darrell. Caffe: 用于快速特征嵌入的卷积架构。 *arXiv 预印本 arXiv:1408.5093*, 2014 年。
- [19] A. Krizhevsky, I. Sutskever 和 GE Hinton. 使用深度卷积神经网络进行 Imagenet 分类。在 *国家知识产权局*, 2012。
- [20] Y. LeCun, L. Bottou, Y. Bengio 和 P. Haffner. 基于梯度的学习应用于文档识别。 *继续-IEEE 刊物*, 86(11):2278–2324, 1998。
- [21] Y. Li, N. Wang, J. Shi, J. Liu 和 X. Hou. 重温用于实际领域适应的批量归一化。 *arXiv 预印本 arXiv:1603.04779*, 2016 年。
- [22] M. Long, G. Ding, J. Wang, J. Sun, Y. Guo 和 PS Yu. 传输稀疏编码以实现稳健的图像表示。在 *CVPR*, 2013。
- [23] M. Long 和 J. Wang. 使用深度适应网络学习可转移特征。在 *ICML*, 2015 年。
- [24] M. Long, J. Wang 和 MI Jordan. 使用联合深转移合适网络进行学习。 *arXiv: arXiv 预印本 1605.06636*, 2016 年。
- [25] M. Long, J. Wang 和 MI Jordan. 具有残差传输网络的无监督域自适应。 *国家知识产权局*, 2016 年。
- [26] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu 和 AY Ng. 使用无监督特征学习读取自然图像中的数字。在 *NIPS 深度学习和无监督特征学习研讨会*, 2011 年卷, 第 5 页, 2011 年。
- [27] TJ 奥尼尔. 对未分类观察的正常歧视。 *美国统计学会杂志*, 73(364):821–826, 1978。
- [28] SJ Pan, IW Tsang, JT Kwok 和 Q. Yang. 通过传输组件分析进行域适应。 *IEEE 神经网络汇刊*, 22(2):199–210, 2011。
- [29] K. Saenko, B. Kulis, M. Fritz 和 T. Darrell. 使视觉类别模型适应新领域。在 *ECCV*, 2010 年。
- [30] K. Saito, Y. Ushiku 和 T. Harada. 用于无监督域适应的非对称三重训练。 *arXiv 预印本 arXiv:1702.08400*, 2017。
- [31] B. Sun, J. Feng 和 K. Saenko. 令人沮丧的轻松域适应的回归。在 *美国汽车工业协会*, 2016 年。
- [32] B. Sun 和 K. Saenko. 深珊瑚: 深度域适应的相关对齐。 *arXiv 预印本 arXiv:1607.01719*, 2016 年。
- [33] T. Tommasi, M. Lanzi, P. Russo 和 B. Caputo. 学习视觉域转移的根源。 *arXiv 预印本 arXiv: 1607.06144*, 2016 年。
- [34] T. Tommasi 和 T. Tuytelaars. 跨数据集分析的测试平台。在 *ECCV*, 2014 年。

- [35] E. Tzeng、J. Hoffman、T. Darrell 和 K. Saenko。跨域和任务的同时深度转移。在 *ICCV*, 2015 年。
- [36] E. Tzeng、J. Hoffman、N. Zhang、K. Saenko 和 T. Darrell。深度域混淆：最大化域不变性。*arXiv 预印本 arXiv:1412.3474*, 2014 年。
- [37] M. Yamada、L. Sigal 和 M. Raptis。不留下任何偏差：用于判别 3d 姿态估计的协变量偏移适应。在 *ECCV*, 2012.
- [38] X. Zeng, W. Ouyang, M. Wang, and X. Wang. 用于行人检测的特定场景分类器的深度学习。在 *ECCV*, 2014 年。