Mind-the-Gap! Unsupervised Domain Adaptation for Text-Video Retrieval

Qingchao Chen[1,4*], **Yang Liu** [2,3*] and **Samuel Albanie** [3]

[1] National Institute of Health Data Science, Peking University, Beijing, China
[2] Wangxuan Institute of Computer Technology, Peking University, Beijing, China
[3] Visual Geometry Group, University of Oxford, Oxford, UK
[4] Department of Engineering Science, University of Oxford, Oxford, UK
qingchao.chen@eng.ox.ac.uk, yangl@robots.ox.ac.uk, albanie@robots.ox.ac.uk

Abstract

When can we expect a text-video retrieval system to work effectively on datasets that differ from its training domain? In this work, we investigate this question through the lens of unsupervised domain adaptation in which the objective is to match natural language queries and video content in the presence of domain shift at query-time. Such systems have significant practical applications since they are capable generalising to new data sources without requiring corresponding text annotations. We make the following contributions: (1) We propose the UDAVR (Unsupervised Domain Adaptation for Video Retrieval) benchmark and employ it to study the performance of text-video retrieval in the presence of domain shift. (2) We propose Concept-Aware-Pseudo-Query (CAPQ), a method for learning discriminative and transferable features that bridge these cross-domain discrepancies to enable effective target domain retrieval using source domain supervision. (3) We show that CAPQ outperforms alternative domain adaptation strategies on UDAVR.

Introduction

Given a natural language query and a pool of videos, the goal of text-video retrieval is to rank the videos according to how well their content fits the query. Recent years have seen substantial progress on popular benchmarks for assessing text-video retrieval (Chen and Dolan 2011; Rohrbach et al. 2015; Xu et al. 2016) through effective use of multimodal cues (Mithun et al. 2018) and powerful pretrained models (Miech, Laptev, and Sivic 2018; Liu et al. 2019). These impressive gains have been driven by access to large quantities of labeled data for supervised learning. To date, much of the work in this area has relied on the assumption that the training data and test data arise from the same domain. As a consequence, the use of text-video retrieval methods in *novel domains* mandates the gathering of corresponding annotation such that models can be retrained or fine-tuned on the target data.

One possible solution to this challenge can be found in zero-shot text-video retrieval, in which one assumes no access to any training content from the target domain. Indeed, recent methods employing large-scale pretraining (Miech et al. 2019, 2020) have shown promising early results pursuing such an approach. It is reasonable to believe, however, that for many problems of interest, the zero-shot assumption is overly constraining. In this work, we therefore consider instead the less restrictive formulation in which a model has access to labelled data on a source domain and only unlabelled data on the target domain of interest, a setting commonly referred to as Unsupervised Domain Adaptation (UDA).

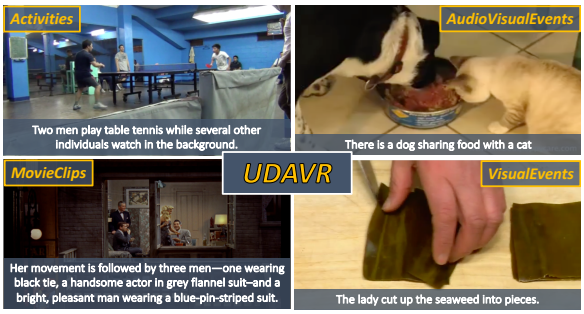

Figure 1: The Unsupervised Domain Adaptation for Video Retrieval (UDAVR) benchmark. In this work, we repurpose existing datasets from across four domains to study the task of text-video retrieval without target domain supervision. The videos from these four domains differ not only in visual composition and duration, but also in the focus and style of their text descriptions.

Motivated by the considerable importance of video data in both academic and commercial settings, recent works have explored video-based UDA scenarios for action recognition (Jamal et al. 2018; Chen et al. 2019b; Munro and Damen 2020). To date, however, unsupervised domain adaptation for text-video retrieval has received limited attention in the literature.

In this work we propose a new benchmark to enable investigation of the text-video retrieval task under this regime. To this end we draw on existing datasets from four domains highlighted in Fig. 1: these domains are *Activities* (sourced from ActivityNet-Captions (Krishna et al. 2017)), *AudioVisualEvents* (sourced from MSR-VTT (Xu et al. 2016)), *MovieClips* (sourced from LSMDC (Rohrbach et al. 2017a)) and *VisualEvents* (sourced from MSVD (Chen and Dolan 2011)).

[*] indicates equally contributed first and corresponding author.

A naive application of empirical risk minimisation suffers from two kinds of domain shift in adapting retrieval models: *video content/style shift* and *description distribution shift*. The former is visible in Fig. 1, in which we observe the shift from a hand-held camcorder showing a game of table tennis (top-left) to the cinematic style of Alfred Hitchcock (bottom-left). The latter, a shift in description distribution, is a key difference from other applications of UDA. Note that differences in description distributions are typically driven by differences in the description style among the teams of annotators that produced each domain. We observe that the Descriptive Video Services (which target a rich storyline for the visually impaired) employed by (Rohrbach et al. 2017b) that are responsible for the movie clip description in Fig. 1 (bottom-left) can produce descriptions that differ substantially from those produced by Amazon Mechanical Turk workers (such as those used by the efficient pipelines used for the three remaining domains in Fig. 1).

To tackle these challenges, we propose the concept-Aware-Pseudo-Query (CAPQ) framework for cross-domain text-video retrieval, which comprises two technical contributions: a concept preservation regulariser that seeks to enhance the *transferability* of the learned embeddings (their invariance to both visual and description distribution shift); and a pseudo-labelling algorithm that aims to ensure that they are as *discriminative* as possible in order to boost retrieval performance on a target domain without access to the description distribution. Additionally, to mitigate the *hubness problem* (Radovanovic, Nanopoulos, and Ivanovic 2010) (in which a small portion of the data samples become "popular" i.e. they form the nearest neighbors of many samples) that can arise from a naive application of pseudo-labelling (Liu and Ye 2019), we propose an iterative, mutual-exclusion selection mechanism that avoids over-exploitation of a small number of pseudo-label candidates.

In summary, we make the following contributions: (1) We propose, to the best of our knowledge, the first benchmark for natural language text-video unsupervised domain adaptation, UDAVR, and employ it to assess the suitability of existing methods for this task. (2) We propose CAPQ, a method which employs source supervision and unlabelled target data to achieve good target domain retrieval performance. (3) We demonstrate that CAPQ outperforms source-only generalisation as well as alternative domain adaptation strategies such as variants of maximum mean discrepancy, adversarial learning strategy and transportation modelling on the proposed benchmark.

Related Work

Text-Video Retrieval. In recent years, cross-modal *joint embeddings* (Otani et al. 2016; Dong, Li, and Snoek 2016) have emerged as a popular mechanism for text-video retrieval. Further developments have sought to develop structured joint spaces (Mithun et al. 2018; Wray et al. 2019), explore large-scale pretraining (Miech et al. 2019) and integrate multiple modalities, learning from experts that operate on both visual and audio sensory data (Miech, Laptev, and Sivic 2018; Liu et al. 2019; Mithun et al. 2018; Gabeur et al. 2020). This latter methodology has proven particularly effective for text-video retrieval, and we use it as a test-bed for our approach. However, differently from these works (which train and test on the same domain), we explore the task of UDA for text-video retrieval.

Unsupervised Domain Adaptation.

There has been significant research interest in developing UDA techniques, including variants of MMD (Long et al. 2015; Long, Wang, and Jordan 2016; Sun and Saenko 2016; Long et al. 2016; Tzeng et al. 2014; Yan et al. 2017; Venkateswara et al. 2017), adversarial learning (Long et al. 2018; Liu and Tuzel 2016; Tzeng et al. 2017; Bousmalis et al. 2016; Liu, Breuel, and Kautz 2017; Russo et al. 2017; Chen and Liu 2020) and transportation plan modelling (Courty et al. 2017; Chen et al. 2018; Bhushan Damodaran et al. 2018; Haeusser et al. 2017) to measure and reduce domain discrepancy. In the experiments section, we compare our proposed approach with representative methods from each of these methods on the UDAVR benchmark. More recently, UDA methods have been extended to object detection (Hsu et al. 2020) and variants of action recognition on video data (Jamal et al. 2018; Chen et al. 2019b; Munro and Damen 2020). Perhaps the setting most closely related to ours is (Cao et al. 2018) who consider the task of text-image retrieval. Differently from our approach, they restrict the query text to be a single word (falling within a predefined set of categories), rather than a open-set free-form sentence query. We also note that pseudo-labelling techniques (Zhang et al. 2018; French, Mackiewicz, and Fisher 2017) have been investigated in prior UDA work for image classification. However, such approaches make use of discrete class labels and are consequently not directly applicable in our video-text retrieval application setting (for which discrete class boundaries are unavailable).

Method

Overall Framework

In this section, we describe our overall framework for domain adaptation in the video retrieval setting and introduce some basic notation. We assume access to a source domain $S = \{v^S, t^S\}$ of paired text and video samples and a target domain $T = \{v^T\}$ of unpaired videos. The goal of this paper is to construct a model which is able to learn discriminative and transferable features that bridge the cross-domain discrepancy and learn a good joint embedding space for target domain retrieval. Crucially, it must do so in the absence of knowledge of the the target domain text distribution t^T.

As noted in the introduction, key challenges to learning a good joint embedding space under this setting include: (1) achieving robustness to visual content/style shift and description distribution shifts; (2) learning discriminative features via the source domain that are transferable to an unknown open set (free-form target queries).

We propose the CAPQ framework to tackle these challenges (Fig. 2). CAPQ comprises a feature extractor F, a cross-domain video encoder ϕ_{vid}, a text encoder ϕ_{text}, a concept selector ψ_{C} and a hallucinator ψ_{H}. We discuss these

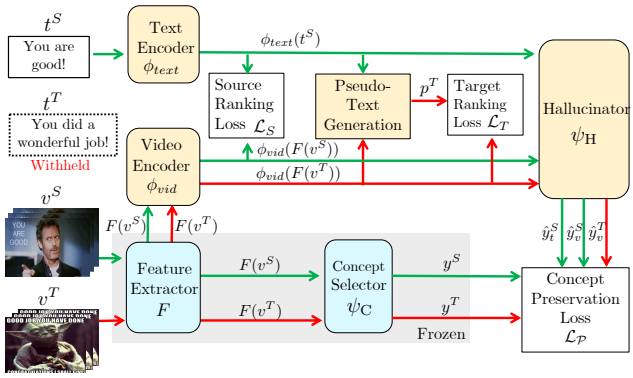

Figure 2: The components of the proposed CAPQ framework. Green and red arrows denote the flow of information from the source and target domains, respectively. First, we adopt a collection of frozen, pretrained expert models, F, to extract generic features $F(v^S)$ and $F(v^T)$, covering an array of semantic concepts. Next, text and video encoders project generic features to the text-video joint embedding space to produce transferable and discriminative features $\phi_{\text{vid}}(v^S)$, $\phi_{\text{vid}}(v^T)$ and $\phi_{\text{text}}(t^S)$. These properties are achieved through the multi-modality concept preservation loss $\mathcal{L}_\mathcal{P}$ and ranking losses $\mathcal{L}_\mathcal{S}$, $\mathcal{L}_\mathcal{T}$ in both source and target domains. To enable the ranking loss in target domain without target text t^T, we design a novel pseudo-text selection module to select "pseudo text" p^T for $\phi_{\text{vid}}(v^T)$ from the pool of unbiased source text embedding $\phi_{\text{text}}(t^S)$.

components and their interactions in the following.

Feature Extractor: We first adopt a *frozen* feature extractor F, which comprises a collection of models (often referred to in the literature as experts (Miech, Laptev, and Sivic 2018; Liu et al. 2019)) that have been pre-trained (on tasks such as image classification, action recognition, etc.) to extract features from source and target domain videos. Descriptors $F(v^S), F(v^T)$ are intended to form a generic representation of the content (this is achieved by employing a pretrained models that cover a wide range of semantic concepts).

Video and Text Encoder: The video encoder ϕ_{vid} takes a generic video descriptor $F(v)$ as input and projects it into the joint text-video embedding space. Similarly, the text encoder ϕ_{text} first maps each query sentence t to a set of feature vectors using pre-trained word-level embeddings (Mikolov et al. 2013), then aggregates the resulting word embeddings via NetVLAD (Arandjelovic et al. 2016) and projects the results to produce the final text representation $\phi_{\text{text}}(t)$ for the retrieval task. We aim to learn a good joint embedding space, where the features are both transferable and discriminative to optimize the performance of target domain retrieval. For simplicity, we denote $\phi_{\text{vid}}(F(v))$ as $\phi_{\text{vid}}(v)$ in the following section unless otherwise specified.

Transferable: To make the feature transferable, we aim to reduce both the video embedding shift and description distribution shift between two domains. This can be achieved by leveraging the constraints of the multi-modality experts' pretrained models via a concept preservation loss. The Video Encoder ϕ_{vid} and Text Encoder ϕ_{text}, *Concept Selector* ψ_{C} and *Hallucinator* ψ_{H} work cooperatively to minimize the concept preservation loss $\mathcal{L}_P$, which aims to preserve previously acquired knowledge by penalising joint space embeddings that are unable to retain the discriminative signal provided by the pre-trained models. In this way, $\mathcal{L}_P$ tends to encourage the joint visual and text embedding to be concept-aware and domain-agnostic. In doing so, we can (1) implicitly reduce the source and video embedding shift in the joint space under this constraint; (2) encourage the text encoder to map annotations of a given style $A_S(v)$ to a more generic text embedding $\phi_{\text{text}}(t)$, where the function $A_S(\cdot)$ represents the description style associated with the source domain annotation. The objective of this design is to give $\phi_{\text{text}}(t)$ access to a wider coverage of diverse semantic concepts contained in the paired video v that can in principle be leveraged to answer unknown queries from the target domain.

Discriminative: The features should be discriminative, i.e., the embeddings for paired text and video should lie close together, while embeddings for text and video that do not match should lie far apart. Intuitively, we can use available training pairs $\{v^S, t^S\}$ via a ranking loss $\mathcal{L}_S$ to make the embedding as discriminative as possible, but only in the source, rather than the target domain. To address this, we propose a pseudo-text mutually-exclusive selection mechanism, selecting from the best collection of unbiased text embeddings $\phi_{\text{text}}(t)$ (as noted above, these embeddings are designed to minimise the influence of the source domain annotation style) and assign it to the target video as a target pseudo query embedding p^T. We then refine the joint video-text embedding space by minimizing the second ranking loss L_T between the pseudo pairs in the target domain $\{v^T, p^T\}$.

Combining the above two designs together, we reach the overall training objective of our CAPQ framework:

$$\min_{\phi_{\text{vid}},\phi_{\text{C}},\psi_{\text{H}}} \mathcal{L} = \mathcal{L}_\mathcal{S} + \lambda_1 \mathcal{L}_\mathcal{P} + \lambda_2 \mathcal{L}_\mathcal{T}. \quad (1)$$

where λ_1 and λ_2 are the weighting coefficients.

Concept Preservation

The Concept Preservation loss $\mathcal{L}_P$ is designed to preserve the previously acquired knowledge to make both video and text features in the joint space transferable. Concretely, the concept selector ψ_{C} first maps a generic video descriptor $F(v)$ (the output of feature extractor module F) to a concept distribution $y = \psi_{\text{C}}(F(v))$ associated with the original pretraining task ($y \in R^C$ where C denotes the number of external concepts in the original classification task). For example, video retrieval systems (e.g. (Miech, Laptev, and Sivic 2017)) often make use of models pretrained to perform image classification on ImageNet (this model would then form part of the feature extractor F). In this case, the concept selector represents the final linear layer of the pretrained model, which is responsible for transforming a fragment of the generic descriptor $F(v)$ to a distribution over the 1000 concepts of ImageNet, $y \in R^{1000}$. We use the same concept selector to project both the source and target generic video descriptors $F(v^S)$ and $F(v^T)$ to their respective concept distributions, i.e., $y^S = \psi_{\text{C}}(F(v^S))$ and $y^T = \psi_{\text{C}}(F(v^T))$, in order to provide a common signal for

both domains. Since video-text pairings are available in the source domain, we propagate the predicted external concept distribution for the source video, y^S, to its paired text embedding. Equivalently, given a video-text pair $\{v^S, t^S\}$, we require that they are mapped to the same concept distribution y^S if the text describes the video.

Next, we use the predicted concept distribution y, as a signal to encourage the joint text-video embeddings to preserve their knowledge of concepts present in the pretrained models. Specifically, we construct a hallucinator, ψ_{H} (implemented as a two layer MLP), whose goal is to make predictions $\hat{y} \in R^C$ from the embeddings ϕ_{vid} and ϕ_{text} that match those of the concept selector, y. Intuitively, to be able to generate these predictions accurately, the embeddings must retain the ability to distinguish between external concepts known to the pretrained models. As for the concept selector, we use the same hallucinator ψ_{H} for both the text and video embeddings, implicitly encouraging feature alignment between $\phi_{\text{vid}}(v)$ and $\phi_{\text{text}}(t)$. We implement this idea through a concept preservation loss, $\mathcal{L}_P$, which is Kullback–Leibler divergence between the concept distribution y and the hallucinator predictions $\hat{y}$, as follows:

$$\mathcal{L}_{\mathcal{P}}(y, \hat{y}) = \frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} -y_{i,c}\log(\frac{\exp(\hat{y}_{i,c})}{\sum_{c'=1}^{C}\exp(\hat{y}_{i,c'})}), \quad (2)$$

where N is the number of samples and C is the number of classes. The concept distribution y provides a soft label to guide feature learning. Note that in the source domain, we can maximally preserve external concept in both the video and text embeddings by minimising $\mathcal{L}_{\mathcal{P}}(y^S, \hat{y}_v^S) + \mathcal{L}_{\mathcal{P}}(y^S, \hat{y}_t^S)$; in the target domain, since the text queries are not available, we instead preserve only the concept in the target video embedding at training, i.e., by minimising $\mathcal{L}_P(y^T, \hat{y}_v^T)$. The overall concept preservation loss is:

$$\mathcal{L}_{\mathcal{P}} = \mathcal{L}_{\mathcal{P}}(y^S, \hat{y}_v^S,) + \mathcal{L}_{\mathcal{P}}(y^S, \hat{y}_t^S) + \mathcal{L}_{\mathcal{P}}(y^T, \hat{y}_v^T,) \quad (3)$$

By preserving the previously acquired knowledge to perform the pretraining classification tasks, the concept preservation loss encourages: (1) feature alignment between $\phi_{\text{vid}}(v_s)$ and $\phi_{\text{vid}}(v_t)$ as they can be transferred to the semantic external concepts by using a shared hallucinator; (2) the output of the text encoder $\phi_{\text{text}}(t)$ to cover a wider range of semantic concepts that could describe the paired video v, as the paired $\phi_{\text{vid}}(v)$ and $\phi_{\text{text}}(t)$ shared the same concept distribution. Note that concept preservation is performed only during training. For inference, the concept selector is discarded (there is consequently no run-time penalty to this approach).

To summarize, the concept preservation loss $\mathcal{L}_{\mathcal{P}}$ encourages learning of domain-agnostic joint video-text embeddings by retaining and preserving semantic concepts that were encountered during pretraining. In addition, $\mathcal{L}_{\mathcal{P}}$ aims to hallucinate from a domain-specific, biased joint embedding to a more diverse and richer embedding by eliminating the domain-specific description styles.

Discussion on Multi-Modality Features: As noted in the related work section, several recent text-video retrieval methods adopt multi-modal pretrained model features to achieve robustness. Note CAPQ is directly applicable to the multi-modal setting by using N_M feature extractors $F \triangleq \{F_n, n \in [1, 2, ..., N_M]\}$. The multi-concept preservation loss extends Eqn. (3) as follows:

$$\hat{\mathcal{L}}_{\mathcal{P}} = \frac{1}{N_M}\sum_{n=1}^{N_M} L_{P,n}. \quad (4)$$

.

Discriminative Joint Space Learning

Given video-text pairs from the source domain, we can train discriminative embeddings for retrieval by minimizing a contrastive margin loss (Karpathy and Fei-Fei 2015) $\mathcal{L}_S$:

$$\mathcal{L}_S = \frac{1}{N_b}\sum_{i=1, j\neq i}^{N_b} \max(0,\ m + s_{i,j}^S - s_{i,i}^S) + \max(0,\ m + s_{j,i}^S - s_{i,i}^S), \quad (5)$$

where N_b is the batch size, m is a margin, and the $s_{i,j}^S$ represents the similarity score between the i^{th} source video v_i^S and the j^{th} text description t_j^S as:

$$s_{i,j}^S = \frac{\phi_{\text{vid}}(v_i^S)\cdot\phi_{\text{text}}(t_j^S)}{|\phi_{\text{vid}}(v_i^S)||\phi_{\text{text}}(t_j^S)|}. \quad (6)$$

To perform well on the cross-domain retrieval task, we propose to refine the joint video-text embedding space (trained from $\mathcal{L}_S$) to fit the requirements of discriminative target domain retrieval. To do so, we propose a pseudo-text selection mechanism, selecting from the collection of unbiased text embeddings $\phi_{\text{text}}(t^S)$ and assign the 'best' to the target video v^T as target pseudo text embedding p^T. We then refine the joint space by minimizing the second ranking loss $\mathcal{L}_T$ between the target video embedding $\phi_{\text{vid}}(F(v^T))$ and the selected pseudo query embedding p^T as a similar form to (5), in which the similarity computed for target domain pairings over the sampled mini-batches $s_{i,j}^T$ can be calculated as:

$$s_{i,j}^T = \frac{\phi_{\text{vid}}(v_i^T)\cdot p_j^T}{|\phi_{\text{vid}}(v_i^T)||p_j^T|}. \quad (7)$$

Mutually-Exclusive Selection Algorithm: Intuitively, given a collection of unbiased text embeddings $\phi_{\text{text}}(t^S)$ (by unbiased we mean, unbiased towards a particular annotator distribution), it is straight-forward to select the best one for each target video feature $\phi_{\text{vid}}(v_j^T)$ by selecting the text embedding that yields the highest similarity score (i.e. maximises $S(\phi_{\text{vid}}(v_j^T), \phi_{\text{text}}(t^S))$). However, this naive selection mechanism is problematic, as the selected pseudo text embedding for one target video v_i, might also generate high similarity scores for other target videos, v_j, where $i \neq j$, particularly in the early stages of training. This observation motivates the design of a *mutually-exclusive* pseudo-text embedding selection process (described below), which operates on a collection of N_Q unbiased text embeddings $\{\phi_{\text{text}}(t_i^S)\}_{i=1}^{N_Q}$, and a minibatch of N_B target video features $\{\phi_{\text{vid}}(v_j^T)\}_{j=1}^{N_B}$. Concretely, we perform a bi-directional softmax operation to refine similarity matrix prior to selection to

enable enforcement of the desired mutual-exclusion property. Given a similarity matrix S, we first amplify the discriminative capability of the text and compute S_{text} by applying the softmax function along the text dimension of S; we also apply the softmax function along the video dimension of S to obtain S_{video} to amplify video discriminability. We then refine the similarity matrix by taking two directions into account via $S' = S_{\text{text}} \cdot S_{\text{video}}$. The selection assignment p_j for the j^{th} video embedding $\phi_{\text{vid}}(v_j^T)$ follows from the refined similarity matrix S' by selecting the unbiased text embedding that yields the highest similarity score. Note that in our proposed selection process, the selected source instance embedding is not the conventional class or cluster category label, but an unbiased text instance in the joint embedding.

The proposed mutually-exclusive pseudo-text selection method is particularly designed for cross-modal retrieval task, because the method first looks through all candidate texts and videos, establishes a smooth similarity graph and finally assigns the "best" pseudo-texts which are **not** the nearest neighbors of other different video queries. This is the key difference compared with conventional pseudo labelling mechanism used in classification tasks, where no penalty is applied when the same pseudo label is assigned for different visual queries (i.e. those that fall within the same cluster).

Analysis

To provide insight into the function of the CAPQ framework, we conduct an analysis of our approach through the lens of non-conservative domain adaptation (Ben-David et al. 2010) inspired by the analysis of (Chen et al. 2019a). Let $\mathcal{H}$ denote the hypothesis class and S, T, $\hat{T}$, the source, target and pseudo-target domain(paired target video samples with the pseudo text selected by CAPQ). The target domain *risk* (expected error), $\epsilon_T(h)$, associated with hypothesis $h \in \mathcal{H}$ can be bounded by three terms:

$$\epsilon_T(h) \leq \epsilon_S(h) + \frac{1}{2} d_{H\Delta H}(S,T) + E, \tag{8}$$

where $\epsilon_S(h)$ denotes source domain risk, $d_{H\Delta H}(S,T)$ is a measure of discrepancy between the distributions S and T, and $E = \epsilon_S(h^{'}, f_S) + \epsilon_T(h^{'}, f_T)$ represents the shared error of the ideal joint hypothesis ($h^{'}$ denotes the ideal joint hypothesis and f_S and f_T represent the true labeling functions for source and target respectively).

In CAPQ, the $\epsilon_S(h)$ term is minimized by the source domain bidirectional loss $\mathcal{L}_S$. To minimize the domain discrepancy $d_{H\Delta H}(S,T)$, we use a common set of robust pretrained, multi-modality visual and language model encodings (as anchors without fine-tuning) which are subsequently aligned by the proposed transferable concept preservation operations. However, it has been noted that minimizing the first two terms is insufficient because E can grow even when the cross-domain marginals are exactly aligned (Chen et al. 2019a) (see also (Wu et al. 2019)). We illustrate in the following that E can be bounded with the proposed mutually-exclusive pseudo-text selection method, and the upper bound can be reduced by the design of CAPQ.

As target annotations are not available, we reformulate the shared error E based on the pseudo video-text pairs in the target domain $\hat{T}$ as follows: $E = min_h\ \epsilon_S(h, f_S) + \epsilon_{\hat{T}}(h, f_T)$, where $\epsilon_{\hat{T}}$ is the expected risk error on $\hat{T}$. Based on the triangle inequality, the upper bound of E can be derived as the following: $E \leq min_h\ \epsilon_S(h, f_S) + \epsilon_{\hat{T}}(h, f_{\hat{T}}) + 2\epsilon_{\hat{T}}(f_S, f_{\hat{T}}) + \epsilon_{\hat{T}}(f_T, f_{\hat{T}})$. In the following discussion, we illustrate how different modules in CAPQ can reduce the bounded terms.

Discussion. First, minimization of $\epsilon_S(h, f_S)$ and $\epsilon_{\hat{T}}(h, f_{\hat{T}})$ is achieved by minimizing the source and pseudo-target ranking losses, $\mathcal{L}_S$ and $\mathcal{L}_T$, respectively.

Second, the proposed mutually-exclusive pseudo-selection algorithm with the pseudo-target ranking loss $\mathcal{L}_T$ aims to progressively align cross-domain visual features at the level of the text descriptions. For example, using the pseudo-selection algorithm, the $\phi_{\text{vid}}(v_j^T)$ may be assigned to the same text description as $\phi_{\text{vid}}(v_i^S)$. In these cases, the risk $\epsilon_{\hat{T}}(f_S, f_{\hat{T}})$ is expected to be minimized.

Third, the term $\epsilon_{\hat{T}}(f_T, f_{\hat{T}})$ represents the false labelling rate of target samples. To minimize this term, we employed two procedures: 1) in the proposed mutually-exclusive pseudo-text selection, the method can robustly pick different source text descriptions to target visual queries. It aims to diversify the candidate pool of pseudo-text selection for robust assignment and to mitigate hubness. 2) leveraging the transferable concept preservation design $\mathcal{L}_P$, the constraint of "learning a visual embedding with external, generic concepts" helps select a reliable pseudo text and can restrict the contribution to inhibit the accumulation of risk $\epsilon_{\hat{T}}(f_T, f_{\hat{T}})$.

Experiments

In this section, we introduce the UDAVR text-video retrieval benchmark and evaluate the proposed framework. We first conduct a detailed domain shift analysis on UDAVR. Then, using this analysis, we select four adaptation directions from among the set of possible adaptations between the four domains. Next we compare our proposed CAPQ with existing retrieval and UDA methods. Finally, we present an ablation study to analyze the model configurations of the proposed methods.

The UDAVR Benchmark

The UDAVR benchmark consists of 160k videos, sourced from four datasets spanning different domains; (1) *Audio-VisualEvents* from the MSR-VTT dataset (Xu et al. 2016), which gathered videos (together with audio tracks) from YouTube using popular user queries; (2) *VisualEvents* from the MSVD dataset (Chen and Dolan 2011) which was similarly sourced from YouTube but was curated such that each video contained a single unambiguous event, did not include overlaid text or subtitles and had the sound track removed before captioning; (3) *Activities* from the ActivityNet-Captions dataset (Krishna et al. 2017) was sourced from web videos (each containing multiple sequential events), with a focus on complex human activities; (4) *MovieClips* from LSMDC (Rohrbach et al. 2015) comprising short video clips from movies. The statistics of each domain are summarised in Table 1, (number of videos/sentences, average length

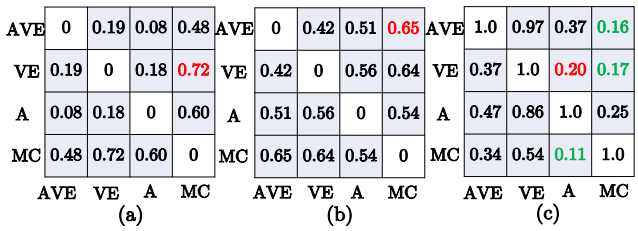

Figure 3: Cross domain statistics.(a) MMD of the pre-trained video feature distribution.(b) The Jensen-Shannon divergence (JSD) of the vocabulary distributions between any pair of datasets.(c) Recovery Matrix:Ratio between source-only and target-only model performances. Row indicates source domain, column indicates target domain.

of videos/sentences, the audio track availability, annotation process). In the following, we first analyze three main factors leading to domain gaps in UDAVR, including *video shift*, *text shift*, and *annotation function shift*. Then we report the performance recovery matrix and finally define a set of adaptation directions which specifically target video shift ($\text{split}_{\text{Video}}$), text shift ($\text{split}_{\text{Text}}$), annotator shift ($\text{split}_{\text{AnnoF}}$) as well as the overall most difficult shift ($\text{split}_{\text{Hard}}$) based on these analysis.

Video Shift: We employ Maximum Mean Discrepancy (MMD) (Gretton et al. 2012) with pretrained video features to measure the video shift across domains (Fig. 3(a)). We select the most difficult direction (with greatest MMD) VisualEvents→MovieClips and denote it $\text{split}_{\text{Video}}$.

Text Shift: We report the Jensen–Shannon divergence (JSD) of vocabulary distribution among all pairs of domains in Fig. 3(b). We select AudioVisualEvents→MovieClips (which exhibits the greatest text shift) and denote it $\text{split}_{\text{Text}}$.

Annotation Function Shift: The annotation function shift across domains is closely linked to annotator style. We select MovieClips→Activities as representative for $\text{split}_{\text{AnnoF}}$ since: (1) As shown in Table 1, annotations from the MovieClips domain are generated via Descriptive Video Service (providing rich captions to accompany the dialogue) while Activities adopt different Amazon Mechanical Turk (AMT) pipelines. (2) Large differences in duration also increase the Annotation Function Shift (i.e., Activities videos require a paragraph summary of all major events in temporal order).

Empirical Difficulty: We adopt the performance recovery matrix to measure the domain shift with all three factors (mentioned above) coupled. We set up a cross-domain video retrieval baseline for each transfer task, i.e., the ratio between source-only and target-only performances (the geometric mean of R@1, R@5 and R@10) with both models tested on target data, as shown in Fig. 3(c). It can be seen that $\text{split}_{\text{Video}}$, $\text{split}_{\text{Text}}$ and $\text{split}_{\text{AnnoF}}$ are most challenging in that they have the three lowest performance recovery ratios in Fig. 3(c). We select VisualEvents→Activities for $\text{split}_{\text{Hard}}$ since it has the lowest performance recovery ratios among the remaining directions.

Results and Comparisons

We evaluate the proposed method using four splits (adaptation directions) of UDAVR defined in the previous section . We adopt the standard retrieval metrics of the target domain dataset, namely R@K (K = 1,10) and median rank (MR), where R@K (recall at K) represents the percentage of test queries for which at least one relevant item is found among the top-K retrieved results. The results of four splits are shown in Tables 2 and 3.

In each split, we report the performance of our CAPQ and comparisons with six existing methods. These include two strong video-retrieval network architecture methods (Miech, Laptev, and Sivic 2017; Liu et al. 2019) and four unsupervised adaptation baselines (Ganin et al. 2016; Long et al. 2015; Sun and Saenko 2016; Courty et al. 2017). Note that neither of the video-retrieval approaches (MoEE (Miech, Laptev, and Sivic 2018) and CE(Liu et al. 2019)) investigated the cross-dataset adaptation problem. Since CE outperforms MoEE slightly in all four datasets in the conventional video-retrieval setting, we adopt CE as our backbone architecture and report the CE results as a solid source-only (SO) baseline. For fair comparison, the adaptation methods we consider, namely Maximum Mean Discrepancy (MMD), adversarial feature alignment (DANN), Deep Coral (D-CORAL) and optimal transport (OT) each use the same backbone network as CAPQ.

To summarise: (1) CAPQ outperforms MMD, DANN, D-CORAL and OT based adaptation methods on all adaptation directions across the four splits. Specifically, our method outperforms prior approaches by a considerable margin, notably achieving a relative gain in geometric mean (R1,R10) of approximately 52%, 27%, 29% and 23% over the second best on each of the four splits respectively. (2) Conventional domain adaptation techniques are effective for addressing the video content and text shifts, but not for annotation function shifts. Specifically, as shown in Table 3, when encountering different annotation styles, a direct application of conventional domain adaptation techniques can under-perform the source-only model.

Ablation Study and Analysis

In this section, we perform ablation studies on the $\text{split}_{\text{Text}}$ adaptation task and investigate the effectiveness of network components and the use of pre-trained features in the concept preservation loss. Finally, we provide qualitative results via visual examples.

Architectural Variants: We conduct a detailed ablation study by examining the effectiveness of each proposed component. As shown in Table 4, each combination of the proposed modules yields improved performance compared with the SO baseline (without adaptation) under all evaluation metrics. We observe that pseudo-text selection contributes the most significant performance gain as an individual module, suggesting that refining the joint text-video embedding space is valuable by minimizing the second ranking loss between the pseudo text queries and the target videos. With both the concept preservation and pseudo-text selection, CAPQ performs the best, achieving a relative

Domain / Source Dataset	#Videos	Video Length	Audio	Sentence Source	#Sentence	Query Length
AudioVisualEvents / MSR-VTT	10,000	20s	Yes	AMT workers	200,000	9.34
VisualEvents / MSVD	1970	10s	No	AMT workers	70,028	7.03
Activities / ActivityNet-Captions	14,926	180s	Yes	AMT workers	54,926	49.76
MovieClips / LSMDC	102,046	4s	Yes	DVS +Scripts	102046	9.75

Table 1: Comparison of four domains in UDAVR.

Method	$split_{Video}$						$split_{Hard}$					
	t2v			v2t			t2v			v2t		
	R1	R10	MR	R1	R10	MR	R1	R10	MR	R1	R10	MR
MoEE	0.8	6.9	195	0.9	3.6	271	2.0	11.3	160	2.3	11.8	135
SO	1.0	7.1	187	0.9	3.7	267	2.2	13.2	110	2.6	13.8	102
MMD	1.5	7.9	230	1.1	5.7	223	2.9	14.6	108	2.5	14.2	98
D-CORAL	0.9	4.2	312	0.4	3.3	327	2.1	12.5	165	2.2	13.2	140
DANN	1.3	7.3	281	1.0	5	246	2.6	14.3	120	2.3	13.7	115
OT	0.8	4.0	313	0.4	3.8	322	2.5	12.2	143	2.2	12.8	120
CAPQ (ours)	2.3	10.5	164	1.7	9.7	162	3.7	19.1	64.3	3.0	16.3	70

Table 2: Performance comparison on $split_{Video}$ and $split_{Hard}$

Method	$split_{Text}$						$split_{AnnoF}$					
	t2v			v2t			t2v			v2t		
	R1	R10	MR	R1	R10	MR	R1	R10	MR	R1	R10	MR
MoEE	1.4	9.9	162	1.1	7.5	196	1.2	7.5	265	1.3	6.7	272
SO	1.5	10.3	162	1.6	9.3	190	1.3	7.7	267	1.4	6.8	275
D-CORAL	2.5	10.5	149	1.9	9.5	177	1.0	6.4	325	0.9	6.4	325
MMD	1.7	10.4	155	2.0	8.3	187	1.2	7.1	265	1.5	9.0	243
DANN	1.9	9.3	178	1.7	9.6	176	1.0	5.6	429	1.1	5.8	311
OT	2.3	10.7	151	1.9	10.2	175	1.0	6.2	325	0.8	6.2	325
CAPQ(ours)	3.1	11.9	137	2.7	12.4	150	1.7	9.2	182	1.9	10.7	142

Table 3: Performance comparison on $split_{Text}$ and $split_{AnnoF}$

gain of 64.2% over the SO baseline (without adaptation). Another interesting observation is: we also report the performance of CAPQ-preserve (no target videos) in the second row, which applies the concept preservation loss only on the source videos and texts (without pseudo-text selection). In this setting, no target video is required during training and we observe that even in this restricted scenario, the method still outperforms the baseline (SO) by a small margin. It indicates that it is useful to make the feature embedding generic, but not sufficient for the cross-domain video retrieval task. We also report the number of parameters in the last column in Table. 4: CAPQ introduces $14M$ more parameters during training (due to the hallucinator). At test-time, since the hallucinator is discarded, there is no computational overhead—inference speed is identical to that of the baseline model (SO).

Qualitative Examples: We also evaluate CAPQ by providing qualitative results and visual examples from $split_{Text}$ in Fig. 4. The first three rows show successful applications of CAPQ and the last row shows a case where it is less effective. Four target domain videos and their corresponding ground truth text descriptions (MovieClips domain) are displayed in the first column for visualization. We observe that CAPQ outperforms the Source Only model by a large margin for each of the first three rows. We also provide the selected pseudo-text in the fourth column using our method. The selected pseudo-text, while not a precise

Method	t2v			v2t			Para-Num
	R1	R10	MR	R1	R10	MR	
SO	1.5	10.3	162	1.6	9.3	190	23.6M
Preserve (no T)	2.2	12.1	168	2.4	10.7	187	37.1M
CAPQ-preserve	2.5	11.7	164	2.5	9.5	185	37.1M
CAPQ-pseudo	2.6	11.3	158	2.4	10.9	163	23.6M
CAPQ	3.1	11.9	137	2.7	12.4	150	37.1M

Table 4: Ablation study on network components ($split_{Text}$)

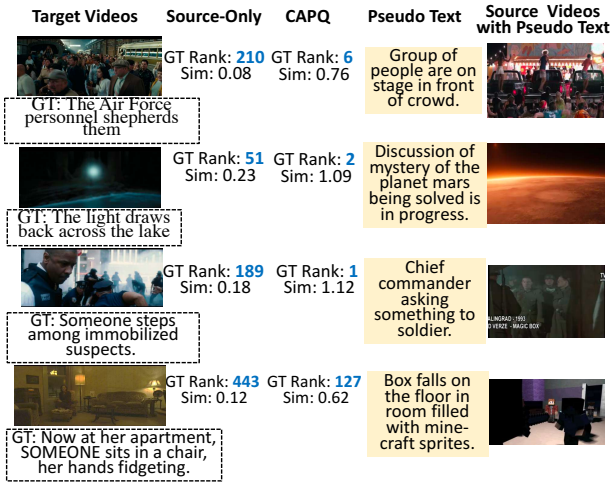

Figure 4: Qualitative results analysis on $split_{Text}$.

match, captures some of the semantic concepts relevant to the target video. In addition, we also show the source domain video frames linked with the selected pseudo text in the last column—the selected source videos share significant commonalities with the target videos. In the last row of Fig. 4, we show an instance for which CAPQ is less effective. In this setting, there is insufficient coverage of the concepts present in the target video for the pseudo-text approach to have high utility. Nevertheless, the selected pseudo-text does still provide some benefit in this setting (in particular the similar scene (room/apartment) and is still identified correctly by CAPQ, providing some boost to performance over the SO model).

Conclusions

In this work, we have proposed a new benchmark and investigated the task of unsupervised domain adaptation for text-video retrieval in this setting. We have introduced the CAPQ framework, and shown that it outperforms standard domain adaptation techniques.

Acknowledgements

The authors gratefully acknowledge the support of Beijing Advanced Discipline Construction Project BMU2019GJJXK001, PKU-Baidu Fund (2019BD017), WICT grant 6202000014/017, the EPSRC Programme Grant CALOPUS EP/R013853/1 and EPSRC Programme Grant Visual AI EP/T028572/1. The authors would also like to thank Andrew Zisserman for helpful suggestions.

References

Arandjelovic, R.; Gronat, P.; Torii, A.; Pajdla, T.; and Sivic, J. 2016. NetVLAD: CNN architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5297–5307.

Ben-David, S.; Blitzer, J.; Crammer, K.; Kulesza, A.; Pereira, F.; and Vaughan, J. W. 2010. A theory of learning from different domains. *Machine learning* 79(1): 151–175.

Bhushan Damodaran, B.; Kellenberger, B.; Flamary, R.; Tuia, D.; and Courty, N. 2018. DeepJDOT: Deep Joint Distribution Optimal Transport for Unsupervised Domain Adaptation. In *The European Conference on Computer Vision (ECCV)*.

Bousmalis, K.; Silberman, N.; Dohan, D.; Erhan, D.; and Krishnan, D. 2016. Unsupervised pixel-level domain adaptation with generative adversarial networks. *arXiv preprint arXiv:1612.05424* .

Cao, Z.; Long, M.; Huang, C.; and Wang, J. 2018. Transfer Adversarial Hashing for Hamming Space Retrieval. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Chen, C.; Xie, W.; Huang, W.; Rong, Y.; Ding, X.; Huang, Y.; Xu, T.; and Huang, J. 2019a. Progressive feature alignment for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 627–636.

Chen, D. L.; and Dolan, W. B. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, 190–200. Association for Computational Linguistics.

Chen, M.-H.; Kira, Z.; AlRegib, G.; Yoo, J.; Chen, R.; and Zheng, J. 2019b. Temporal attentive alignment for large-scale video domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, 6321–6330.

Chen, Q.; and Liu, Y. 2020. Structure-Aware Feature Fusion for Unsupervised Domain Adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 10567–10574.

Chen, Q.; Liu, Y.; Wang, Z.; Wassell, I.; and Chetty, K. 2018. Re-Weighted Adversarial Adaptation Network for Unsupervised Domain Adaptation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 3, 6.

Courty, N.; Flamary, R.; Tuia, D.; and Rakotomamonjy, A. 2017. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence* 39(9): 1853–1865.

Dong, J.; Li, X.; and Snoek, C. G. 2016. Word2visualvec: Image and video to sentence matching by visual feature prediction. *arXiv preprint arXiv:1604.06838* .

French, G.; Mackiewicz, M.; and Fisher, M. 2017. Self-ensembling for visual domain adaptation. *arXiv preprint arXiv:1706.05208* .

Gabeur, V.; Sun, C.; Alahari, K.; and Schmid, C. 2020. Multi-modal transformer for video retrieval. In *European Conference on Computer Vision (ECCV)*, volume 5. Springer.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. 2016. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research* 17(1): 2096–2030.

Gretton, A.; Borgwardt, K. M.; Rasch, M. J.; Schölkopf, B.; and Smola, A. 2012. A kernel two-sample test. *Journal of Machine Learning Research* 13(Mar): 723–773.

Haeusser, P.; Frerix, T.; Mordvintsev, A.; and Cremers, D. 2017. Associative domain adaptation. In *International Conference on Computer Vision (ICCV)*, volume 2, 6.

Hsu, H.-K.; Yao, C.-H.; Tsai, Y.-H.; Hung, W.-C.; Tseng, H.-Y.; Singh, M.; and Yang, M.-H. 2020. Progressive domain adaptation for object detection. In *The IEEE Winter Conference on Applications of Computer Vision*, 749–757.

Jamal, A.; Namboodiri, V. P.; Deodhare, D.; and Venkatesh, K. 2018. Deep Domain Adaptation in Action Space. In *BMVC*, 264.

Karpathy, A.; and Fei-Fei, L. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3128–3137.

Krishna, R.; Hata, K.; Ren, F.; Fei-Fei, L.; and Carlos Niebles, J. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, 706–715.

Liu, F.; and Ye, R. 2019. A strong and robust baseline for text-image matching. *arXiv preprint arXiv:1906.01205* .

Liu, M.-Y.; Breuel, T.; and Kautz, J. 2017. Unsupervised Image-to-Image Translation Networks. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30*, 700–708. Curran Associates, Inc. URL http://papers.nips.cc/paper/6672-unsupervised-image-to-image-translation-networks.pdf.

Liu, M.-Y.; and Tuzel, O. 2016. Coupled generative adversarial networks. In *Advances in neural information processing systems*, 469–477.

Liu, Y.; Albanie, S.; Nagrani, A.; and Zisserman, A. 2019. Use What You Have: Video Retrieval Using Representations From Collaborative Experts. *arXiv preprint arXiv:1907.13487* .

Long, M.; Cao, Y.; Wang, J.; and Jordan, M. I. 2015. Learning transferable features with deep adaptation networks. *arXiv preprint arXiv:1502.02791* 97–105.

Long, M.; CAO, Z.; Wang, J.; and Jordan, M. I. 2018. Conditional Adversarial Domain Adaptation. In Bengio, S.; Wallach, H.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 31*, 1640–1650. Curran Associates, Inc. URL http://papers.nips.cc/paper/7436-conditional-adversarial-domain-adaptation.pdf.

Long, M.; Wang, J.; and Jordan, M. I. 2016. Deep transfer learning with joint adaptation networks. *arXiv preprint arXiv:1605.06636* .

Long, M.; Zhu, H.; Wang, J.; and Jordan, M. I. 2016. Unsupervised domain adaptation with residual transfer networks. In *Advances in Neural Information Processing Systems*, 136–144.

Miech, A.; Alayrac, J.-B.; Smaira, L.; Laptev, I.; Sivic, J.; and Zisserman, A. 2020. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9879–9889.

Miech, A.; Laptev, I.; and Sivic, J. 2017. Learnable pooling with context gating for video classification. *arXiv preprint arXiv:1706.06905* .

Miech, A.; Laptev, I.; and Sivic, J. 2018. Learning a text-video embedding from incomplete and heterogeneous data. *arXiv preprint arXiv:1804.02516* .

Miech, A.; Zhukov, D.; Alayrac, J.-B.; Tapaswi, M.; Laptev, I.; and Sivic, J. 2019. Howto100M: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE International Conference on Computer Vision*, 2630–2640.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.

Mithun, N. C.; Li, J.; Metze, F.; and Roy-Chowdhury, A. K. 2018. Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, 19–27. ACM.

Munro, J.; and Damen, D. 2020. Multi-Modal Domain Adaptation for Fine-Grained Action Recognition. *arXiv preprint arXiv:2001.09691* .

Otani, M.; Nakashima, Y.; Rahtu, E.; Heikkilä, J.; and Yokoya, N. 2016. Learning joint representations of videos and sentences with web image search. In *European Conference on Computer Vision*, 651–667. Springer.

Radovanovic, M.; Nanopoulos, A.; and Ivanovic, M. 2010. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research* 11(sept): 2487–2531.

Rohrbach, A.; Rohrbach, M.; Tandon, N.; and Schiele, B. 2015. A dataset for movie description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3202–3212.

Rohrbach, A.; Torabi, A.; Rohrbach, M.; Tandon, N.; Pal, C.; Larochelle, H.; Courville, A.; and Schiele, B. 2017a. Movie Description. *International Journal of Computer Vision* URL http://link.springer.com/article/10.1007/s11263-016-0987-1?wt_mc=Internal.Event.1.SEM.ArticleAuthorOnlineFirst.

Rohrbach, A.; Torabi, A.; Rohrbach, M.; Tandon, N.; Pal, C.; Larochelle, H.; Courville, A.; and Schiele, B. 2017b. Movie description. *International Journal of Computer Vision* 123(1): 94–120.

Russo, P.; Carlucci, F. M.; Tommasi, T.; and Caputo, B. 2017. From source to target and back: symmetric bi-directional adaptive gan. *arXiv preprint arXiv:1705.08824* 3.

Sun, B.; and Saenko, K. 2016. Deep coral: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision*, 443–450. Springer.

Tzeng, E.; Hoffman, J.; Saenko, K.; and Darrell, T. 2017. Adversarial discriminative domain adaptation. *arXiv preprint arXiv:1702.05464* .

Tzeng, E.; Hoffman, J.; Zhang, N.; Saenko, K.; and Darrell, T. 2014. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474* .

Venkateswara, H.; Eusebio, J.; Chakraborty, S.; and Panchanathan, S. 2017. Deep Hashing Network for Unsupervised Domain Adaptation. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Wray, M.; Larlus, D.; Csurka, G.; and Damen, D. 2019. Fine-grained action retrieval through multiple parts-of-speech embeddings. In *Proceedings of the IEEE International Conference on Computer Vision*, 450–459.

Wu, Y.; Winston, E.; Kaushik, D.; and Lipton, Z. 2019. Domain Adaptation with Asymmetrically-Relaxed Distribution Alignment. In Chaudhuri, K.; and Salakhutdinov, R., eds., *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 6872–6881. Long Beach, California, USA: PMLR.

Xu, J.; Mei, T.; Yao, T.; and Rui, Y. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5288–5296.

Yan, H.; Ding, Y.; Li, P.; Wang, Q.; Xu, Y.; and Zuo, W. 2017. Mind the Class Weight Bias: Weighted Maximum Mean Discrepancy for Unsupervised Domain Adaptation. *arXiv preprint arXiv:1705.00609* .

Zhang, W.; Ouyang, W.; Li, W.; and Xu, D. 2018. Collaborative and adversarial network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3801–3809.