Joint Wasserstein Autoencoders for Aligning Multimodal Embeddings

Shweta Mahajan Teresa Botschen Iryna Gurevych Stefan Roth
Department of Computer Science, TU Darmstadt

Abstract

One of the key challenges in learning joint embeddings of multiple modalities, e.g. of images and text, is to ensure coherent cross-modal semantics that generalize across datasets. We propose to address this through joint Gaussian regularization of the latent representations. Building on Wasserstein autoencoders (WAEs) to encode the input in each domain, we enforce the latent embeddings to be similar to a Gaussian prior that is shared across the two domains, ensuring compatible continuity of the encoded semantic representations of images and texts. Semantic alignment is achieved through supervision from matching image-text pairs. To show the benefits of our semi-supervised representation, we apply it to cross-modal retrieval and phrase localization. We not only achieve state-of-the-art accuracy, but significantly better generalization across datasets, owing to the semantic continuity of the latent space.

1. Introduction

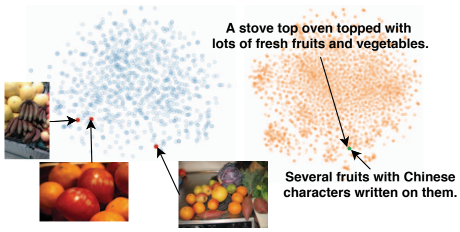
The availability of significant amounts of image-text data on the internet (*e.g*., images with their captions) has posed the question whether it is possible to leverage information from both visual *and* textual sources. To take advantage of such heterogeneous data, one of the fundamental challenges is the joint representation of multiple domains [3]. Powerful multimodal representations are integral to the accuracy of models on cross-domain tasks, such as image captioning [21] or cross-domain retrieval [1, 7, 9, 24, 26, 43].

Multimodal embeddings of images and texts can be obtained by mapping input image and text representations into a common latent space [9, 43]. Learning such representations is often formulated as an image-text matching problem in a fully supervised setup. An alternative approach is to first learn separate latent spaces and to align them later through constraints, *e.g*., supervised information [40]. The benefit is that the latent representations of each modality are learned independently, allowing to take advantage of unsupervised (*i.e*. unpaired) data.

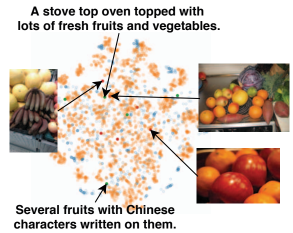
One of the main challenges in multimodal learning is to obtain meaningful latent representations such that they capture semantics that are present across modalities, even if the paired training data does not extensively cover all relevant semantic concepts. For example, the Flickr30k [48] and COCO [27] image captioning datasets do not contain matching image pairs. Lacking within-domain structural constraints, semantic similarity within each modality may not be preserved in the embedding space (Figs. 1b and 2a).

We address this problem by learning *semantically continuous* latent representations of images and texts in their respective embedding spaces, *i.e*. multimodal embeddings that encourage a smooth change in the semantics of the input modalities. We adopt a semi-supervised setting and propose a *joint Wasserstein autoencoder* (jWAE) model, leveraging that regularized autoencoders are known to yield semantically meaningful latent spaces [22, 39]. Specifically, we adopt Gaussian regularization to ensure semantically continuous latent representations of each input modality (*c.f*. Fig. 2b). In contrast to standard Wasserstein autoencoders (WAEs) [39], we share the Gaussian prior across modalities to encourage comparable levels of semantic continuity in both modalities. Unlike variational autoencoders (VAEs) [22], Wasserstein autoencoders map the input data to a point in the latent space, which allows for the coordination of the two modalities through a supervised loss based on matching image-text pairs. The advantage of the shared Gaussian prior on the two modalities is that their semantic representations are comparable and can be better aligned with supervision as illustrated in Fig. 1c.

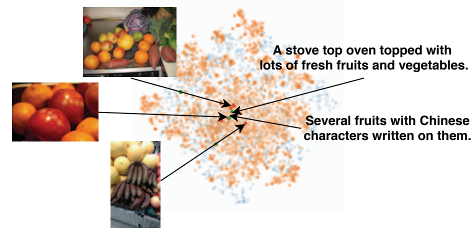
We first evaluate our multimodal embeddings of images and texts on cross-modal retrieval and show that they yield state-of-the-art accuracy on the Flickr30k and COCO datasets. One of the crucial advantages of the semantically continuous representation from our semi-supervised approach is its generalization capability across datasets. The benefit over the state of the art widens when embeddings of one dataset are evaluated on a related, *previously unseen* dataset. Finally, we demonstrate the advantage of our jWAE on phrase localization on the Flickr30k Entities dataset [34], where we again outperform recent methods from the literature.



(a) Input embeddings for images (left) and text (right).



(b) Joint representation of image-text embeddings based on negative sampling [43].



(c) Coordinated representation of image-text embeddings of the proposed jWAE.

Figure 1. *From input embeddings to coordinated embeddings with semantic continuity: (a)* Semantic similarity in the input modalities does not imply proximity in the embedding space. *(b)* Joint representations align the modalities, but the sparsity of supervised information does not achieve continuity of the semantic space. *(c)* Our jWAE based on joint Gaussian regularization leads to semantic continuity.

(a) Matching images in the joint embedding space of [43].

(b) Matching images in the coordinated embedding space with the proposed jWAE.

Figure 2. *Semantic continuity with Gaussian regularization.* Each row shows images that are close in the learned embedding space. *(a)* Without prior knowledge of matching image pairs, supervised cross-modal methods do not achieve semantic continuity. *(b)* Our jWAE based on Gaussian regularization achieves semantic continuity in the coordinated image space in absence of within-domain supervision.

2. Related Work

Supervised multimodal learning. Early work on multimodal learning includes Canonical Correlation Analysis (CCA) [17] and kernel CCA (KCCA) [25], which maximize correlation to learn projections of joint embeddings (*e.g.*, of images and texts). However, these methods do not scale to large datasets [28]. Deep Canonical Correlation Analysis (DCCA) [1] aims to overcome this scalability issue. Yet, optimization is challenging as the covariance matrix has to be estimated during training and is prone to over-fitting.

Many recent works formulate embedding multiple domains into a joint space as a learning-to-rank problem [9, 10, 14, 15, 23, 43, 46]. A ranking hinge loss with a margin is used such that matching cross-domain pairs are ranked higher (*i.e.* are closer in the latent space) than non-matching pairs. Wang *et al.* [43] additionally incorporate structural information on the input representations themselves with a domain-specific ranking loss. This requires prior knowledge about within-domain matching pairs, which was only available for text. Image-text matching has also been studied in a classification setting, employing logistic regression or a softmax with cross-entropy [11, 43].

Gu *et al.* [14] augment the ranking loss with a conditional generative model framework for cross-modal generation to obtain fine-grained multimodal features. Harada *et al.* [16] learn an image-text embedding space with a Gaussian prior using generative adversarial networks. However, the Gaussian latent prior is applied only on the image modality, and the text distribution is matched to the latent image distribution. Moreover, the adversarial framework can suffer from mode collapse. Chi *et al.* [6] match image and text embeddings to the label representation. They assume that image and text have same label, which is limited to tasks where only one concept is required per image.

Wehrmann *et al.* [45] improve sentence representations with a character level inception module and [20, 26] improve image representations for image-text matching models. Huang *et al.* [20] use multi-label classification to extract various concepts in images, requiring additional image annotations. Lee *et al.* [26] propose an attention mechanism for aligning image regions with words in a sentence. This is orthogonal to the underlying multimodal embedding and can be combined with the proposed jWAE framework.

Semi- and unsupervised multimodal learning. Various approaches [31, 40] use autoencoders to obtain latent representations. Embeddings of the two domains can be aligned using distribution matching constraints [40, 42]. Unlike previous work, which does not encourage any continuity in the latent semantic space across modalities, we use *regularized* autoencoders based on generative models, which enforce the latent embeddings from the encoders to match a prior distribution, thereby yielding continuity in the embed-

Text	Supervised approach [43]	Ours (jWAE-MSE)
Two tan dogs play on the grass near the wall.		
A tennis player wearing white is jumping up to hit a ball.		

Table 1. Examples of images retrieved for a given sentence on Flickr30k based on embeddings trained on COCO.

Image	Supervised approach [43]	Ours (jWAE-MSE)
	• Chevrolet car on display at a convention. • Construction in the city at night. • Two firemen beside their fire engine. • An escalator with many people on it, leading out of a tunnel.	• Two firemen beside their fire engine. • BSR trucks and machinery and workers. • An old, beat-up jeep being towed away. • Construction in the city at night.
	• The dog is running around the cow. • A person laying on the ground next to a cow. • Two children pet horses in a field. • Girl atop horse that is chasing a small longhorn.	• A lioness is chasing a black bison across a grassy plain. • A lioness chases a black animal with horns. • Two brown dogs play. • Horse jockeys racing on horses in a race.

Table 2. Examples of the top-4 captions retrieved for a given image on Flickr30k based on embeddings trained on COCO.

ding space. We use the latent representations obtained from these models as a basis for our multimodal approach.

Deep generative models aim to minimize the difference between the model and the empirical data distribution, and have been successfully applied, *e.g.*, to image generation tasks. Generative adversarial networks (GANs) [13] generate the model distribution in a one step procedure where decoders are input with random (Gaussian) noise to construct the data distribution. Regularized autoencoders such as variational autoencoders (VAEs) [22] model the data distribution through a two step procedure. The empirical distribution is first mapped to a latent space via encoders and then mapped back to the data space via decoders. VAEs minimize the reconstruction error between the input and output representation and balance this with the discrepancy between the encoded representation in the latent space and a prior distribution, *e.g.* a Gaussian, for *each* input. Recently, [29, 39] proposed autoencoder-based frameworks where the discrepancy between the encoded distribution of *all* input representations and prior distribution is minimized. This forces the entire encoded distribution to match the prior. Such regularization captures the semantics of entire input distribution in a continuous latent representation, desirable for encoding each modality in multimodal learning.

3. Motivation & Background

Current approaches formulate the task of learning multimodal representations of images and text in an image-text matching framework [9, 44], in which a ranking loss is minimized in a fully supervised setting. For matching image-text pairs (x_l, y_l) and non-matching images $x_{l'}$ or texts $y_{l'}$ with similarity function $s(x, y)$, the ranking formulation based on the max-margin hinge loss with margin m is defined as

$$\begin{aligned}\mathcal{L}_{\text{MH}} = \sum_l \psi\Big(& \max\left[0, m + s(x_l, y_{l'}) - s(x_l, y_l)\right]\Big) \\ & + \psi\Big(\max\left[0, m + s(x_{l'}, y_l) - s(x_l, y_l)\right]\Big).\end{aligned} \tag{1}$$

Here, ψ is either the sum of hinge losses over all the negative samples for a given matching pair, or the maximum over all hinge losses. The dependence on the choice of negative examples limits the robustness and generalization of the obtained multi-modal embeddings; they fit to the particularities of a dataset, which can be seen from the example retrievals in Tables 1 and 2 on the Flickr30k dataset for an embedding space trained on the COCO dataset. While methods like domain adaptation rely on data from source *and* target datasets to adapt a model to perform better on a specific target dataset [19, 41], in this work we show the performance of the embeddings where both supervised and unsupervised losses are trained *only* on the source dataset, commonly referred to as *generalization*.

To overcome the limitation of existing methods in expressing semantic coherence within a domain and across multiple domains, we propose to employ Gaussian regularization on the latent distribution. By virtue of the autoencoder framework, structurally similar input representations are close to each other in the low-dimensional latent space; Gaussian regularization encourages *continuity* in the space of encoded representations. Semantic alignment of these spaces is further obtained with supervision. We refer to the resultant embeddings as *semantically continuous*. Our model consists of three main components. First, each input distribution is mapped to a Gaussian distribution where semantically similar representations are close to each other within the domain. This is illustrated in Fig. 2b, where images that are close in the embedding space are semantically related. Second, we share this Gaussian prior across domains, leading to compatible levels of continuity in both domains. Third, the latent representations of images and text are semantically aligned with supervised information. As shown in Fig. 1c, images and texts representing similar semantics come closer in the proposed joint embedding space. Moreover, this offers significantly better generalization across datasets as seen in Tables 1 and 2, where the example captions retrieved with our semi-supervised approach are semantically more related to the given image than when only employing supervised ranking.

4. Approach

We propose a semi-supervised approach for improving the semantic alignment between two modalities, such as images and texts. Let $X = \{x_i\}_{i=1}^{N_X}$ and $Y = \{y_j\}_{j=1}^{N_Y}$ denote unpaired input images and texts, respectively. Further, let $S = \{(x_l, y_l)\}_{l=1}^{N_S}$ be matching image and text pairs. We assume the latent space of each domain to be of dimension d. Encoders $f_v : \mathbb{R}^{d_v} \to \mathbb{R}^d$ and $f_t : \mathbb{R}^{d_t} \to \mathbb{R}^d$ map visual data (images) and text to their respective d-dimensional latent spaces. $g_v : \mathbb{R}^d \to \mathbb{R}^{d_v}$ and $g_t : \mathbb{R}^d \to \mathbb{R}^{d_t}$ are the decoders that map the latent representations back to images and text. We denote the latent representations of images and texts by $\tilde{v}$ and $\tilde{t}$, respectively. Next, we let $P_L \sim \mathcal{N}(0, I_d)$ be a unit Gaussian prior in the d-dimensional space with identity covariance matrix I_d and denote the encoded image and text distributions as $F_v = \{f_v(x_i)\}_{i=1}^{N_X}$ and $F_t = \{f_t(y_j)\}_{j=1}^{N_Y}$, respectively. Similarly, we define the output (reconstructed) model distributions as $G_v = \{g_v(\tilde{v}_i)\}_{i=1}^{N_X}$ and $G_t = \{g_t(\tilde{t}_j)\}_{j=1}^{N_Y}$. Following the common abuse of notation, we let F_v and F_t denote both the encoded activations and the distribution over encodings.

4.1. Wasserstein autoencoder backbone

We build the latent representation of each domain, images or text, on Wasserstein autoencoders [39]. We first describe the WAE backbone for the image pipeline and then extend it to text.

Generative models such as VAEs and WAEs minimize the discrepancy between the true data distribution X and the model (reconstructed) distribution G_v. In such models, latent variables z, sampled from a fixed prior distribution P_L in the latent space, are mapped to the space of the original data with parameterized functions g_v, and f-divergences, such as the KL-divergence or the Jensen-Shannon divergence, between the distributions are minimized. In high-dimensional spaces, estimating the model distribution directly by sampling in the latent domain would require a large number of samples from the latent distribution. Therefore, representing the model distribution by random sampling of the latent distribution is computationally expensive.

Variational autoencoders make sampling efficient by introducing a proposal distribution for each x_i. Specifically, $F_v(x_i)$ is a latent distribution that generates latent representations z likely to produce x_i. The VAE minimizes

$$\mathcal{L}_{\text{VAE}} = \mathcal{D}_{\text{KL}}\big(F_v(x_i)||P_L\big) - \mathbb{E}_{F_v(x_i)}\big[\log G_v(\tilde{v}_i)\big], \quad (2)$$

where the first term encourages the latent variables z over each x_i to match the prior distribution P_L. While this helps latent samples to be representative of a data point, it does not capture the full underlying true data distribution. As has been pointed out by [39] for a Gaussian prior, this results in overlapping Gaussians in the latent space from different input data points. This also is the cause of blurry images from VAEs in image generation tasks. Moreover, mapping the input to a latent *distribution* is problematic when using supervision in the latent space, as intended here.

To model the entire data distribution in the latent space, minimizing $\mathcal{D}(F_v||P_L)$ with $F_v = \int F_v(x)\,dX$ is thus desirable [39]. To that end, given the input distribution X and the model distribution G_v, Wasserstein autoencoders (WAEs) minimize the optimal transport cost $W_c(X, G_v)$ between the two distributions. Optimal transport with a cost function $c(s,t) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ is defined as

$$W_c(X, G_v) = \inf_{\gamma \in \Gamma(X, G_v)} \mathbb{E}_{(s,t)\sim\gamma}\big[c(s,t)\big], \quad (3)$$

where $\Gamma(X, G_v)$ is the set of all possible joint distributions (couplings) of (s,t) whose marginals are X and G_v, respectively. In generative models with a deterministic mapping from the latent distribution P_L to G_v, the optimal transport cost between X and G_v reduces to finding a conditional distribution F_v such that $\int F_v(x)dX = P_L$ [4, 39]. This constraint is enforced as a regularization term by minimizing $\mathcal{D}(F_v||P_L)$. This yields Wasserstein autoencoders as

$$\mathcal{L}_{\text{WAE}} = \inf_{F_v \in \mathcal{F}} \mathbb{E}_X \mathbb{E}_{F_v}\big[c(x, g_v(x))\big] + \lambda \mathcal{D}\big(F_v||P_L\big). \quad (4)$$

Here, $c(s,t) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ is a cost function and $\mathcal{F}$ is a set of probabilistic encoders without any constraints (non-parametric), which model $p(z|x)$. We take it to be the set of all fully-connected networks with fixed size.

Choosing $\mathcal{D}$ as the KL-divergence between F_v and P_L would require closed forms of F_v and P_L. Since the closed form of F_v is not available, we instead minimize the Jensen-Shannon divergence between the prior and encoded latent distributions $\mathcal{D}_{\text{JS}}(F_v||P_L)$ using a GAN-based formulation, which allows to conveniently use samples from the distribution. We refer to $\mathcal{D}_{\text{JS}}$ as Gaussian regularization since we consider the prior to be a unit Gaussian.

Note that the popular Wasserstein GANs [2] minimize the optimal transport cost (Eq. 3) between the data distribution and the model distribution using the dual formulation of the optimal transport cost directly from the latent space

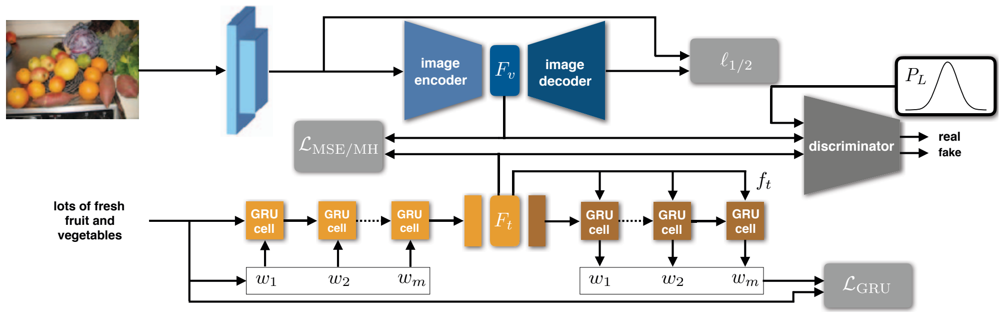


Figure 3. Architecture of our joint Wasserstein Autoencoder.

to the data space. That is, they do not have an encoder that maps the input to latent representations. However, for learning coordinated representations of two input data distributions with a shared prior, an encoder that maps data points to the latent space is desirable. We thus build on WAEs here.

4.2. Joint WAE

For learning coordinated representations, we are now interested in formulating continuous d-dimensional embedding spaces for each modality, which are aligned through constraints. To this end, we propose to share the prior on the latent representations. Specifically, the latent representations of each domain are constrained to be close to the same unit Gaussian distribution; that is we minimize the discrepancy $\mathcal{D}\left(F_v||P_L\right)$ and $\mathcal{D}\left(F_t||P_L\right)$ between the encoded representations of each modality and the Gaussian prior P_L.

We now formulate our joint Wasserstein autoencoder in terms of a classical encoder-decoder setting. To that end, we first note that when f_v and g_v in a regular WAE (Eq. 4) are parameterized with encoders and decoders in a deep neural network framework, the first term of Eq. (4) reduces to minimizing the reconstruction error between the true input representation and the decoded representation.

To apply the WAE formulation to sentences, we first extend it to gated recurrent units (GRUs) [37] where f_t is the encoded output of the GRU encoder and Gaussian regularization is applied to the encoded distribution, F_t, of all the sentences. f_t serves as the input hidden state to the GRU decoder and the reconstruction loss for the sentence is

$$\mathcal{L}_{\text{GRU}} = -\sum_{m=0}^{M-1} \log p_{g_t}\left(w_m^j \middle| w_{0:m-1}^j, f_t(y_j); g_t\right), \quad (5)$$

where w_m^j is the ground truth word and $p_{g_t}(w_m^j|w_{0:m-1}^j, f_t)$ is the output probability of word w_m^j in sentence y_j given the decoder g_t and hidden state f_t.

Recalling that F_v and F_t denote the encoded distribution of images and text, respectively, we formulate the unsupervised part of our joint WAE loss as

$$\begin{aligned}\mathcal{L}_{\text{jWAE}} = \lambda_1 \sum_{i=1}^{N_X} \left\| x_i - g_v\big(f_v(x_i)\big)\right\| \\ - \lambda_2 \sum_{j=1}^{N_Y}\sum_{m=0}^{M-1} \log p_{g_t}\left(w_m^j \middle| w_{0:m-1}^j, f_t(y_j); g_t\right) \\ + \lambda_3 \mathcal{D}_{\text{JS}}(F_v||P_L) + \lambda_4 \mathcal{D}_{\text{JS}}(F_t||P_L). \quad (6)\end{aligned}$$

Here, $\{\lambda_1, \ldots, \lambda_4\}$ are the regularization parameters and $\|.\|$ is ℓ_1 or ℓ_2 norm. Modality specific reconstruction error terms are crucial components in preventing mode collapse and encouraging diversity in the latent representations of each domain. For sentences encoded with fully connected encoders using pre-trained sentence encodings like average of word2vec [30], we use an analogous formulation with the mean-squared error as the reconstruction loss.

Gaussian regularization itself does not induce any semantic coupling between the cross-modal distributions. To ensure that the latent spaces not only have a compatible continuity but rather align semantic concepts across modalities, we add a supervised loss minimizing the distance between latent representations of matching image-text pairs. Embeddings of the two domains can now be directly aligned with the mean-squared error

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N_S}\sum_{l=1}^{N_S} \left\| f_v(x_l) - f_t(y_l)\right\|^2. \quad (7)$$

The overall loss function of our approach is then given as

$$\mathcal{L}_{\text{jWAE-MSE}} = \mathcal{L}_{\text{jWAE}} + \mathcal{L}_{\text{MSE}}. \quad (8)$$

Alternatively, we also show the effect of Gaussian regularization with the max-margin hinge loss [9] from Eq. (1) as supervised loss function

$$\mathcal{L}_{\text{jWAE-MH}} = \mathcal{L}_{\text{jWAE}} + \mathcal{L}_{\text{MH}}. \quad (9)$$

The overall model architecture is illustrated in Fig. 3.

Implementation. The outputs from the encoders (*i.e.* the encoded distributions F_v and F_t) along with $z \sim P_L$ are input to the discriminator. The number of samples from the prior distribution P_L equals the sum of the samples output by the two encoders. The discriminator distinguishes between the encoded distributions and the joint Gaussian prior. Note that we implement a single discriminator network for two generator networks, which makes our architecture computationally efficient. This is possible as the same prior distribution is used for images and texts. The discriminator is a fully-connected three layer neural network with leaky ReLU non-linearities after the first two layers, which enables a better flow of gradients during optimization [36]. The generator network (the encoder) of each pipeline tries to "fool" the discriminator network by generating encodings close to the Gaussian prior distribution.

In general, both encoders and decoders consist of two fully connected layers; ReLU non-linearities are applied after the first layers. For the text pipeline based on GRUs, the encoder is a bi-directional GRU with two layers. The output of the GRU is encoded in the latent space after application of a linear fully-connected layer. The decoder is a uni-directional GRU with word dropout encouraging meaningful latent representations of sentences.

5. Experiments

To show the applicability of our jWAE framework across *different tasks*, we evaluate the learned multimodal embeddings on cross-modal retrieval and phrase localization.

5.1. Cross-modal retrieval

Visual input. We consider pretrained VGG-19 [38] and ResNet-152 [18] models for image features. For VGG-19, we extract the 4096-dimensional feature vector from the first fully connected layer and for ResNet-152 the 2048-dimensional activations from the fully connected layer.

Textual input. Following [43], we use the mean of 300-dimensional word2vec [30] features of the words in the sentence. Alternatively, we use nonlinear Fisher vectors from a hybrid Gaussian-Laplacian mixture model (HGLMM) [24]. For GRU, one-hot encodings of the words are projected with an embedding layer, which is initialized either randomly or with pre-trained word2vec embeddings.

Datasets. Two popular benchmark datasets for evaluating multimodal visual and textual representations are Flickr30k and COCO [9, 24, 43]. Flickr30k [48] is comprised of 31783 images with five captions per image. We use the splits of [21, 44]; validation and test splits consist of 1000 images with 5 captions each. The remaining images are used for training. COCO [27] is larger and more diverse than Flickr30k. It consists of 82783 training images with five captions each. 5000 images from the validation set are retained for validation purposes and the remaining 30504 images are used for training. Similar to [21, 24, 44], we use 1000 images with their captions in the test split.

Network training. We train the network (see supplemental material for architectural details) using the Adam optimizer with a learning rate of 1E-4. The discriminator is trained with a learning rate of 5E-5. The batch size is taken as 64 or 128. For jWAE-MSE, the regularization parameters are set to $\lambda_1 = \lambda_2 = 1.0$ for the reconstruction terms in Eq. (6) and $\lambda_3 = \lambda_4 = 0.2$ for the Gaussian regularization. For jWAE-MH, the parameters are $\lambda_1 = 0.5$, $\lambda_2 = 0.005$, $\lambda_3 = \lambda_4 = 0.01$ for most of the experiments.

Evaluation metric. In cross-modal retrieval tasks, Recall@K is a standard performance measure and defined as the fraction of instances for which their ground truth is in the top-K based on a similarity score (cosine similarity).

Baselines & methods. We compare the accuracy of our approach against several state-of-the-art methods for image-to-text and text-to-image retrieval, particularly the Embedding Network of [43, 44] and VSE++ [9], which use different formulations of the ranking loss, and the attention-based Stacked Cross Attention Network (SCAN) [26]. For evaluation with SCAN, we integrate the jWAE-MH framework with the best-performing setting on the respective dataset. To compare our approach against methods without negative sampling, we also include the Similarity Network of [43] and the Canonical Correlation Analysis (CCA) approach of [24]. For our jWAE framework, we demonstrate the effect of Gaussian regularization with supervision through the mean-squared error (jWAE-MSE, Eq. 8) as well as a margin-based hinge loss (jWAE-MH, Eq. 9). We extend jWAE for the SCAN method (jWAE-MH+SCAN t-i/i-t) where in i-t attention is applied on words with respect to each image region and in t-i image regions are attended with respect to each word in the sentence. We also include results from a MMD loss for matching text and image distributions (MMD-MSE) [40] and an ablation of our method without the Gaussian regularization (MSE), both with the usual reconstruction error for autoencoders.

Results. In Table 3, we show the results of our method for cross-modal retrieval on the Flickr30k and COCO datasets.

Our jWAE framework leads to competitive results compared to the current state of the art in image-to-text retrieval. For example, jWAE-MH outperforms VSE++ with respect to top-1 recall by 1.0% points on Flickr30k and by 2.0% points on COCO. For the Embedding Network with VGG+w2v features, jWAE-MH has 3.5% better accuracy on COCO. Our semi-supervised representations also improve the top-1 recall of SCAN t-i by 2.2% and 2.1% points on Flickr30k and COCO, respectively. For text-to-image retrieval, improving the top-1 recall is more challenging. Yet, we also achieve an improvement in top-1 recall of 1.3% and

	Method (Features)	Recall on Flickr30k						Recall on COCO					
		Image-to-text			Text-to-image			Image-to-text			Text-to-image		
		@1	@5	@10	@1	@5	@10	@1	@5	@10	@1	@5	@10
Without Negative Sampling	CCA (VGG+HGLMM) [24]	34.4	61.0	72.3	24.4	52.1	65.6	37.7	66.6	79.1	24.9	58.8	76.5
	CCA (Mean Vector) [24]	24.8	52.5	64.3	20.5	46.3	59.3	33.2	61.8	75.1	24.2	56.4	72.4
	Sim. Network (VGG+HGLMM) [43]	16.6	38.8	51.0	7.4	23.5	33.3	30.9	61.1	76.2	14.0	30.0	37.8
	MMD-MSE (VGG+w2v) [40]	35.3	60.8	72.8	16.5	40.2	52.9	42.9	73.8	84.4	19.6	50.2	66.6
With Negative Sampling	Emb. Network (VGG+w2v) [43]	35.7	62.9	74.4	25.1	53.9	66.1	40.7	74.2	85.3	33.5	68.7	83.2
	Emb. Network (VGG+HGLMM) [43]	40.3	68.9	79.9	29.7	60.1	72.1	50.1	79.7	89.2	39.6	75.2	86.9
	2WayNet (VGG+HGLMM) [7]	49.8	67.5	–	36.0	55.6	–	55.8	75.2	–	39.7	63.3	–
	VSE++ (Resnet+GRU) [9]	52.9	80.5	87.2	39.6	70.1	79.5	64.6	90.0	95.7	52.0	84.3	92.0
	GXN (ResNet+GRU) [15]	56.8	–	89.6	41.5	–	80.1	68.5	–	97.9	56.6	–	94.5
	SCO (Resnet+GRU) [20]	55.5	82.0	89.3	41.1	60.5	80.1	69.9	92.9	97.5	56.7	**87.5**	**94.8**
	SCAN t-i (ResNet+GRU) [26]	61.8	87.5	93.7	45.8	74.4	83.0	70.9	**94.5**	97.8	56.4	87.0	93.9
	SCAN i-t (ResNet+GRU) [26]	67.9	89.0	94.4	43.9	74.2	82.8	69.2	93.2	97.5	54.4	86.0	93.6
	MSE (VGG+w2v)	33.9	61.3	73.2	15.9	40.3	52.8	45.0	76.5	87.0	21.3	52.3	68.1
	jWAE-MSE (VGG+w2v)	35.7	61.6	73.6	17.3	41.8	55.3	43.2	75.1	85.7	21.5	53.5	69.3
	jWAE-MH (VGG+w2v)	35.4	62.5	74.4	24.1	50.4	62.6	44.2	77.4	87.7	31.3	66.0	81.3
	jWAE-MSE (VGG+HGLMM)	40.3	66.3	77.2	20.3	46.5	58.9	50.3	79.4	88.3	25.2	57.5	73.3
	jWAE-MH (ResNet+GRU)	53.9	82.2	87.4	40.7	72.4	81.9	66.6	91.4	96.6	53.1	84.5	92.0
	jWAE-MH+SCAN t-i (ResNet+GRU)	64.0	89.4	**95.2**	**47.1**	**75.9**	**84.0**	**72.0**	**94.5**	**98.3**	**57.1**	**87.5**	94.1
	jWAE-MH+SCAN i-t (ResNet+GRU)	**68.5**	**89.6**	94.2	44.2	74.2	83.2	69.8	93.3	98.1	54.6	86.1	93.2

Table 3. Cross-modal retrieval results (in %) on the Flickr30k dataset [48] as well as the COCO dataset [27] with 1000 test images.

Method (all ResNet+GRU)	Image-to-text		Text-to-image	
	R@1	R@5	R@1	R@5
VSE++ [9] (baseline)	64.6	90.0	52.0	84.3
VAE-MH	57.8	86.9	46.3	80.3
jWAE-MH	**66.6**	**91.4**	**53.1**	**84.5**

Table 4. Comparison of jWAE with VAE on the COCO dataset.

Flickr30k(Train) ⇒ COCO(Test)

Method	Image-to-text		Text-to-image	
	R@5	R@10	R@5	R@10
CCA (VGG+w2v) [24]	13.3	20.1	10.3	16.0
Embed. Network (VGG+w2v) [43]	37.6	49.5	32.5	45.8
MSE (VGG+w2v)	40.1	52.9	21.7	33.8
SCAN (ResNet+GRU) [26]	59.2	70.0	53.2	66.6
jWAE-MSE (VGG+w2v)	42.8	55.3	28.3	40.9
jWAE-MH (VGG+w2v)	43.4	58.4	34.5	49.2
jWAE-MH+SCAN (ResNet+GRU)	**64.4**	**76.9**	**55.2**	**68.7**

Table 5. Generalization results of models trained only on the Flickr30k training set and evaluated on the COCO test set.

0.7% points over SCAN t-i, and an improvement of 1.1% for VSE++ on Flickr30k and COCO. This shows that irrespective of the network architecture and complexity of input features, our jWAE improves the current state of the art. To the best of our knowledge, our jWAE framework improves (over) the currently leading cross-modal retrieval methods.

We observe that the recall for text-to-image retrieval of jWAE-MSE is not as competitive and is comparable to methods that do not use negative sampling. This can be attributed to the nature of the datasets where five sentences compete for the same image. Moreover, given a sentence there can be many images that can be described reasonably well by the sentence. jWAE-MH bridges the gap between maximizing top-K recall by ranking matching image-text pairs higher than non-matching pairs *and* improving accuracy of the embeddings by encouraging semantic continuity.

In Table 4, we additionally compare jWAE-based Gaussian regularization to traditional VAEs using VSE++ as baseline. While jWAE-MH improves the accuracy of the VSE++ baseline, VAE-MH results in a decrease of top-K recall. The reason is that VAEs map *each* input to a Gaussian latent *distribution*. This hinders the application of a point-wise supervised loss in the latent space. In order to match the latent representations of two modalities with the MH loss, we instead require mapping to a *point* in the latent space. jWAE enforces global Gaussian priors on the latent distributions while mapping the input to a point in the embedding space. Therefore and unlike VAEs, jWAEs are suitable for semi-supervised learning of joint embeddings.

To show that our method learns meaningful representations with continuity in the latent space, we test the cross-dataset generalization capability of our method against various retrieval approaches: CCA [24], the Embedding Network [43], and SCAN [26]. For cross-dataset generalization, the model is trained on the training set of one dataset, *e.g*. COCO (Flickr30k), and tested on the test set of another dataset, *i.e*. Flickr30k (COCO). Note, we do not train to improve the accuracy on specific target dataset. We find that previous methods based on global representations [43] have low generalization performance with top-10 recall as low as 12.2% when testing on Flickr30k for a model trained

COCO(Train) ⇒ Flickr30k(Test)				
	Image-to-text		Text-to-image	
Method	R@5	R@10	R@5	R@10
Embed. Network (VGG+w2v) [43]	33.7	45.5	8.4	12.2
MSE (VGG+w2v)	43.6	56.9	24.3	34.2
SCAN (ResNet+GRU) [26]	73.7	82.6	61.3	72.4
jWAE-MSE (VGG+w2v)	48.5	60.0	29.1	40.7
jWAE-MH (VGG+w2v)	51.1	63.3	37.0	49.1
jWAE-MH+SCAN (ResNet+GRU)	**80.0**	**87.0**	**66.7**	**75.9**

Table 6. Generalization results of model trained only on the COCO training set and evaluated on the Flickr30k test set.

Methods	R@1	R@5	R@10
MCB [11]	48.7	–	–
GroundeR [35]	47.8	–	–
Embedding Network [43]	51.0	70.4	75.5
Similarity Network [43]	51.0	70.3	75.0
SPC [33]	55.4	–	–
IGOP [47]	53.9	–	–
CITE [32]	59.2	–	–
jWAE-MSE	52.5	**75.0**	**81.5**
jWAE-MSE+CITE	**60.4**	–	–

Table 7. Phrase localization on the Flickr30k Entities dataset [34].

on COCO. Fine-grained representations based on attention [26] generalize better compared to [43]. Following Tables 5 and 6, the jWAE-MH framework significantly improves the generalization across datasets further, owing to the semantic continuity from the Gaussian regularization. For image-to-text and text-to-image retrieval we improve the top-5 recall by 5.2% and 2.0% points, respectively, generalizing from Flickr30k to COCO and by 6.3% and 5.4% points, respectively, for generalizing from COCO to Flickr30k.

In Fig. 2, we show modality-specific semantic continuity, where, *e.g.*, in the third row, for an image with "a man and a horse", matching images retrieved by our approach are of 'person-animal interaction' whereas for the supervised approach [43] matching images show the concept 'person'. Similarly, for cross-modal semantic continuity in Table 1, [43] retrieves an image with a concept 'two' while jWAE is able to retrieve the image representative of the given sentence "Two tan dogs play on grass near the wall".

5.2. Phrase localization

We next analyze the benefit of our jWAE framework for phrase localization on the Flickr30k Entities dataset [34]. Phrase localization associates (grounds) a phrase to a region in the image using bounding boxes [5, 35, 43, 47]. Following [43], we formulate phrase localization as a retrieval problem where given an image and a phrase from its associated sentence, the phrase is mapped to the regions in the image. Bounding box proposal regions are extracted with Edge Box [49]. Since we are mainly interested in evaluating the quality of our multimodal embeddings rather than the specific task, we compared to other embedding-based approaches [35, 43]. Additionally, we integrate our jWAE framework with Conditional Image Text Embedding (CITE) [32], which builds on top of the embeddings from the Similarity Network. We also include [33], which uses additional image and language constraints, and [47], which considers all possible bounding boxes based on image concepts like segmentation, word priors, and detection scores.

Dataset and input. The Flickr30k Entities dataset [34] augments the captions of images in Flickr30k with 244k mentions of distinct entities across sentences. The mentions are associated with 276k bounding boxes. Similar to [32, 35, 43], we extract 4096-dimensional visual features from Fast R-CNN [12], finetuned on the PASCAL VOC 2007–2012 datasets [8]. We use proposal regions with IoU ≥ 0.7 as a positive region for a phrase during training. For encoding phrases, PCA is applied to HGLMM features to reduce the dimensionality to 6000 [24].

Results. We compare our method with [11, 35, 43] where multimodal embeddings are evaluated for phrase localization. Following these methods, we use 200 or 500 Edge Box proposals per image. An IoU of at least 0.5 is required for a proposal region to match the ground truth bounding box for a phrase. As shown in Table 7, our method outperforms previous multimodal embedding networks for the phrase localization task by 1.5% for top-1 recall. The gap compared to [43] widens to around 5% for the top-5 and top-10 recall. Moreover, using jWAE as the embedding framework in CITE [32] similarly improves the top-1 recall by 1.2% with new state of the art results for phrase localization. This again highlights the improved accuracy of the embeddings obtained from our semi-supervised jWAE approach.

6. Conclusion

We presented a novel joint Wasserstein autoencoder framework for modeling continuous multimodal representations of images and texts with Gaussian regularization, allowing to better capture the semantic structure in latent representations. Our experiments show that our multimodal embeddings push the current state of the art under full supervision. A key advantage of our method is its generalization capability across datasets, where it significantly outperforms recent methods. We thus believe our semi-supervised approach provides an important step toward learning generalizable multimodal representations, which are a crucial component, *e.g.*, for image captioning [10] in the real world.

Acknowledgement. This work has been supported by the German Research Foundation as part of the Research Training Group *Adaptive Preparation of Information from Heterogeneous Sources (AIPHES)* under grant No. GRK 1994/1.

References

[1] Galen Andrew, Raman Arora, Jeff A. Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *ICML*, 2013. 1, 2

[2] Martín Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN. *ICML*, 2017. 4

[3] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *TPAMI*, 41(2), 2019. 1

[4] Olivier Bousquet, Sylvain Gelly, Ilya Tolstikhin, Carl-Johann Simon-Gabriel, and Bernhard Schölkopf. From optimal transport to generative modeling: The VEGAN cookbook. *arXiv:1705.07642*, 2017. 4

[5] Kan Chen, Rama Kovvuri, Jiyang Gao, and Ram Nevatia. MSRC: Multimodal spatial regression with semantic context for phrase grounding. In *ICMR*, 2017. 8

[6] Jingze Chi and Yuxin Peng. Dual adversarial networks for zero-shot cross-media retrieval. In *IJCAI*, 2018. 2

[7] Aviv Eisenschtat and Lior Wolf. Linking image and text with 2-way nets. In *CVPR*, 2017. 1, 7

[8] Mark Everingham, Luc J. Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The PASCAL visual object classes (VOC) challenge. *IJCV*, 88(2), 2010. 8

[9] Fartash Faghri, David J. Fleet, Ryan Kiros, and Sanja Fidler. VSE++: Improving visual-semantic embeddings with hard negatives. In *BMVC*, 2018. 1, 2, 3, 5, 6, 7

[10] Andrea Frome, Gregory S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *NIPS*, 2013. 2, 8

[11] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *EMNLP*, 2016. 2, 8

[12] Ross B. Girshick. Fast R-CNN. In *ICCV*, 2015. 8

[13] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial networks. In *NIPS*, 2014. 3

[14] Jiuxiang Gu, Jianfei Cai, Shafiq R. Joty, Li Niu, and Gang Wang. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. In *CVPR*, 2018. 2

[15] Guibing Guo, Songlin Zhai, Fajie Yuan, Yuan Liu, and Xingwei Wang. VSE-ens: Visual-semantic embeddings with efficient negative sampling. In *AAAI*, 2018. 2, 7

[16] Tatsuya Harada, Kuniaki Saito, Yusuke Mukuta, and Yoshitaka Ushiku. Deep modality invariant adversarial network for shared representation learning. In *ICCV Workshops*, 2017. 2

[17] David R. Hardoon, Sándor Szedmák, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12), 2004. 2

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[19] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A. Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *ICML*, 2018. 3

[20] Yan Huang, Qi Wu, Chunfeng Song, and Liang Wang. Learning semantic concepts and order for image and sentence matching. In *CVPR*, 2018. 2, 7

[21] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *TPAMI*, 39(4), 2017. 1, 6

[22] Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *ICLR*, 2014. 1, 3

[23] Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *TACL*, 2015. 2

[24] Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. Associating neural word embeddings with deep image representations using Fisher vectors. In *CVPR*, 2015. 1, 6, 7, 8

[25] Pei Ling Lai and Colin Fyfe. Kernel and nonlinear canonical correlation analysis. *Int. J. Neural Syst.*, 10(5), 2000. 2

[26] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *ECCV*, 2018. 1, 2, 6, 7, 8

[27] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 1, 6, 7

[28] Zhuang Ma, Yichao Lu, and Dean P. Foster. Finding linear structure in large datasets with scalable canonical correlation analysis. In *ICML*, 2015. 2

[29] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, and Ian J. Goodfellow. Adversarial autoencoders. *ICLR Workshops*, 2016. 3

[30] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *ICLR Workshops*, 2013. 5, 6

[31] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. Multimodal deep learning. In *ICML*, 2011. 2

[32] Bryan A. Plummer, Paige Kordas, M. Hadi Kiapour, Shuai Zheng, Robinson Piramuthu, and Svetlana Lazebnik. Conditional image-text embedding networks. In *ECCV*, 2018. 8

[33] Bryan A. Plummer, Arun Mallya, Christopher M. Cervantes, Julia Hockenmaier, and Svetlana Lazebnik. Phrase localization and visual relationship detection with comprehensive image-language cues. In *ICCV*, 2017. 8

[34] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *IJCV*, 123(1), 2017. 1, 8

[35] Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. Grounding of textual phrases in images by reconstruction. In *ECCV*, 2016. 8

[36] Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *NIPS*, 2016. 6

[37] Mike Schuster and Kuldip K. Paliwal. Bidirectional recurrent neural networks. *TSP*, 45(11), 1997. 5

[38] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015. 6

[39] I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schölkopf. Wasserstein auto-encoders. In *ICLR*, 2018. 1, 3, 4

[40] Yao-Hung Hubert Tsai, Liang-Kang Huang, and Ruslan Salakhutdinov. Learning robust visual-semantic embeddings. In *ICCV*, 2017. 1, 2, 6, 7

[41] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *CVPR*, 2017. 3

[42] Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen. Adversarial cross-modal retrieval. In *ACM MM*, 2017. 2

[43] Liwei Wang, Yin Li, Jing Huang, and Svetlana Lazebnik. Learning two-branch neural networks for image-text matching tasks. *TPAMI*, 41(2), 2019. 1, 2, 3, 6, 7, 8

[44] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *CVPR*, 2016. 3, 6

[45] Jonatas Wehrmann and Rodrigo C. Barros. Bidirectional retrieval made simple. In *CVPR*, 2018. 2

[46] Jason Weston, Samy Bengio, and Nicolas Usunier. WSABIE: Scaling up to large vocabulary image annotation. In *IJCAI*, 2011. 2

[47] Raymond Yeh, Jinjun Xiong, Wen-Mei W. Hwu, Minh Do, and Alexander G. Schwing. Interpretable and globally optimal prediction for textual grounding using image concepts. In *NIPS*, 2017. 8

[48] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *ACL*, 2014. 1, 6, 7

[49] C Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *ECCV*, 2014. 8