Biconditional Generative Adversarial Networks for Multiview Learning with Missing Views

Anastasiia Doinychko, Massih-Reza Amini

Biconditional Generative Adversarial Networks for Multiview Learning with Missing Views

Anastasiia Doinychko[1,2], Massih-Reza Amini[1]
[1]Université Grenoble Alpes
LIG/IMAG, 700 av. centrale
38401 Saint-Martin d'Hères
France
[2]Mentor Graphics
110 Rue Blaise Pascal
38330 Montbonnot-Saint-Martin
France
FirstName.Lastname@univ-grenoble-alpes.fr

Abstract

In this paper, we present a conditional GAN with two generators and a common discriminator for multiview learning problems where observations have two views, but one of them may be missing for some of the training samples. This is for example the case for multilingual collections where documents are not available in all languages. Some studies tackled this problem by assuming the existence of view generation functions to approximately complete the missing views; for example Machine Translation to translate documents into the missing languages. These functions generally require an external resource to be set and their quality has a direct impact on the performance of the learned multiview classifier over the completed training set. Our proposed approach addresses this problem by jointly learning the missing views and the multiview classifier using a tripartite game with two generators and a discriminator. Each of the generators is associated to one of the views and tries to fool the discriminator by generating the other missing view conditionally on the corresponding observed view. The discriminator then tries to identify if for an observation, one of its views is completed by one of the generators or if both views are completed along with its class. Our results on a subset of Reuters RCV1/RCV2 collections show that the discriminator achieves significant classification performance; and that the generators learn the missing views with high quality without the need of any consequent external resource.

1 Introduction

We address the problem of multiview learning with Generative Adversarial Networks (GANs) in the case where some observations may have missing views without there being an external resource to complete them. This is a typical situation in many applications where different sources generate different views of

samples unevenly; like text information present in all Wikipedia pages while images are more scarce. Another example is multilingual text classification where documents are available in two languages and share the same set of classes while some are just written in one language. Previous works supposed the existence of view generating functions to complete the missing views before deploying a learning strategy [2]. However, the performance of the global multiview approach is biased by the quality of the generating functions which generally require external resources to be set. The challenge is hence to learn an efficient model from the multiple views of training data without relying on an extrinsic approach to generate altered views for samples that have missing ones.

In this direction, GANs provide a propitious and broad approach with a high ability to seize the underlying distribution of the data and create new samples [11]. These models have been mostly applied to image analysis and major advances have been made on generating realistic images with low variability [7, 15, 16]. GANs take their origin from the game theory and are formulated as a two players game formed by a generator G and a discriminator D. The generator takes a noise z and produces a sample $G(z)$ in the input space, on the other hand the discriminator determines whenever a sample comes from the true distribution of the data or if it is generated by G. Other works included an inverse mapping from the input to the latent representation, mostly referred to as BiGANs, and showed the usefulness of the learned feature representation for auxiliary discriminant problems [8, 9]. This idea paved the way for the design of efficient approaches for generating coherent synthetic views of an input image [21, 14, 6].

In this work, we propose a GAN based model for bilingual text classification, called `Cond`^{2}`GANs`, where some training documents are just written in one language. The model learns the representation of missing versions of bilingual documents jointly with the association to their respective classes, and is composed of a discriminator D and two generators G_1 and G_2 formulated as a tripartite game. For a given document with a missing version in one language, the corresponding generator induces the latter conditionally on the observed one. The training of the generators is carried out by minimizing a regularized version of the cross-entropy measure proposed for multi-class classification with GANs [19] in a way to force the models to generate views such that the completed bilingual documents will have high class assignments. At the same time, the discriminator learns the association between documents and their classes and distinguishes between observations that have their both views and those that got a completed view by one of the generators. This is achieved by minimizing an aggregated cross-entropy measure in a way to force the discriminator to be certain of the class of observations with their complete views and uncertain of the class of documents for which one of the versions was completed. The regularization term in the objectives of generators is derived from an adapted feature matching technique [17] which is an effective way for preventing from situations where the models become unstable; and which leads to fast convergence.

We demonstrate that generated views allow to achieve state-of-the-art results on a subset of Reuters RCV1/RCV2 collections compared to multiview

approaches that rely on Machine Translation (MT) for translating documents into languages in which their versions do not exist; before training the models. Importantly, we exhibit qualitatively that generated documents have meaningful translated words bearing similar ideas compared to the original ones; and that, without employing any large external parallel corpora to learn the translations as it would be the case if MT were used. More precisely, this work is the first to :

- Propose a new tripartite GAN model that makes class prediction along with the generation of high quality document representations in different input spaces in the case where the corresponding versions are not observed (Section 3.2);
- Achieve state-of-the art performance compared to multiview approaches that rely on external view generating functions on multilingual document classification; and which is another challenging application than image analysis which is the domain of choice for the design of new GAN models (Section 4.2);
- Demonstrate the value of the generated views within our approach compared to when they are generated using MT (Section 4.2).

2 Related work

Multiview learning has been an active domain of research these past few years. Many advances have been made on both theoretic and algorithmic sides [5, 12]. The three main families of techniques for (semi-)supervised learning are (kernel) Canonical Correlation Analysis (CCA), Multiple kernel learning (MKL) and co-regularization. CCA finds pairs of highly correlated subspaces between the views that is used for mapping the data before training, or integrated in the learning objective [3, 10]. MKL considers one kernel per view and different approaches have been proposed for their learning. In one of the earliest work, [4] proposed an efficient algorithm based on sequential minimization techniques for learning a corresponding support vector machine defined over a convex non-smooth optimization problem. Co-regularization techniques tend to minimize the disagreement between the single-view classifiers over their outputs on unlabeled examples by adding a regularization term to the objective function [18]. Some approaches have also tackled the tedious question of combining the predictions of the view specific classifiers [20]. However all these techniques assume that views of a sample are complete and available during training and testing.

Recently, many other studies have considered the generation of multiple views from a single input image using GANs [14, 21, 23] and have demonstrated the intriguing capacity of these models to generate coherent unseen views. The former approaches rely mostly on an encoder-encoder network to first map images into a latent space and then generate their views using an inverse mapping. This is a very exciting problem, however, our learning objective differs from

these approaches as we are mostly interested in the classification of muti-view samples with missing views. The most similar work to ours that uses GANs for multiview classification is probably [6]. This approach generates missing views of images in the same latent space than the input image, while `Cond`^{2}`GANs` learns the representations of the missing views in their respective input spaces conditionally on the observed ones which in general are from other feature spaces. Furthermore, `Cond`^{2}`GANs` benefits from low complexity and stable convergence which has been shown to be lacking in the previous approach.

Another work which has considered multiview learning with incomplete views, for also document classification, is [2]. The authors proposed a Rademacher complexity bounds for a multiview Gibbs classifier trained on multilingual collections where the missing versions of documents have been generated by Machine Translation systems. Their bounds exhibit a term corresponding to the quality of the MT system generating the views. The bottleneck is that MT systems depend on external resources, and they require a huge amount of parallel collections containing documents and their translations in all languages of interest for their tuning. For rare languages, this can ultimately affect the performance of the learning models. Regarding these aspects our proposed approach differs from all the previous studies, as we do not suppose the existence of parallel training sets nor MT systems to generate the missing versions of the training observations.

3 Cond^{2}GANs

In the following sections, we first present the basic definitions which will serve to our problem setting, and then the proposed model for multiview classification with missing views.

3.1 Framework and problem setting

We consider multiclass classification problems, where a bilingual document is defined as a sequence $\mathbf{x} = (x^1, x^2) \in \mathcal{X}$ that belongs to one and only one class $\mathbf{y} \in \mathcal{Y} = \{0,1\}^K$. The class membership indicator vector $\mathbf{y} = (y_k)_{1 \leq k \leq K}$, of each bilingual document, has all its components equal to 0 except the one that indicates the class associated with the example which is equal to one. Here we suppose that $\mathcal{X} = (\mathcal{X}_1 \cup \{\perp\}) \times (\mathcal{X}_2 \cup \{\perp\})$, where $x^v = \perp$ means that the v-th view is not observed. Hence, each observed view $x^v \in \mathbf{x}$ is such that $x^v \neq \perp$ and it provides a representation of $\mathbf{x}$ in a corresponding input space $\mathcal{X}_v \subseteq \mathbb{R}^{d_v}$. Following the conclusions of the co-training study [5], our framework is based on the following main assumption :

Assumption 1 ([5]) *Observed views are not completely correlated, and are equally informative.*

Furthermore, we assume that each example $(\mathbf{x}, \mathbf{y})$ is identically and independently distributed (i.i.d.) according to a fixed yet unknown distribution $\mathcal{D}$ over

$\mathcal{X} \times \mathcal{Y}$, and that at least one of its views is observed. Additionally, we suppose to have access to a training set $\mathcal{S} = \{(\mathbf{x}_i, \mathbf{y}_i); i \in \{1, \ldots, m\}\} = \mathcal{S}_F \sqcup \mathcal{S}_1 \sqcup \mathcal{S}_2$ of size m drawn i.i.d. according to $\mathcal{D}$, where $\mathcal{S}_F = \{\left(\left(x_i^1, x_i^2\right), \mathbf{y}_i\right) \mid i \in \{1, \ldots, m_F\}\}$ denotes the subset of training samples with their both complete views and $\mathcal{S}_1 = \{\left(\left(x_i^1, \perp\right), \mathbf{y}_i\right) \mid i \in \{1, \ldots, m_1\}\}$ (respectively $\mathcal{S}_2 = \{\left(\left(\perp, x_i^2\right), \mathbf{y}_i\right) \mid i \in \{1, \ldots, m_2\}\}$) is the subset of training samples with their second (respectively first) view that is not observed (i.e. $m = m_F + m_1 + m_2$).

It is possible to address this problem using existing techniques; for example, by learning singleview classifiers independently on the examples of $\mathcal{S} \sqcup \mathcal{S}_1$ (respectively $\mathcal{S} \sqcup \mathcal{S}_2$) for the first (respectively second) view. To make prediction, one can then combine the outputs of the classifiers [20] if both views of a test example are observed; or otherwise, use one of the outputs corresponding to the observed view. Another solution is to apply multiview approaches over the training samples of $\mathcal{S}_F$; or over the whole training set $\mathcal{S}$ by completing the views of examples in $\mathcal{S}_1$ and $\mathcal{S}_2$ beforhand using external view generation functions.

3.2 The Tripartite Game

As an alternative, the learning objective of our proposed approach is to generate the missing views of examples in $\mathcal{S}_1$ and $\mathcal{S}_2$, jointly with the learning of the association between the multiview samples (with all their views complete or completed) and their classes. The proposed model consists of three neural networks that are trained using an objective implementing a three players game between a discriminator, D, and two generators, G_1 and G_2. The game that these models play is depicted in Figure 1 and it can be summarized as follows. At each step, if an observation is chosen with a missing view, the corresponding generator – G_1 (respectively G_2) if the first (respectively second) view is missing – produces the view from random noise conditionally on the observed view in a way to fool the discriminator. On the other hand, the discriminator takes as input an observation with both of its views complete or completed and, classifies it if the views are ini-

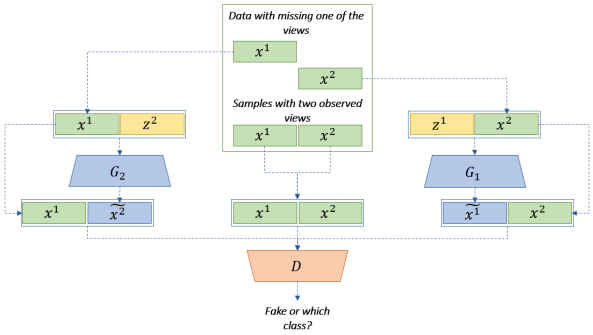


Figure 1: A visual representation of the proposed GAN model composed of three neural networks; a discriminator D and two generators G_1 and G_2. The missing view of an observation is completed by the corresponding generator conditionally on its observed view. The discriminator is trained to recognize between observations having their views completed and those with complete initial views as well as their classes.

tially observed or tells if a view was produced by one of the generators. Formally, both generators G_1 and G_2 take as input; samples from respectively the training subsets $\mathcal{S}_2$ and $\mathcal{S}_1$; as well as random noise drawn from uniform distribution defined over the input space of the missing view and produce the corresponding pseudo-view, which is missing; i.e. $G_1(z^1, x^2) = \tilde{x}^1$ and $G_2(x^1, z^2) = \tilde{x}^2$. These models are learned in a way to replicate the conditional distributions $p(x^1|x^2, z^1)$ and $p(x^2|x^1, z^2)$; and inherently define two probability distributions, denoted respectively by p_{G_1} and p_{G_2}, as the distribution of samples if both views where observed i.e. $(\tilde{x}^1, x^2) \sim p_{G_1}(x^1, x^2)$, $(x^1, \tilde{x}^2) \sim p_{G_2}(x^1, x^2)$. On the other hand, the discriminator takes as input a training sample; either from the set $\mathcal{S}_F$, or from one of the training subsets $\mathcal{S}_1$ or $\mathcal{S}_2$ where the missing view of the example is generated by one of the generators accordingly. The task of D is then to recognize observations from $\mathcal{S}_1$ and $\mathcal{S}_2$ that have completed views by G_1 and G_2 and to classify examples from $\mathcal{S}$ to their true classes. To achieve this goal we add a fake class, $K+1$, to the set of classes, $\mathcal{Y}$, corresponding to samples that have one of their views generated by G_1 or G_2. The dimension of the discriminator's output is hence set to $K+1$ which by applying softmax is supposed to estimate the posterior probability of classes for each multiview observation (with complete or completed views) given in input. For an observation $\mathbf{x} \in \mathcal{X}$, we use $D_{K+1}(\mathbf{x}) = p_D(y = K+1|\mathbf{x})$ to estimate the probability that one of its views is generated by G_1 or G_2. As the task of the generators is to produce good quality views such that the observation with the completed view will be assigned to its true class with high probability, we follow [17] by supplying the discriminator to not get fooled easily as stated in the following assumption :

Assumption 2 ([17]) *An observation* $\mathbf{x}$ *has one of its views generated by* G_1 *or* G_2*; if and only if* $D_{K+1}(\mathbf{x}) > \sum_{k=1}^{K} D_k(\mathbf{x})$.

In the case where; $D_{K+1}(\mathbf{x}) \leq \sum_{k=1}^{K} D_k(\mathbf{x})$ the observation $\mathbf{x}$ is supposed to have its both views observed and it is affected to one of the classes following the rule; $\max_{k=\{1,\dots,K\}} D_k(\mathbf{x})$. The overall learning objective of `Cond`^{2}`GANs` is to train the generators to produce realistic views indistinguishable with the real ones, while the discriminator is trained to classify multiview observations having their complete views and to identify view generated samples. If we denote by p_{real} the marginal distribution of multiview observations with their both views observed (i.e. $(x^1, x^2) = p_{real}(x^1, x^2)$); the above procedure resumes to the following tripartite minmax game with value function $V(D, G_1, G_2)$:

$$\begin{aligned} \max_{D} \min_{G_1, G_2} V(D, G_1, G_2) = \; & \mathbb{E}_{(x^1,x^2)\sim p_{real}} \left[\log p_D(y < K+1|x^1, x^2)\right] \\ & + \frac{1}{2}\mathbb{E}_{(\tilde{x}^1,x^2)\sim p_{G_1}} \left[\log p_D(y = K+1|\tilde{x}^1, x^2)\right] \\ & + \frac{1}{2}\mathbb{E}_{(x^1,\tilde{x}^2)\sim p_{G_2}} \left[\log p_D(y = K+1|x^1, \tilde{x}^2)\right] \end{aligned} \quad (1)$$

Note that, following Assumption 1, we impose the generators to produce equally informative views by assigning the same weight to their corresponding terms in V (Eq. 1).

3.3 Analyses and convergence

From the outputs of the discriminator we build an auxiliary function $\mathbf{D}$ equal to the sum of the first K outputs associated to the true classes :

$$\forall \mathbf{x} \in \mathcal{X}; \mathbf{D}(\mathbf{x}) = \sum_{k=1}^{K} p_D(y = k \mid \mathbf{x}) \tag{2}$$

In this following, we provide a theoretical analysis of `Cond`^{2}`GANs` involving the auxiliary function $\mathbf{D}$ (Eq. 2) under nonparametric hypotheses.

Proposition 1 *For fixed generators G_1 and G_2, the minmax game defined in (Eq. 1) leads to the following optimal discriminator $\mathbf{D}^*_{G_1,G_2}$:*

$$\mathbf{D}^*_{G_1,G_2}(x^1, x^2) = \frac{p_{real}(x^1, x^2)}{p_{real}(x^1, x^2) + p_{G_{1,2}}(x^1, x^2)}, \tag{3}$$

where $p_{G_{1,2}}(x^1, x^2) = \frac{1}{2}(p_{G_1}(x^1, x^2) + p_{G_2}(x^1, x^2))$.

Proof. The proof follows from [11]. Let

$$\forall \mathbf{x} = (x^1, x^2), \mathbf{D}(\mathbf{x}) = \sum_{k=1}^{K} D_k(\mathbf{x})$$

From Assumption 2, and the fact that for any observation $\mathbf{x}$ the outputs of the discriminator sum to one i.e. $\sum_{k=1}^{K+1} D_k(\mathbf{x}) = 1$, the value function V writes :

$$V(\mathbf{D}, G_1, G_2) = \iint \log(\mathbf{D}(x^1, x^2)) p_{real}(x^1, x^2) dx^1 dx^2 +$$
$$\frac{1}{2} \iint \log(1 - \mathbf{D}(x^1, x^2)) p_{G_1}(x^1, x^2) dx^1 dx^2 + \frac{1}{2} \iint \log(1 - \mathbf{D}(x^1, x^2)) p_{G_2}(x^1, x^2) dx^1 dx^2$$

For any $(\alpha, \beta, \gamma) \in \mathbb{R}^3 \backslash \{0, 0, 0\}$; the function $z \mapsto \alpha \log z + \frac{\beta}{2} \log(1 - z) + \frac{\gamma}{2} \log(1 - z)$ reaches its maximum at $z = \frac{\alpha}{\alpha + \frac{1}{2}(\beta + \gamma)}$, which ends the proof as the discrimintaor does not need to be defined outside the supports of p_{data}, p_{G_1} and p_{G_2}. □

By plugging back $\mathbf{D}^*_{G_1,G_2}$ (Eq. 3) into the value function V we have the following necessary and sufficient condition for attaining the global minimum of this function :

Theorem 1 *The global minimum of the function $V(G_1, G_2)$ is attained if and only if*

$$p_{real}(x^1, x^2) = \frac{1}{2}(p_{G_1}(x^1, x^2) + p_{G_2}(x^1, x^2)). \tag{4}$$

At this point, the minimum is equal to $-\log 4$.

Proof. By plugging back the expression of $\mathbf{D}^*$ (Eq. 3), into the value function V, it comes

$$V(\mathbf{D}^*, G_1, G_2) = \iint \log\left(\frac{p_{real}(x^1, x^2)}{p_{real}(x^1, x^2) + p_{G_{1,2}}(x^1, x^2)}\right) p_{real}(x^1, x^2) dx^1 dx^2 + \iint \log\left(\frac{p_{G_{1,2}}(x^1, x^2)}{p_{real}(x^1, x^2) + p_{G_{1,2}}(x^1, x^2)}\right) p_{G_{1,2}}(x^1, x^2) dx^1 dx^2$$

Which from the definition of the Kullback Leibler (KL) and the Jensen Shannon divergence (JSD) can be rewritten as

$$\begin{aligned} V(\mathbf{D}^*, G_1, G_2) &= -\log 4 + KL\left(p_{real} \,\middle\|\, \frac{p_{real} + p_{G_{1,2}}}{2}\right) + KL\left(p_{G_{1,2}} \,\middle\|\, \frac{p_{real} + p_{G_{1,2}}}{2}\right) \\ &= -\log 4 + 2JSD\big(p_{real} \,\|\, p_{G_{1,2}}\big) \end{aligned}$$

The JSD is always positive and $JSD\big(p_{real} \,\|\, p_{G_{1,2}}\big) = 0$ if and only if $p_{real} = p_{G_{1,2}}$ which ends the proof $\square$

From Equation 4, it is straightforward to verify that $p_{real}(x^1, x^2) = p_{G_1}(x^1, x^2) = p_{G_2}(x^1, x^2)$ is a global Nash equilibrium but it may not be unique. In order to ensure the uniqueness, we add the Jensen-Shannon divergence between the distribution p_{G_1} and p_{real} and p_{G_2} and p_{real} the value function V (Eq. 1) as stated in the corollary below.

Corollary 1 *The unique global Nash equilibrium of the augmented value function :*

$$\bar{V}(\mathbf{D}, G_1, G_2) = V(\mathbf{D}, G_1, G_2) + JSD(p_{G_1}||p_{real}) + JSD(p_{G_2}||p_{real}), \tag{5}$$

is reached if and only if

$$p_{real}(x^1, x^2) = p_{G_1}(x^1, x^2) = p_{G_2}(x^1, x^2), \tag{6}$$

where $V(\mathbf{D}, G_1, G_2)$ is the value function defined in Equation (1) *and $JSD(p_{G_1}||p_{real})$ is the Jensen-Shannon divergence between the distribution p_{G_1} and p_{real}.*

Proof. The proof follows from the positivness of JSD and the necessary and sufficient condition for it to be equal to 0. Hence, $\bar{V}(\mathbf{D}, G_1, G_2)$ reaches it minimum $-\log 4$, iff $p_{G_1} = p_{real} = p_{G_2}$. $\square$

This result suggests that at equilibrium, both generators produce views such that observations with their completed view follow the same real distribution than those which have their both views observed.

3.4 Algorithm and Implementation

In order to avoid the collapse of the generators [17], we perform minibatch discrimination by allowing the discriminator to have access to multiple samples in combination. From this perspective, the minmax game (Eq. 1) is equivalent to the maximization of a cross-entropy loss, and we use minibatch training to learn the parameters of the three models. The corresponding empirical errors estimated over a minibatch $\mathcal{B}$ that

contains m_b samples from each of the sets $\mathcal{S}_F$, $\mathcal{S}_1$ and $\mathcal{S}_2$ are :

$$\mathcal{L}_D(\mathcal{B}) = -\frac{1}{m_b}\sum_{\mathbf{x}\in\mathcal{B}\cap\mathcal{S}_F}\frac{1}{K+1}\sum_{k=1}^{K} y_k \log\left[D_k(x^1,x^2)\right] \tag{7}$$
$$-\frac{1}{2m_b}\sum_{\mathbf{x}\in\mathcal{B}\cap\mathcal{S}_1}\log\left[D_{K+1}(G_1(z^1,x^2),x^2))\right] - \frac{1}{2m_b}\sum_{\mathbf{x}\in\mathcal{B}\cap\mathcal{S}_2}\log\left[D_{K+1}(x^1,G_2(x^1,z^2))\right]$$

$$\mathcal{L}_{G_v}(\mathcal{B}) = -\frac{1}{m_b}\sum_{\mathbf{x}\in\mathcal{B}\cap\mathcal{S}_v}\frac{1}{K+1}\sum_{k=1}^{K} y_k \log\left[D_k(G_v(z^v,x^{3-v}),x^{3-v})\right] + \mathcal{L}^v_{FM}; v\in\{1,2\} \tag{8}$$

In order, to be inline with the premises of Corollary 1; we empirically tested different solutions and the most effective one that we found was the feature matching technique proposed in [17], which addressed the problem of instability for the learning of generators by adding a penalty term $\mathcal{L}^v_{FM} = \|\mathbb{E}_{p_{real}} f(x^1,x^2) - \mathbb{E}_{p_{G_v}} f(x^{3-v},G_v(x^v))\|$, $v\in\{1,2\}$ to their corresponding objectives (Eq. (8)). Where, $\|.\|$ is the ℓ_2 norm and f is the sigmoid activation function on an intermediate layer of the discriminator. The overall algorithm of `Cond`^{2}`GANs` is shown above. The parameters of the three neural networks are first initialized using *Xavier*. For a given number of iterations T, minibatches of size $3m_b$ are randomly sampled from the sets $\mathcal{S}_F$, $\mathcal{S}_1$ and $\mathcal{S}_2$. Minibatches of noise vectors are randomly drawn from the uniform distribution. Models parameters of the discriminator and both generators are then sequentially updated using *Adam* optimization algorithm [13]. We implemented our method by having two layers neural networks for each of the components of `Cond`^{2}`GANs`. These neural nets are composed of 200 nodes in hidden layers with a sigmoid activation function. Since the values of the generated samples are supposed to approximate any possible real value, we do not use the activation function in the outputs of both generators.[1]

Minibatch stochastic training of `Cond`^{2}`GANs`

Input: A training set $\mathcal{S} = \mathcal{S}_F \sqcup \mathcal{S}_1 \sqcup \mathcal{S}_2$
Initialization: Size of minibatches, m_b
Use $Xavier$ initializer to initialize discriminator and generators parameters, respectively $\theta_d^{(0)}, \theta_{g_1}^{(0)}, \theta_{g_2}^{(0)}$
for $i = 0 \dots T-1$ **do**
  Sample randomly a minibatch $\mathcal{B}_i$ of size $3m_b$ from $\mathcal{S}_1$, $\mathcal{S}_2$ and $\mathcal{S}_F$; create minibatches of noise vector z^1, z^2 from $\mathcal{U}(-1,1)$
  $\theta_d^{(i+1)} \leftarrow Adam(\mathcal{L}_D(\mathcal{B}_i), \theta_d^{(i)}, \alpha, \beta)$ # Update of D
  $\theta_{g1}^{(i+1)} \leftarrow Adam(\mathcal{L}_{G_1}(\mathcal{B}_i), \theta_{g1}^{(i)}, \alpha, \beta)$ # Update of G_1
  $\theta_{g2}^{(i+1)} \leftarrow Adam(\mathcal{L}_{G_2}(\mathcal{B}_i), \theta_{g2}^{(i)}, \alpha, \beta)$ # Update of G_2
end

4 Experiments

In this Section, we present experimental results aimed at evaluating how the generation of views by `Cond`^{2}`GANs` can help to take advantage of existing training examples, with many having an incomplete view, in order to learn an efficient classification function. We perform experiments on a publicly available collection, extracted from Reuters

[1] We will release the code for reproducibility and research purpose.

RCV1/RCV2, that is proposed for multilingual multiclass text categorization[2] (Table 1). The dataset contains feature vectors of documents originally presented in five languages (EN, FR, GR, IT, SP). In our experiments, we consider four pairs of languages with always English as one of the views ((EN,FR),(EN,SP),(EN,IT),(EN,GR)). Documents in different languages belong to one and only one class within the same set of classes ($K = 6$); and they also have translations into all the other languages. These translations are obtained from a state-of-the-art Statistical Machine Translation system [22] trained over the Europal parallel collection using about 8.10^6 sentences for the 4 considered pairs of languages.[3]

Table 1: The statistics of RCV1/RCV2 Reuters data collection used in our experiments.

Language	# docs	(%)	dim
EN	18,758	16.78	21,531
FR	26,648	23.45	24,893
GR	29,953	26.80	34,279
IT	24,039	21.51	15,506
SP	12,342	11.46	11,547
Total	111,740		

Class	Size (all lang.)	(%)
C15	18,816	16.84
CCAT	21,426	19.17
E21	13,701	12.26
ECAT	19,198	17.18
GCAT	19,178	17.16
M11	19,421	17.39

4.1 Experimental Setup

In our experiments, we consider the case where the number of training documents having their two versions is much smaller than those with only one of their available versions (i.e. $m_F \ll m_1 + m_2$). This corresponds to the case where the effort of gathering documents in different languages is much less than translating them from one language to another. To this end, we randomly select $m_F = 300$ samples having their both views, $m_1 = m_2 = 6000$ samples with one of their views missing and the remaining samples without their translations for test. In order to evaluate the quality of generated views by Cond^{2}GANs we considered two scenarios. In the first one (denoted by $\mathbf{T}_{\text{EN}\bar{v}}$), we test on English documents by considering the generation of these documents with respect to the other view ($v \in \{\text{FR}, \text{GR}, \text{IT}, \text{SP}\}$) using the corresponding generator. In the second scenario (denoted by $\mathbf{T}_{\bar{\text{EN}}v}$), we test on documents that are written in another language than English by considering their generation on English provided by the other generator. For evaluation, we test the following four classification approaches along with Cond^{2}GANs; one singleview approach and four multiview approaches. In the singleview approach (denoted by $\mathbf{c}_v$) classifiers are the same than the discriminator and they are trained on the part of the training set with examples having their corresponding view observed. The multiview approaches are MKL [4], co-classification (co-classif) [1], unanimous vote (mv_b) [2]. Results are evaluated over the test set using the accuracy and the F_1 measure which is the harmonic average of

[2] https://archive.ics.uci.edu/ml/datasets/Reuters+RCV1+RCV2+Multilingual, +Multiview+Text+Categorization+Test+collection

[3] http://www.statmt.org/europarl/

precision and recall. The reported performance are averaged over 20 random (train/test) sets, and the parameters of Adam optimization algorithm are set to $\alpha = 10^{-4}$, $\beta = 0.5$.

4.2 Experimental Results

On the value of the generated views. We start our evaluation by comparing the F_1 scores over the test set, obtained with `Cond`^{2}`GANs` and a neural network having the same architecture than the discriminator D of `Cond`^{2}`GANs` trained over the concatenated views of documents in the training set where the missing views are generated by Machine Translation. Figure 2 shows these results. Each point represents a class, where its abscissa (resp. ordinate) represents the test F_1 score of the Neural Network trained using MT (resp. one of the generators of `Cond`^{2}`GANs`) to complete the missing views. All of the classes, in the different language pair scenarios, are above the line of equality, suggesting that the generated views by `Cond`^{2}`GANs` provide higher value information than translations provided by MT for learning the Neural Network. This is an impressive finding, as the resources necessary for the training of MT is large (8.10^6 pairs of sentences and their translations); while `Cond`^{2}`GANs` does both view completion and discrimination using only the available training data. This is mainly because both generators induce missing views with the same distribution than real pairs of views as stated in Corollary 1.

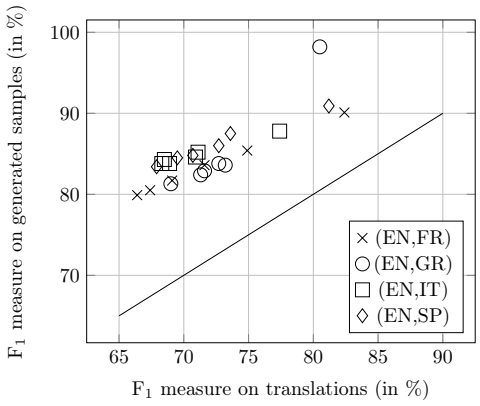


Figure 2: F_1-score per class measured for test predictions made by a Neural-Network, with the same architecture than the discriminator of `Cond`^{2}`GANs`, and trained over documents where their missing views are generated by MT, or by G_1 or G_2.

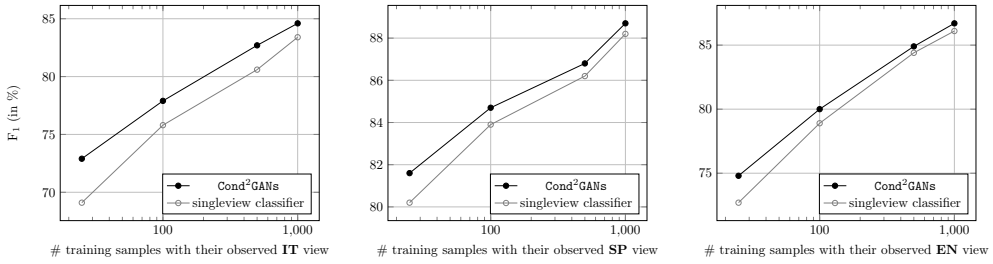
Comparison between multiview approaches. We now examine the gains, in terms of accuracy, of learning the different multiview approaches on a collection where for other approaches than `Cond`^{2}`GANs` the missing views are completed by one of the generators of our model. Table 2 summarizes these results obtained by `Cond`^{2}`GANs`, `MKL`, `co-classif`, and $\texttt{mv}_b$ for both test scenarios. In all cases `Cond`^{2}`GANs`, provides significantly better results than other approaches. This provides empirical evidence of the effectiveness of the joint view generation and class prediction of `Cond`^{2}`GANs`. Furthermore, `MKL`, `co-classif` and `Cond`^{2}`GANs` are binary classification models and tackle the multiclass classification case with one vs all strategy making them to suffer from class imbalance problem. Results

obtained using the F_1 measure are in line with those of Table 2 and they are not reported for the sake of space.

Table 2: Test classification accuracy averaged over 20 random training/test sets. For each of the pairs of languages, the best result is in bold, and a $^{\downarrow}$ indicates a result that is statistically significantly worse than the best, according to a Wilcoxon rank sum test with $p < .01$.

Approaches	(EN,$v =$ FR)		(EN,$v =$ GR)		(EN,$v =$ IT)		(EN,$v =$ SP)	
	$\mathbf{T}_{\texttt{EN}\bar{v}}$	$\mathbf{T}_{\bar{\texttt{EN}}v}$	$\mathbf{T}_{\texttt{EN}\bar{v}}$	$\mathbf{T}_{\bar{\texttt{EN}}v}$	$\mathbf{T}_{\texttt{EN}\bar{v}}$	$\mathbf{T}_{\bar{\texttt{EN}}v}$	$\mathbf{T}_{\texttt{EN}\bar{v}}$	$\mathbf{T}_{\bar{\texttt{EN}}v}$
MKL	75.6$^{\downarrow}$	77.3$^{\downarrow}$	79.4$^{\downarrow}$	79.6$^{\downarrow}$	78.4$^{\downarrow}$	79.8$^{\downarrow}$	81.2$^{\downarrow}$	83.5$^{\downarrow}$
co-classif	81.4$^{\downarrow}$	83.2$^{\downarrow}$	84.3$^{\downarrow}$	81.6$^{\downarrow}$	82.7$^{\downarrow}$	82.5$^{\downarrow}$	85.1$^{\downarrow}$	86.2$^{\downarrow}$
mv$_b$	83.1$^{\downarrow}$	84.5$^{\downarrow}$	85.2$^{\downarrow}$	79.9$^{\downarrow}$	84.3$^{\downarrow}$	82.1$^{\downarrow}$	84.4$^{\downarrow}$	86.2$^{\downarrow}$
Cond^{2}GANs	**85.3**	**85.1**	**86.6**	**82.9**	**85.3**	**84.5**	**86.5**	**88.3**

Impact of the increasing number of observed views. In Figure 3, we compare F_1 measures between Cond^{2}GANs and one of the single-view classifiers with an increasing number of training samples, having the view corresponding to the singleview classifier observed; while the number of training examples with the other observed view is fixed. With an increasing number of training samples, the corresponding singleview classifier gains in performance. On the other hand, Cond^{2}GANs can leverage the lack of information from training examples having their other view observed, making that the difference of performance between these models for small number of training samples is higher.



(a) $m_{\texttt{EN}} = 6000$ (b) $m_{\texttt{EN}} = 6000$ (c) $m_{\texttt{GR}} = 6000$

Figure 3: F_1 measure of Cond^{2}GANs and a singleview classifier ($\mathbf{c}_v$) for an increasing number of training samples with the corresponding view that is observed. The number of training examples corresponding to the other view ($m_{\not v} = 6000$); and the number of training examples with their both views observed is $m_F = 300$.

5 Conclusion

In this paper we presented Cond^{2}GANs for multiview multiclass classification where observations may have missing views. The model consists of three neural-networks implementing a three players game between a discriminator and two generators. For an observation with a missing view, the corresponding generator produces the view conditionally on the other observed one. The discriminator is trained to recognize observations with a generated view from others having their views complete and to classify the latter into one of the existing classes. We evaluate the effectiveness of our approach on another challenging application than image analysis which is the domain of choice for the design of new GAN models. Our experiments on a subset of Reuters RCV1/RCV2 show the effectiveness of Cond^{2}GANs to generate high quality views allowing to achieve significantly better results, compared to the case where the missing views are generated by Machine Translation which requires a large collection of sentences and their translations to be tuned. As future study, we will be working on the generalization of the proposed model to more than 2 views. One possible direction is the use of an aggregation function of available views as a condition to the generators.

Acknowledgment

This work has been partially supported by MIAI@Grenoble Alpes (ANR-19-P3IA-0003).

References

[1] Amini, M.R., Goutte, C.: A co-classification approach to learning from multilingual corpora. Machine Learning Journal **79**(1-2), 105–121 (2010)

[2] Amini, M.R., Usunier, N., Goutte, C.: Learning from multiple partially observed views - an application to multilingual text categorization. In: Advances in Neural Information Processing Systems 22, pp. 28–36 (2009)

[3] Bach, F.R., Jordan, M.I.: Kernel independent component analysis. Journal of Machine Learning Research pp. 1–48 (2003)

[4] Bach, F.R., Lanckriet, G.R.G., Jordan, M.I.: Multiple kernel learning, conic duality, and the smo algorithm. In: Proceedings of the 21^{st} International Conference on Machine Learning (ICML) (2004)

[5] Blum, A., Mitchell, T.: Combining labeled and unlabeled data with co-training. In: Proceedings of the Eleventh Annual Conference on Computational Learning Theory (COLT). pp. 92–100 (1998)

[6] Chen, M., Denoyer, L.: Multi-view Generative Adversarial Networks. In: Springer (ed.) ECML PKDD 2017. Machine Learning and Knowledge Discovery in Databases, vol. 10535, pp. 175–188 (2017)

[7] Denton, E.L., Chintala, S., Szlam, A., Fergus, R.: Deep generative image models using a laplacian pyramid of adversarial networks. In: Advances in Neural Information Processing Systems 28, pp. 1486–1494 (2015)

[8] Donahue, J., Krähenbühl, P., Darrell, T.: Adversarial feature learning. In: International Conference on Representation Learning (ICLR) (2017)

[9] Dumoulin, V., Belghaz, I., Poole, B., Lamb, A., Arjovsky, M., Mastropietro, O., , Courville, A.: Adversarially learned inference. In: International Conference on Representation Learning (ICLR) (2017)

[10] Farquhar, J., Hardoon, D., Meng, H., Shawe-taylor, J.S., Szedmák, S.: Two view learning: Svm-2k, theory and practice. In: Advances in Neural Information Processing Systems 18, pp. 355–362

[11] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems 27, pp. 2672–2680 (2014)

[12] Goyal, A., Morvant, E., Germain, P., Amini, M.R.: PAC-Bayesian Analysis for a two-step Hierarchical Multiview Learning Approach. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 205–221 (2017)

[13] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: International Conference on Representation Learning (ICLR) (2015)

[14] Ma, L., Jia, X., Sun, Q., Schiele, B., Tuytelaars, T., Van Gool, L.: Pose guided person image generation. In: Advances in Neural Information Processing Systems 30, pp. 406–416 (2017)

[15] Odena, A., Olah, C., Shlens, J.: Conditional image synthesis with auxiliary classifier GANs. In: Proceedings of the 34th International Conference on Machine Learning (ICML). pp. 2642–2651 (2017)

[16] Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. In: International Conference on Representation Learning (ICLR) (2016)

[17] Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., Chen, X.: Improved Techniques for Training GANs. In: Advances in Neural Information Processing Systems 29, pp. 2234–2242 (2016)

[18] Sindhwani, V., Rosenberg, D.S.: An RKHS for multi-view learning and manifold co-regularization. In: Proceedings of the 25^{th} International Conference on Machine Learning (ICML) (2008)

[19] Springenberg, J.T.: Unsupervised and semi-supervised learning with categorical generative adversarial networks. In: International Conference on Representation Learning (ICLR) (2016)

[20] Tian, L., Nie, F., Li, X.: A unified weight learning paradigm for multi-view learning. In: Proceedings of Machine Learning Research. pp. 2790–2800 (2019)

[21] Tian, Y., Peng, X., Zhao, L., Zhang, S., Metaxas, D.N.: CR-GAN: Learning Complete Representations for Multi-view Generation. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI). pp. 942–948 (2018)

[22] Ueffing, N., Simard, M., Larkin, S., Johnson, H.: Nrc's portage system for wmt 2007. In: Proceedings of the Second Workshop on Statistical Machine Translation. pp. 185–188 (2007)

[23] Zhao, B., Wu, X., Cheng, Z., Liu, H., Feng, J.: Multi-view image generation from a single-view. In: Proceedings of the 26th ACM International Conference on Multimedia (MM). pp. 383–391 (2018)