Multimodal Storytelling via Generative Adversarial Imitation Learning

Zhiqian Chen[1], Xuchao Zhang[1], Arnold P. Boedihardjo[2], Jing Dai[3], Chang-Tien Lu[1]

[1]Computer Science Department, Virginia Tech, Falls Church, Virginia

[2]U. S. Army Corps of Engineers

[3]Google Inc.

[1]{czq,xuczhang,ctlu}@vt.edu,[2]arnold.p.boedihardjo@erdc.dren.mil, [3]jddai@google.com

Abstract

Deriving event storylines is an effective summarization method to succinctly organize extensive information, which can significantly alleviate the pain of information overload. The critical challenge is the lack of widely recognized definition of storyline metric. Prior studies have developed various approaches based on different assumptions about users' interests. These works can extract interesting patterns, but their assumptions do not guarantee that the derived patterns will match users' preference. On the other hand, their exclusiveness of single modality source misses cross-modality information. This paper proposes a method, multimodal imitation learning via generative adversarial networks(MIL-GAN), to directly model users' interests as reflected by various data. In particular, the proposed model addresses the critical challenge by imitating users' demonstrated storylines. Our proposed model is designed to learn the reward patterns given user-provided storylines and then applies the learned policy to unseen data. The proposed approach is demonstrated to be capable of acquiring the user's implicit intent and outperforming competing methods by a substantial margin with a user study.

1 Introduction

As the Internet becomes more pervasive, information overload becomes increasingly more severe. Even with the help of search engines such as Google and Yahoo, people cannot easily understand a series of coherent news events. For example, a person who desires to learn about the *2016 Presidential Election* needs iteratively search through several keywords many times and review numerous news documents so that he or she can generate a cohesive picture, e.g., knowledge graph. **Storytelling** is an efficient way to solve this issue of information overload. By inferring the entity nodes connections, the original documents can be represented through a knowledge graph which consists of a set of storylines. Current works [Kumar *et al.*, 2008; Hossain *et al.*, 2012; Shahaf *et al.*, 2012a; 2012b; Lin *et al.*, 2012] mainly focus
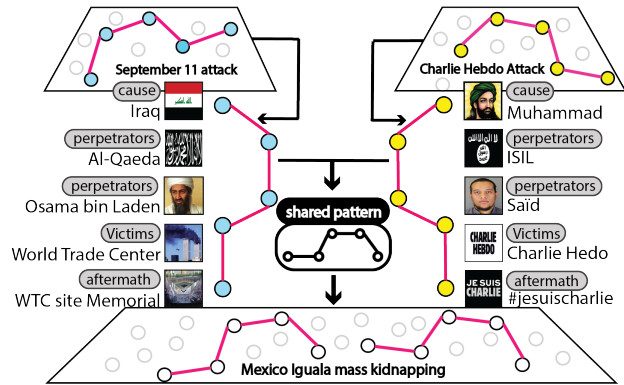

Figure 1: Imitation Storytelling

on this task but employ strict assumptions that may not capture meaningful stories. For instance, [Shahaf and Guestrin, 2010] argues that maximizing the weakest link makes good storylines, which can be reduced to density-based clustering. However, to model users' preferred stories, it requires to understand the evolution patterns, not merely to keep strong coherence.

Specifically, existing research in this area suffers from several shortcomings: **(1) Strong assumptions can lead to poor storylines.** Most related works manually design coherence metrics which are directly assumed to be associated with good storylines. However, a high consistency does not guarantee story quality, since good story might have other properties such interest, novelty, or user-preferred style. Therefore, the coherence based metric is not sufficient for modeling storylines. **(2) Lack of multimodal learning**. Existing works only focus on unimodal data such as Text Storytelling or Visual Storytelling, overlooking the cross-modality information. Multimodal learning can find entity linkages that a unimodal may miss. **(3) Absence of benchmark dataset**: Few prior works are reported to provide a publicly available dataset for imitation storytelling.

This paper focuses on directly imitating user-provided storylines rather than designing any indirect measures. The basic idea is to learn the connectivity features and structure in storylines so that the agent can reveal similar stories in other domains. This approach is illustrated in Figure 1. The *September 11 attack* event contains a storyline which shows the key

entities related to the cause, perpetrators, victims, and the aftermath. Similarly, the event *Charlie Hebdo Attack* also has a story of a similar type. We argue that the two similar storylines share the same structure in a certain embedding space. Therefore, we can reveal similar stories in another event domain(*Mexico Iguala mass kidnapping* at the bottom). The inherent benefit of utilizing multimodal data is that humans often make inferences between the images and texts so as to resolve ambiguities.

Deep reinforcement learning can learn multi-step decisions. However, a critical difficulty in reinforcement learning is designing a reward function for optimizing the agent. Unlike game application in which there exists a responsive environment, it is dramatically difficult to design a reward function for storytelling due to the unavailability of such responses. Instead of proposing reward function, we introduce a typical Inverse Reinforcement Learning(IRL), imitation learning, to learn the latent policy. Imitating from demonstrations is a strategy in which an agent learns a hidden policy for a dynamic environment by observing demonstrations delivered by an expert agent. While IRL is often with two issues: instability and implicit policy. To solve these problems, Generative Adversarial Networks(GAN)[Goodfellow *et al.*, 2014] mechanism is employed to solve the instability issue. GAN yields an internal generator model which can output policy explicitly after training, which addresses the second issue. Therefore, it is promising to integrate IRL with GAN to learn policy from users' demonstrations.

Different from previous work, our study treats storytelling as an imitation learning. Specifically, a policy is acquired from one event domain, and then transfer the policy to learn a storyline from another event. Furthermore, our work takes full advantage of multimodal data to improve the imitation performance. In this paper, we define a Multimodal Imitation Storytelling Task(MIST), and then propose a multimodal generative adversarial method that derives latent policy behind users' demonstrations without explicitly designing a reward function. The main contributions as follows:

- **Proposing an imitation learning method for storytelling**: To avoid the difficulty in designing reward function for storytelling, we enforce generative adversarial model on imitation learning. Using this learning strategy, the model can robustly model latent connectivity patterns.

- **Designing a multimodal model integrated with GAN based imitation learning**: Inspired by human's ability to link multiple entities through visual similarity, we propose a multimodal method across textual and visual modality with imitation learning. Our model learns reward functions from these two modalities and their correlation.

- **Creating a benchmark dataset for multimodal imitation storytelling**: A new multimodal storytelling dataset is collected from multiple attacks and civil unrest events. Under several selected topics, storylines are manually extracted and validated. Both texts and images are included in our dataset.

The rest of this paper is organized as follows. Section 2 reviews the related works. A detailed description of the proposed method is given in Section 3. Experiments on multiple public datasets and case study are presented in Section 4. We conclude the paper in Section 5.

2 Related Works

Storytelling: The storyline generation problem was first studied by Kumar et al. [Kumar *et al.*, 2008] as a generic redescription mining technique, by which a series of redescription between the given disjoint and dissimilar object sets and corresponding subsets are discovered. Storytelling is an efficient way to solve the issue of information overload. By extracting critical and connected entities, the original document is structurally summarized. Current works contain two categories: Textual Storytelling[Kumar *et al.*, 2008; Hossain *et al.*, 2012; Fang *et al.*, 2011; Voskarides *et al.*, 2015; Lee *et al.*, 2012; Shahaf *et al.*, 2012a; 2012b; Lin *et al.*, 2012] and Visual Storytelling[Kim *et al.*, 2014; Park and Kim, 2015; Wang *et al.*, 2012]. Few works are reported to extract storylines based on both text and image. Current methods often suggest assumptions between good storyline and explicit metrics, such as average similarity or weakest similarity of all the neighbor nodes. However, these assumptions limit the generating meaningful stories since a user may have unique notions of good storylines. A few researchers employ Latent Dirichlet Allocation(LDA)[Zhou *et al.*, 2015; Huang and Huang, 2013] to extract stories in unsupervised fashion. However, it is difficult for LDA to accurately model sequential data.

Reinforcement Learning: Starting from AlphaGo [Silver *et al.*, 2016] and Atari[Mnih *et al.*, 2015], numerous game applications [Lample and Chaplot, 2016; He *et al.*, 2016; Oh *et al.*, 2015; Narasimhan *et al.*, 2016] enjoy the property of deep reinforcement learning in imitating sequential patterns. However, it is often difficult to design rewards function, especially for this real world problem. Other possible solutions include behavior cloning[Pomerleau, 1991], and Inverse Reinforcement Learning(IRL)[Russell, 1998; Ng *et al.*, 2000] which can derive the underlying cost function under which the expert data is uniquely optimal. Unfortunately, behavior cloning learns a policy as a supervised learning problem over state-action pairs from expert trajectories and required large amounts of data because of compounding error caused by covariate shift. While IRL has stability issue and does not explicitly tell us how to act. Several works[Yu *et al.*, 2017; Ho and Ermon, 2016] employ Generative Adversarial Networks(GAN) solve IRL issue. Hence we leverage GAN and IRL to effectively imitate user-provided storylines.

Multimodal Learning is to derive the patterns in associated cross-modality information. An example is that readers often estimate the relationship among news articles using the top images inside because if the images contain the same entity, they are probably involved in the same event. However, current works[Kiros *et al.*, 2015][Srivastava and Salakhutdinov, 2012][Ngiam *et al.*, 2011] focus on joint representation learning, paying little attention to sequence problem.

Different from the previous study, our work treats storytelling as a combination problem of imitation learning and

multimodal learning. By employing GAN based imitation learning, our proposed model can learn and show the hidden policy. Moreover, this work takes full advantage of joint constraint on cross-modality data to improve the imitation performance.

3 Multimodal Imitation Storytelling

This section formally defines the task of imitation storytelling and then describes our proposed MIL-GAN model. In particular, a multimodal and GAN based imitation learning is elaborated in 3.2, the multimodal method we applied is introduced in 3.3.

3.1 Problem Setup

Based on a set of event document-storyline pairs $\{\mathcal{D}, \mathcal{S}\}$, our goal is to reveal the user generated policy $\pi^{G_{user}}$. Each $\mathcal{D}$ is a documents collection and each $\mathcal{S}$ consist of several entity nodes $\{e\}$

Let $\mathcal{M} = \{txt, img, mm\}$ be the types of modalities appearing in $\mathcal{D}$ and $\mathcal{S}$, which denote textual data, image and multimodal relationship between respectively. In the proposed method, a generator $\pi^{G_{\theta}}$ controlled by multimodal parameters $\boldsymbol{\theta} = \{\theta^{txt}, \theta^{img}, \theta^{mm}\}$ is used to approximate the users' policy generator $\pi^{G_{user}}$.

Following REINFORCE algorithm[Williams, 1992], the collected rewards R_T along the sequence starting from the initial state $\boldsymbol{e_0} = \{e^{txt}, e^{img}, e^{mm}\}$ are maximized so as to optimize the generator. Utilizing GAN model, a discriminator $\mathcal{D}_{\phi}$ yields rewards as the state score of the proposed generator. Since the Q value function is influenced by both the generator and discriminator, let $Q_{D_{\phi}}^{G_{\theta}}$ denote the Q function. The learning process iteratively updates the objective function via gradient w.r.t. the parameters $\boldsymbol{\theta}$ until convergence.

3.2 Imitation Learning via GAN

The intermediate rewards are set to zero for storytelling task because it's difficult to evaluate the generated storyline until it's complete. Following [Sutton *et al.*,],the objective of policy $\pi\left(s_t | \mathcal{S}_{1:t-1}\right)$ is to generate a sequence starting from $\boldsymbol{s_0}$ to maximize the expected reward:

$$J\left(\theta\right) = \mathbb{E}\left[\boldsymbol{R_T} | \boldsymbol{e_0}, \boldsymbol{\theta}\right] \qquad (1)$$

Because multimodal constraints involves several separate relationship embedded in text, image and their bi-directional coherence, the objective function is formulated as:

$$
\begin{aligned}
J\left(\boldsymbol{\theta}\right) &= \mathbb{E}\left[\sum_{i \in \mathcal{M}} \lambda^i R_T^i \middle| \boldsymbol{e_0}, \boldsymbol{\theta}\right] = \sum_{i \in \mathcal{M}} \lambda^i \, \mathbb{E}^i\left[R_T^i \middle| e_0^i, \theta^i\right] \\
&= \sum_{i \in \mathcal{M}} \lambda^i \sum_{s_1 \in \mathcal{S}^i} \pi^{G_{\theta^i}}\left(s_1 | e_0^{txt}\right) \cdot Q_{D_{\phi}^i}^{G_{\theta^i}}\left(e_0^i, s_1\right) \quad (2)
\end{aligned}
$$

where $Q_{D_{\phi}}^{G_{\theta}}\left(state, action\right)$ is the action-value function that estimates the expected accumulative reward from initial state e_0. The Q value is from the rating on sequence generated by $\pi^{G_{\theta}}$. When the discriminator consider the generated sequences real, the Q value will increase. The basic
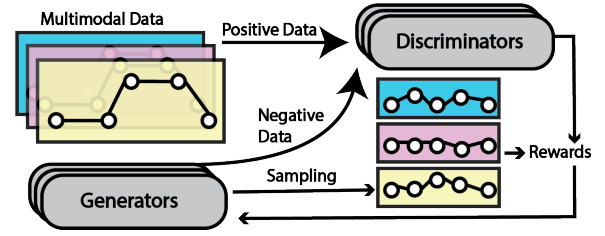

Figure 2: Illustration of Multimodal GAN: Large arrows show the data flow in GAN training, while small arrows indicate the process inside policy gradient

idea of estimating Q value is to treat it as the probability of the sequence being real considered by the discriminator, i.e., $Q_{D_{\phi}}^{G_{\theta}} = D_{\phi}$. Due to the model limitation, only the sequence of fixed length can be evaluated by discriminator. However, incomplete sequence does provide partial information for rating. To estimate this incomplete portion of the sequence, the action-value function value is assigned with the expected rewards obtained from LSTM sampling controlled by $\pi^{G_{\theta}}$. Therefore, the $Q_{D_{\phi}}^{G_{\theta}}\left(\mathcal{S}_{1:t-1}, s_t\right)$ value is estimated as:

$$
\begin{cases}
D_{\phi}\left(\mathcal{S}\right) & \text{if } len(\mathcal{S}) = T \\
\mathbb{E}_{\pi^{G_{\theta}}}\left[D_{\phi}\left(LSTM(\mathcal{S}, \pi^{G_{\theta}})\right)\right] & else
\end{cases} \quad (3)
$$

By using such Q values, the generator can update the objective function w.r.t. its parameters. Once generator finishes update, the sampling sequences mixed with the real data are fed into discriminator. Following the typical rules of GAN and an improvement[Arjovsky *et al.*, 2017] that the original log operation of discriminator is removed, the discriminator will then improve the function:

$$\min_{\phi} -\mathbb{E}_{Y \sim P_{data}}\left[D_{\phi}\left(\mathcal{S}\right)\right] - \mathbb{E}_{Y \sim \pi^{G_{\theta}}}\left[\left(1 - D_{\phi}\left(\mathcal{S}\right)\right)\right] \quad (4)$$

As illustrated in Figure 2, the GAN iteratively updates the multimodal generators using Formula 3, and improves the multimodal discriminators by Formula 4. Following policy gradient theorem(episodic case) [Sutton *et al.*,], the gradient of the objective function $J(\theta)$ w.r.t. the parameters θ can be expressed as:

Theorem 3.1 (Multimodal Policy Gradient Theorem). *The gradient of the objective function $\nabla_{\theta} J\left(\boldsymbol{\theta}\right)$ w.r.t. the parameter $\boldsymbol{\theta}$ via policy gradient is:*

$$\sum_{i \in \mathcal{M}} \lambda^i \cdot \mathbb{E}_{G_{\theta_i}}\left[\sum_{s_t \in \mathcal{S}_i} \nabla_{\theta^i} \pi^{G_{\theta_i}}\left(s_t | \mathcal{S}_{1:t-1}\right) \cdot Q_{D_{\phi^i}}^{G_{\theta^i}}\left(\mathcal{S}_{1:t-1}, s_t\right)\right]$$

Proof of Theorem 3.1. Let v denote the state value function. To keep the notation simple, we leave it implicit in all cases that v, π and Q are controlled by the generator G^{θ_i}. For each single modality, due to that intermediate rewards are set to zero, and state transferring probability γ is one-hot deterministic, we have:

$$Q\left(\mathcal{S}_{1:t-1}, s_t\right) = 0 + \sum_{s' \in \mathcal{S}} \gamma_{ss'} v\left(s'\right) = v\left(\mathcal{S}_{1:t}\right) \quad (5)$$

$$v\left(\mathcal{S}_{1:t-1}\right) = \sum_{s_i \in \mathcal{S}} \pi\left(s_t | \mathcal{S}_{1:t-1}\right) Q\left(\mathcal{S}_{1:t-1}, s_t\right) \quad (6)$$

Sum over modalities $\mathcal{M}$ on Eq. (6) is exactly Eq. (2), i.e., $\nabla_\theta J = \sum_{i \in \mathcal{M}} \lambda^i \cdot \nabla_{\theta^i} v(s_0)$. Substitute v in Eq. (6) and Q in (5) iteratively:

$$\nabla_\theta J = \sum_{i \in \mathcal{M}} \lambda^i \cdot \nabla_{\theta^i} v(s_0)$$

$$= \sum_{i \in \mathcal{M}} \lambda^i \cdot \nabla_{\theta^i} \left[\sum_{s_1 \in \mathcal{S}} \pi(s_1|s_0) Q(s_0, s_1) \right]$$

$$= \sum_{i \in \mathcal{M}} \lambda^i \sum_{s_1 \in \mathcal{S}} [\nabla_\theta \pi(s_1|s_0) Q(s_1|s_0) + \pi(s_1|s_0) \nabla_\theta Q(s_0, s_1)]$$

$$= \sum_{i \in \mathcal{M}} \lambda^i \sum_{s_1 \in \mathcal{S}} [\nabla_\theta \pi(s_1|s_0) Q(s_1|s_0) + \pi(s_1|s_0) \nabla_\theta v(\mathcal{S}_{1:1})]$$

$$= \sum_{i \in \mathcal{M}} \lambda^i \sum_{s_1 \in \mathcal{S}} \Big[\nabla_\theta \pi(s_1|s_0) Q(s_1|s_0) + \pi(s_1|s_0) \nabla_{\theta^i} \big[\sum_{s_1 \in \mathcal{S}} \pi(s_1|s_0) Q(s_0, s_1) \big] \Big]$$

$$= \sum_{i \in \mathcal{M}} \lambda^i \sum_{s_1 \in \mathcal{S}} \Big[\nabla_\theta \pi(s_1|s_0) Q(s_1|s_0) + \pi(s_1|s_0) \sum_{s_1 \in \mathcal{S}} [\nabla_\theta \pi(s_1|s_0) Q(s_1|s_0) + \pi(s_1|s_0) \nabla_\theta Q(s_0, s_1)] \Big]$$

We can re-write accumulative policy output as probability:

$$Pr(\mathcal{S}_{1:t}|s_0) = \prod_i^t \pi(s_i|\mathcal{S}_{1:i-1}) = \pi(s_1|s_0)\pi(s_2|\mathcal{S}_{1:1})..\pi(s_t|\mathcal{S}_{1:t-1}) \quad (7)$$

After repeating unrolling and apply Eq. (7), it is then immediate that:

$$\nabla_\theta J = \sum_{i \in \mathcal{M}} \lambda^i \sum_{s_1 \in \mathcal{S}} \Big[\nabla_\theta \pi(s_1|s_0) Q(s_1|s_0)$$
$$+ Pr(\mathcal{S}_{1:1}|s_0) \sum_{s_1 \in \mathcal{S}} \nabla_\theta \pi(s_1|s_0) Q(s_1|s_0)$$
$$+ \pi(\mathcal{S}_{1:2}|s_0) \nabla_\theta Q(s_0, s_1) \Big]$$

$$= \sum_{i \in \mathcal{M}} \lambda^i \Big[\sum_{t=1}^{T} \sum_{\mathcal{S}_{1:t-1}} Pr(\mathcal{S}_{1:t-1}|s_0; \pi) \sum_{s_t \in \mathcal{S}} \nabla_\theta \pi(s_t|\mathcal{S}_{1:t-1}) Q(\mathcal{S}_{1:t-1}, s_t)$$
$$+ \sum_{s_t \in \mathcal{S}} \pi(s_t|\mathcal{S}_{1:t-1}) \nabla_\theta Q(\mathcal{S}_{1:t-1}, s_t) \Big]$$

In storytelling, each node can only be assigned with one entity, the policy function in the last term given a specific s_t reach $\lim_{T \to \infty} \pi(s_t|\mathcal{S}_{1:t-1}) = \inf \pi(s_t|\mathcal{S}_{1:t-1}) = 0$. Accordingly, the above equation becomes:

$$\nabla_\theta J = \sum_{i \in \mathcal{M}} \lambda^i \sum_{t=1}^{T} \sum_{\mathcal{S}_{1:t-1}} Pr(\mathcal{S}_{1:t-1}|s_0; \pi) \sum_{s \in \mathcal{S}} \nabla_\theta \pi(s_t|\mathcal{S}_{1:t-1}) Q(\mathcal{S}_{1:t-1}, s)$$

$$= \sum_{i \in \mathcal{M}} \lambda^i \cdot \mathbb{E}_{G_{\theta_i}} \left[\sum_{s_t \in \mathcal{S}_i} \nabla_{\theta^i} \pi^{G_{\theta_i}}(s_t|\mathcal{S}_{1:t-1}) \cdot Q_{D_{\phi^i}}^{G_{\theta^i}}(\mathcal{S}_{1:t-1}, s_t) \right]$$

<div align="right">QED</div>

Using weighted likelihood [Sutton *et al.*,], the expectation inside Formula 3.1 can be decomposed using unbiased estimation. To keep the notation simple, we leave it implicit in all cases that π is a function of s_t given $\mathcal{S}_{1:t-1}$, and $Q_{D_\phi}^{G_\theta}$ is a function of s_t and $\mathcal{S}_{1:t-1}$. Thus, the gradient can be estimated

as:

$$\nabla_\theta J(\boldsymbol{\theta}) \simeq \sum_{i \in \mathcal{M}} \lambda^i \cdot \frac{1}{T} \sum_{t=1}^{T} \sum_{s_t \in \mathcal{S}} \nabla_{\theta^i} G_{\theta^i} \cdot Q_{D_{\phi^i}}^{G_{\theta^i}}$$

$$= \sum_{i \in \mathcal{M}} \frac{\lambda^i}{T} \sum_{t=1}^{T} \sum_{s_t \in \mathcal{S}} G_{\theta^i} \nabla_{\theta^i} \log G_{\theta^i} \cdot Q_{D_{\phi^i}}^{G_{\theta^i}}$$

$$= \sum_{i \in \mathcal{M}} \frac{\lambda^i}{T} \sum_{t=1}^{T} \mathbb{E}_{s_t \sim G_{\theta^i}} \left[\nabla_{\theta^i} \log G_{\theta^i} \cdot Q_{D_{\phi^i}}^{G_{\theta^i}} \right]$$

The gradient can then be applied on the previous parameters as: $\theta^i \leftarrow \theta^i + \alpha \nabla_{\theta^i} J(\theta^i)$, where α is learning rate. Algorithm 1 presents full details of the proposed method. First, we pre-train G^i on input set $\mathcal{S}$. Then, the generator and discriminator are trained alternatively and periodically. When training the discriminator, positive examples are from the given dataset, while negative examples are sampled from the generator.

Algorithm 1: Multimodal Imitation Storytelling

Input: initialize generator policy π^{G_θ} and discriminator D_ϕ with random weights; multimodal storyline dataset $Data = \{E(ntity)_{1:T}, I(mg)_{1:T}\}$; event documents as knowledge base $\mathcal{D}$

Output: Generator π^{G_θ} and discriminator D_ϕ; Newly-generated storyline sequences by π^{G_θ} on unseen test dataset.

1   // Pre-processing
2   Derive embeddings V_e on $\mathcal{D}$ and map entity E in V_e;
3   Compute representations V_i of based on $\mathcal{D}$ using multimodal learning, and map each of $I_{1:T}$ to V_i;
4   Generate the sequence $\mathcal{S}$ including V_e, V_i and $V_i - V_e$;
5   // Pre-training
6   Train π^{G_θ} using LSTM on $\mathcal{S}$;
7   Generate negative samples using π^{G_θ} for training D_ϕ;
8   Feed D_ϕ with negative samples and real data;
9   Train D_ϕ via minimizing the cross entropy;
10   // GAN training
11   **repeat**
12   **for** π^{G_θ} *training* **do**
13   Sample a sequence $\mathcal{S}_{1:T}^{\pi^{G_\theta}} = (s_1, ..., s_T) \sim \pi^{G_\theta}$;
14   **for** *t in 1:T* **do**
15   Derive $Q(s_s; \mathcal{S}_{1:t-1})$ using Eq. (3)
16   **end**
17   Update generator: $\theta^i \leftarrow \theta + \alpha \nabla_\theta J(\theta)$
18   **end**
19   **for** D_ϕ *training* **do**
20   Generate negative examples by sampling $\mathcal{S}_{1:T}^{\pi^{G_\theta}} = (s_1, ..., s_T) \sim \pi^{G_\theta}$;
21   Use given positive examples $\mathcal{S}$;
22   Train discriminator D_ϕ by Eq. (4)
23   **end**
24   **until** *GAN converges*;
25   Apply π^{G_θ} on unseen dataset $Data'$ starting from randomly selected entity $E' \in Data'$

3.3 Multimodal Learning

The task contains three types of modalities which are text, image, and their coherence. Entity words are encoded using word embeddings algorithm such as Word2Vec. Whereas our model expresses images in a semantic sentential space instead of word space since images contain massive information.

A SVD-based semantic embedding model is employed to derive vectors for images conditioning on contextual words. Denote the image vector $\mathcal{I} \in M_k(\mathbb{R})$ from a multimodal vector space and is associated with sentential description $\mathcal{SEN} = \{w_1, ..., w_N\}$ where w_i indicates words. To represent the vectors $\mathcal{I}$ is to condition on the embedding vector of the description $\mathcal{SEN}$. We aim to model the conditional distribution $Pr(w_n|w_{1:n-1}, \mathcal{I})$ of the following word w_n given context from the contextual words and the vector $\mathcal{I}$. Vocabulary embedding matrix with associated image vectors are represented using a tensor $\Psi = \mathcal{I}_k \otimes \mathcal{SEN}_{v \times k} \in M_{v \times k \times k}(\mathbb{R})$ where $\otimes$ mean tensor product in which any word in description is associated with related images. Given $\mathcal{I}$, the model predicts word representation of the following words as a function of $\mathcal{I}$ and contextual words using tensor decomposition:

$$\Psi(\mathcal{I}) = U_{jv}^T \cdot \Sigma(\sigma_{jk}, \mathcal{I}) \cdot V_{jk} \tag{8}$$

where the dimensionality of j is tunable. Σ denote a function which retain all the arguments on the diagonal after multiplication. This idea share resemblance with SVD in which the middle matrix is a diagonal matrix. The context is then represented as an intermediate variable which subject to the expectation of weights distribution among word pairs : $C = \mathbb{E}_{v \times v} \left[U_{jk}^T V_{fv}(w_i) \right]$. Combining with I, we get another intermediate factor which encode context and image using Hadamard product $\psi = (UC) \cdot (V\mathcal{I})$. Using with softmax, the conditional probability Pr can be calculated. The true following words are used as true label for backpropagation method such as SGD, this neural networks iteratively update w.r.t. the parameters. The training parameters include $U_{jv}, \sigma_{jk}, V_{jk}$, and weights distribution $\mathbb{E}_{vv}$. The this model encodes image in a vector space V_i. Finally, the third modality is the sequential difference between word embeddings and image vectors.

4 Benchmark and Experiment Setting

The proposed method is evaluated on newly-proposed storytelling dataset[1]. To guide the model to discover desirable stories, manually labeled storylines are compiled for GAN training. Generator obtained in one event dataset was tested on another event corpus. This experiment shows if the generator is capable of deriving transferable storyline. Please note that different event datasets share no entities.

4.1 Benchmark Description and Metrics

Training set contain events from two major categories: *Homicide* and *Protest*. *Homicide* contains 1310 storylines, while *Protest* chooses 934 storylines. Two short examples

Table 1: Similarity performance (T./I. denotes Text/Image respectively, T.I. means the combination of T. and I.)

Ran.	SS	PG	LSTM
-3860.02	33714.67	34009.27	338876.34
T.	I.	T.I.	MIL-GAN
34263.59	12483.20	36143.34	**36697.33**

are shown in Figure 1. They are from the most famous historical events, such as *9/11 Attack* , *Orlando Nightclub Attack*, *Occupy Wall Street*, and *Protest led by Martin Luther King*. Each event contains several hundred documents. All the articles are taken from Google news and Wikipedia. Generators will be trained in the two categories separately. The **Test set** includes two event *2014 Iguala mass kidnapping*[2] and Malaysia Airlines Flight 370[3], since the two test events involve both *homicide* and *protest* sub-events. For data augmentation purpose, slicing window is employed to divide raw data into minimal sequence.

Metrics: First the proposed method with several baselines are tested under convergence performance. Secondly, our evaluation conducts user studies via Amazon Mechanical Turk(AMT) cloud sourcing service, since the ultimate goal is to validate imitation behaviors.

4.2 Training Setting

Words are expressed using Word2Vec in each independent event corpus, while images are initially represented in VGG19 and then transferred to sentential space using multimodal learning. To normalize their shape, word vectors along each storyline are reduced by the vector of the first word in its storyline. Likewise, image vectors are reduced by the first image in each storyline. The same operation is also conducted on the multimodal sequence. For GAN model, the generator is implemented using LSTM model which accept continuous values. TextCNN[Zhang and LeCun, 2015] is used for discriminator since such CNN is effective for both text and image. Baselines include Random(**Ran.**), Scheduled sampling(**SS**)[Bengio *et al.*, 2015], policy gradient(**PG**), and **LSTM**.

4.3 Initial Analysis

We evaluate our result with several established alternatives: random, scheduled sampling[Bengio *et al.*, 2015] and policy gradient with similarity. Unfortunately, the baselines do not share the same metric or objective function. Instead, they were compared regarding accumulative normalized similarity on the training set:

$$sum_{sim} = \sum_{\mathcal{A} \in Data} \sum_{\mathcal{B} \in model} \frac{\mathcal{A}\mathcal{B}}{\|\mathcal{A}\|_F \cdot \|\mathcal{B}\|_F}$$

where, sample $\mathcal{A}$ denotes sequence matrix from the user-provided data, and $\mathcal{B}$ is the output sequence matrix from models. The result is shown in Table (1).

---

[1] https://gist.github.com/aquastar/03dadfd751f5862ea0b44bb66996b490

[2] https://en.wikipedia.org/wiki/2014_Iguala_mass_kidnapping
[3] https://en.wikipedia.org/wiki/Malaysia_Airlines_Flight_370

Using the multimodal GAN, we improve the text-only model. One interesting point is that the image-only modality does not performance very well, but with other modalities, the performance increases. This makes sense because an image may contain too much information which is likely to lead to confusion. For instance, an image of original World Trade Center being attached could be either a *victim* object(the building) or an *attacker* (the attacking flight). With textual label such as victim, the image would be confined in a smaller and accurate semantic space.

The balance parameters $\lambda_{i=1,2,3}$ are all initialized to 1. After fine tuning, good performance often appear if more weights were assigned on text part. One good set example is [0.6, 0.3, 0.1] for $[V_e, V_i, V_i - V_e]$ separately.

However, result of the training set can only show that the potential ability to achieve success, but it is not the ultimate evaluation. In next section, we conducted user study via Amazon Mechanical Turk(AMT) Service to directly assess users' satisfaction.

4.4 Evaluation via user study

Evaluating storytelling is a difficult task, due to the fact that there is no established golden standard, and even ground truth is hard to elicit. The ultimate goal of imitation storytelling is to help users discover customized storylines given users' examples. Therefore, a third-party user study via AMT is conducted, since it allows us to obtain accurate statistical information with a large sample size of users. It aims to test whether the generated storyline matches users' interests. We evaluate our result with the baselines mentioned in the last subsection.

In the study, workers were asked which storyline fit the given storyline best. Due to our dynamic reward per assignment in AMT, every minimal unit of cost is treated as one evaluation, so 1000 rating were collected. For each test event, several starting entities were randomly chosen. Storylines were generated from these starting entities using both our proposed method and baselines. For each unit task in AMT, workers were asked to choose the best one among those generated candidates gave event background knowledge such as Wikipedia or key news articles:

- **1st Step**: AMT will be given E0, which denotes an event background and user-provided storyline S0 which is a user-define storyline. S0 contain a topic which is based on E0.

- **2nd Step**: After AMT workers study the relationship between the storyline and event, they will be given another event E1 and a few machine generated storyline S1/S2/S3/S4. They are also related.

- **3rd Step**: Based on their understanding of the relationship S0-E0, choose the best one that would have similar relationship: E1-S[?] among (S1,S2,S3,S4)

The final result as shown in Table 2 demonstrates that our proposed MIL-GAN was significantly preferred by users. T. slightly improve the preference beyond Schedule Sampling. Policy Gradient fails to capture the users' behavior.

Table 2: Preference statistics from AMT

PG	SS	T.	MIL-GAN
17.3%	23.2%	**28.1%**	**31.4%**

4.5 case analysis

In this subsection, several typical examples are presented, and detailed analysis will be elaborated.

case 1: Mexico Murderer
- **SS**: Abarca→Flipe→Flipe→Guerrero
- **PG**: Abarca→Pineda→student→student
- **T.**: Abarca→gang→student→Guerrero
- **MIL-GAN**: Abarca→Flipe→Ayotzinapa→Guerrero

case 1: Mexico Drug Cartel
- **SS**: Cartel→Flipe→Flipe→Guerrero
- **PG**: Cartel→gang→gang→Sinaloa
- **T.**: Cartel→Drug→Pineda→gang
- **MIL-GAN**: Cartel→Pineda→Abarca→Flipe

case 2: MH370
- **SS**: Zaharie→wife→wife→cockpit
- **PG**: Zaharie→Flight→Fariq→training
- **T.**: Zaharie→Flight→Aisa→Search
- **MIL-GAN**: Zaharie→MH370→India→debris

We found that **(1)**The generated sequences from the baselines **SS** and **PG** often contains more vague words such as *student* and *gang*. These two baselines tend to overfit and often yield repeat nodes as in case 1 and 2. This suggests that the models without GAN are likely to overfit. **(2)** With GAN, the proposed models seems succeeded to avoid the overfit issue. Compared with **SS** and **PG**, **T.** and **MIL-GAN** tend to generate specific word rather than ambiguous ones. For example, in case 1, 2 and 3, **T.** and **MIL-GAN** generated specific words such as *Aisa* and *debris*. Compared with **T.**, **MIL-GAN** generates even more specific entity name such as *Ayotzinapa*, *Flipe* and *India*. This implies that multimodal constraints does improve the unimodal performance.

5 Conclusion

In this paper, we proposed a multimodal imitation learning approach for generating storyline on unseen events. To avoid the reward function designing, GAN based imitation learning is introduced to learn the latent policy given users' demonstrations. To bridge the information gap between text and image, our model effectively integrates generative adversarial nets and multimodal learning via deterministic policy gradient. The different modalities learn from each other and potentially resolve confusion from each single modality. In our experiments, we utilized user-provided demonstration to explicitly illustrate the advantage of the proposed method beyond baselines. Associating with multimodal perspective, our model succeeds to capture the latent patterns across different modalities, and therefore reveal more satisfying storylines towards users' interests.

References

[Arjovsky *et al.*, 2017] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein GAN. *ArXiv e-prints*, jan 2017.

[Bengio *et al.*, 2015] Samy Bengio, Oriol Vinyals, et al. Scheduled sampling for sequence prediction with recurrent neural networks. In *NIPS*, pages 1171–1179, 2015.

[Fang *et al.*, 2011] Lujun Fang, Anish Das Sarma, Cong Yu, and Philip Bohannon. Rex: explaining relationships between entity pairs. *VLDB*, 5(3):241–252, 2011.

[Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Yoshua Bengio, et al. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014.

[He *et al.*, 2016] Ji He, Mari Ostendorf, et al. Deep reinforcement learning with a combinatorial action space for predicting popular reddit threads. *EMNLP*, 2016.

[Ho and Ermon, 2016] Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In *NIPS*, pages 4565–4573, 2016.

[Hossain *et al.*, 2012] M Shahriar Hossain, Patrick Butler, Arnold P Boedihardjo, and Naren Ramakrishnan. Storytelling in entity networks to support intelligence analysts. In *SIGKDD*, pages 1375–1383. ACM, 2012.

[Huang and Huang, 2013] Lifu Huang and Lian'en Huang. Optimized event storyline generation based on mixture-event-aspect model. In *EMNLP*, pages 726–735, 2013.

[Kim *et al.*, 2014] Gunhee Kim, Leonid Sigal, et al. Joint summarization of large-scale collections of web images and videos for storyline reconstruction. In *CVPR*, pages 4225–4232, 2014.

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

[Kumar *et al.*, 2008] Deept Kumar, Naren Ramakrishnan, Richard F Helm, and Malcolm Potts. Algorithms for storytelling. *KDD*, 20(6):736–751, 2008.

[Lample and Chaplot, 2016] Guillaume Lample and Devendra Singh Chaplot. Playing fps games with deep reinforcement learning. *AAAI*, 2016.

[Lee *et al.*, 2012] Heeyoung Lee, Marta Recasens, et al. Joint entity and event coreference resolution across documents. In *EMNLP-CONLL*, pages 489–500. Association for Computational Linguistics, 2012.

[Lin *et al.*, 2012] Chen Lin, Chun Lin, et al. Generating event storylines from microblogs. In *CIKM*, pages 175–184. ACM, 2012.

[Mnih *et al.*, 2015] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

[Narasimhan *et al.*, 2016] Karthik Narasimhan, Adam Yala, et al. Improving information extraction by acquiring external evidence with reinforcement learning. *EMNLP*, 2016.

[Ng *et al.*, 2000] Andrew Y Ng, Stuart J Russell, et al. Algorithms for inverse reinforcement learning. In *ICML*, pages 663–670, 2000.

[Ngiam *et al.*, 2011] Jiquan Ngiam, Aditya Khosla, et al. Multimodal deep learning. In *ICML*, pages 689–696, 2011.

[Oh *et al.*, 2015] Junhyuk Oh, Xiaoxiao Guo, Honglak Lee, et al. Action-conditional video prediction using deep networks in atari games. In *NIPS*, pages 2863–2871, 2015.

[Park and Kim, 2015] Cesc C Park and Gunhee Kim. Expressing an image stream with a sequence of natural sentences. In *NIPS*, pages 73–81, 2015.

[Pomerleau, 1991] Dean A Pomerleau. Efficient training of artificial neural networks for autonomous navigation. *Neural Computation*, 3(1):88–97, 1991.

[Russell, 1998] Stuart Russell. Learning agents for uncertain environments. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 101–103. ACM, 1998.

[Shahaf and Guestrin, 2010] Dafna Shahaf and Carlos Guestrin. Connecting the dots between news articles. In *SIGKDD*, pages 623–632. ACM, 2010.

[Shahaf *et al.*, 2012a] Dafna Shahaf, Carlos Guestrin, and Eric Horvitz. Metro maps of science. In *SIGKDD*, pages 1122–1130. ACM, 2012.

[Shahaf *et al.*, 2012b] Dafna Shahaf, Carlos Guestrin, and Eric Horvitz. Trains of thought: Generating information maps. In *WWW*, pages 899–908. ACM, 2012.

[Silver *et al.*, 2016] David Silver, Aja Huang, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

[Srivastava and Salakhutdinov, 2012] Nitish Srivastava and Ruslan R Salakhutdinov. Multimodal learning with deep boltzmann machines. In *NIPS*, pages 2222–2230, 2012.

[Sutton *et al.*,] Richard S Sutton, David A McAllester, Satinder P Singh, Yishay Mansour, et al. Policy gradient methods for reinforcement learning with function approximation.

[Voskarides *et al.*, 2015] Nikos Voskarides, Edgar Meij, et al. Learning to explain entity relationships in knowledge graphs. 2015.

[Wang *et al.*, 2012] Dingding Wang, Tao Li, and Mitsunori Ogihara. Generating pictorial storylines via minimum-weight connected dominating set approximation in multi-view graphs. In *AAAI*. Citeseer, 2012.

[Williams, 1992] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

[Yu *et al.*, 2017] L Yu, W Zhang, et al. Seqgan: sequence generative adversarial nets with policy gradient. volume 31. AAAI, 2017.

[Zhang and LeCun, 2015] Xiang Zhang and Yann LeCun. Text understanding from scratch. *arXiv preprint arXiv:1502.01710*, 2015.

[Zhou *et al.*, 2015] Deyu Zhou, Liangyu Chen, et al. An unsupervised framework of exploring events on twitter: Filtering, extraction and categorization. In *AAAI*, 2015.