(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
03 December 2020 (03.12.2020)

WIPO | PCT

(10) International Publication Number
WO 2020/239208 A1

(51) International Patent Classification:
G06T 11/00 (2006.01)

(21) International Application Number:
PCT/EP2019/063853

(22) International Filing Date:
28 May 2019 (28.05.2019)

(25) Filing Language: English

(26) Publication Language: English

(71) Applicants: **TOYOTA MOTOR EUROPE** [BE/BE]; Avenue du Bourget 60, 1140 BRUSSELS (BE). **MAX-PLANCK-INSTITUT FÜR INFORMATIK** [DE/DE]; Saarland Informatics Campus, Campus E1 4, 66123 SAARBRÜCKEN (DE).

(72) Inventors: **OLMEDA REINO, Daniel**; c/o TOYOTA MOTOR EUROPE, European Technical Administration Division, Avenue du Bourget 60, 1140 BRUSSELS (BE). **BHATTACHARYYA, Apratim**; Zum Bartenberg 1, Dudweiler, 66125 SAARBRÜCKEN (DE). **FRITZ, Mario**; Gustav-Bruch-Straße 10, 66123 SAARBRÜCKEN (DE). **SCHIELE, Bernt**; St.Ingberter Str. 39, 66125 SAARBRÜCKEN (DE).

(74) Agent: **HEWEL, Christoph** et al.; CABINET BEAU DE LOMENIE, 158 Rue de l'Université, 75340 PARIS CEDEX 07 (FR).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report (Art. 21(3))*

(54) Title: METHOD AND SYSTEM FOR TRAINING A MODEL FOR IMAGE GENERATION

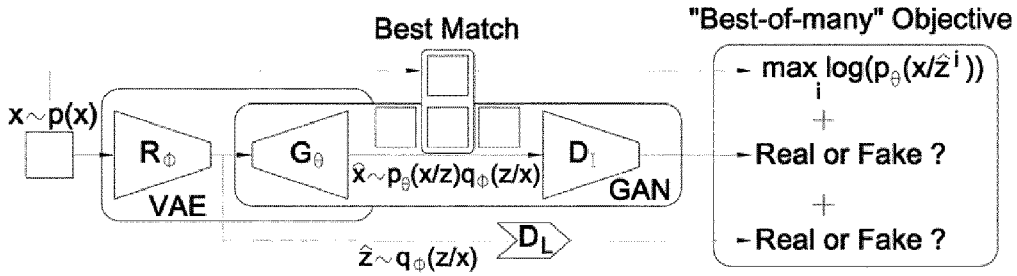

Fig. 3

(57) Abstract: The invention relates to a method and system for training a model for image generation. The model comprises a hybrid variational auto-encoder (VAE) - generative adversarial network (GAN) framework. The method comprises the steps of: a - multiple input (S01) of an input image into the VAE which outputs in response multiple distinct output image samples, b - determine (S02) the best of the multiple output image samples as a best- of-many sample, the best-of-many sample having the minimum reconstruction cost, c - train (S03) the model based on a predefined training objective, the predefined training objective integrating the best-of-many sample reconstruction cost and a GAN-based synthetic likelihood term.

WO 2020/239208 A1

Method and system for training a model for image generation

FIELD OF THE DISCLOSURE

[0001] The present disclosure is related to the field of image processing, in particular to a method for training a model for image generation, the model comprising a hybrid variational auto-encoder (VAE) – generative adversarial network (GAN) framework.

BACKGROUND OF THE DISCLOSURE

[0002] Generative Adversarial Networks (GANs) have achieved state-of-the-art performance, with respect to realism, in generative modeling of image distributions, cf.:

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozairy, Aaron Courville, Yoshua Bengioz (2014) "Generative Adversarial Nets", Advances in neural information processing systems, Pages 2672-2680.

[0003] GANs do not explicitly estimate the data likelihood. Instead, it aims to "fool" an adversary, so that the adversary is unable to distinguish between images from the true distribution and the generated images. This leads to the generation of very realistic images. However, there is no incentive to cover the whole data distribution. Entire modes of the true data distribution can be missed – commonly referred to as the mode collapse problem.

[0004] In contrast, auto-encoders explicitly maximize data log-likelihood and are forced to cover all modes. However, auto-encoder latent distributions are discontinuous and hard to estimate and thus do not allow for sampling. Variational Auto-encoders (VAEs) enable generation using auto-encoders by constraining the latent space to be Gaussian, cf.:

D. P. Kingma and M. Welling. Auto-encoding variational bayes. ICLR, 2014.

[0005] This allows for generation using the decoder by sampling through the latent space. However, the usual log-likelihood estimate using L_1 reconstruction cost leads to the generation of blurry images. Therefore, there has been a spur of recent work which aim to combine VAEs and GANs to jointly overcome each others shortcomings, cf. e.g.:

M. Rosca, B. Lakshminarayanan, D. Warde-Farley, and S. Mohamed. Variational approaches for auto-encoding generative adversarial networks. arXiv preprint arXiv:1706.04987, 2017.

[0006] Notably in this work, the VAE objective with the L_1 reconstruction likelihood is combined with a GAN discriminator based synthetic likelihood leading to image quality at par with plain GANs.

[0007] However, the reconstruction log-likelihood and the latent space constraint in the VAE objective are at odds, which makes it difficult to achieve both at the same time. This problem is further exacerbated with the addition of the synthetic likelihood in hybrid VAE-GANs. This forces the encoder to trade-off between the two and makes latent spaces drift from true Gaussian. This leads to the degradation in the quality and diversity of generated images at test time.

SUMMARY OF THE DISCLOSURE

[0008] Currently, it remains desirable to enable an encoder to maintain both the latent representation constraint and high data log-likelihood and at the same time enhance the realism of generated images. In particular, it remains desirable to achieve high data log-likelihood and low divergence to the latent prior at the same time while generating realistic images.

[0009] Therefore, according to the embodiments of the present disclosure, a (desirably computer-implemented) method of training a model for image generation is provided. The model comprises (or is) a hybrid variational auto-encoder (VAE) – generative adversarial network (GAN) framework (i.e. architecture). The method comprises the steps of:

a – multiple input of an input image (i.e. of the same input image) into the VAE which outputs in response multiple distinct output image samples,

b – determine the best of the multiple output image samples as a best-of-many sample, the best-of-many sample having the minimum reconstruction cost, and

c – train the model based on a predefined training objective, the predefined training objective integrating the best-of-many sample reconstruction cost and a GAN-based synthetic likelihood term.

[0010] By providing such a method, a novel objective is proposed which integrates a "Best-of-Many" sample reconstruction cost and a synthetic likelihood term. This proposed objective enables the hybrid VAE-GAN framework to achieve high data log-likelihood and low divergence to the latent prior at the same time.

[0011] In other words, the constraints on the VAE can be relaxed, giving the encoder multiple chances to draw samples with high reconstruction likelihood –

only the best sample being penalized so that it can achieve both good reconstructions and maintain a latent space close to Gaussian. Furthermore, a synthetic likelihood term can be integrated in the novel objective to yield a novel hybrid VAE-GAN framework. The GAN-based synthetic likelihood term integrated to the objective can enhance the realism of generated images.

[0012] The model may be trained by using only the best-of-many sample for training the model and by disregarding the further multiple output image samples.

[0013] The model may be trained based on the best-of-many sample in relation to the input image according to a predefined VAE objective.

[0014] The model may be a (or may comprise at least one) deep neural network.

[0015] In particular the model may comprise a variational auto-encoder (VAE) including a recognition network and a generator and a generative adversarial network (GAN) including a generator and a discriminator.

[0016] The variational auto-encoder (VAE) and the generative adversarial network (GAN) may share a common generator. Hence, the model is desirably "hybrid" in the sense that the VAE and the GAN share the same Generator G_θ.

[0017] The model may be trained in step c based on the GAN-based synthetic likelihood term to learn generating sharper images by leveraging a discriminator of the GAN which is jointly trained to distinguish between real and generated images.

[0018] During each training iteration the latent distribution of the input image may be sampled by multiple input of the input image into a recognition network which outputs in response respective regions in a latent space, and generation of respective output image samples in the image space by inputting the respective regions in the latent space into a generator.

[0019] The output image samples are inputted into a discriminator of the GAN which outputs the GAN-based synthetic likelihood term.

[0020] More in particular or as an alternative only the worst of the multiple output image samples may be inputted into a discriminator of the GAN which outputs the GAN-based synthetic likelihood term. With regard to the multiple output image samples, the term "worst" may mean the least realistic of the multiple output image samples.

[0021] The GAN-based synthetic likelihood term may have a Lipschitz constant. This Lipschitz constant may be constrained to be equal to a predetermined value, in particular equal to 1, using e.g. Spectral Normalization.

[0022] The present disclosure further relates to a (computer) system for training a model for image generation. The model comprises a hybrid variational auto-encoder (VAE) – generative adversarial network (GAN) framework. The system comprises:

a module A configured for a multiple input of an input image into the VAE which outputs in response multiple distinct output image samples,

a module B for determining the best of the multiple output image samples as a best-of-many sample, the best-of-many sample having the minimum reconstruction cost, and

a module C for training the model based on a predefined training objective, the predefined training objective integrating the best-of-many sample reconstruction cost and a GAN-based synthetic likelihood term.

[0023] The system may comprise the model, i.e. a hybrid variational auto-encoder (VAE) – generative adversarial network (GAN) framework.

[0024] The system may comprise further (sub-) modules and features corresponding to the features of the method described above.

[0025] The present disclosure further relates to a (computer) system for generating an image sample, comprising the trained model of step c of the method described above or of the trained module D of the system described above.

[0026] Furthermore the present disclosure relates to a computer program including instructions for executing the steps of a method, as described above, when said program is executed by a computer.

[0027] This program can use any programming language and take the form of source code, object code or a code intermediate between source code and object code, such as a partially compiled form, or any other desirable form.

[0028] Finally, the present disclosure relates to a recording medium readable by a computer and having recorded thereon a computer program including instructions for executing the steps of a method, as described above.

[0029] The information medium can be any entity or device capable of storing the program. For example, the medium can include storage means such

as a ROM, for example a CD ROM or a microelectronic circuit ROM, or magnetic storage means, for example a diskette (floppy disk) or a hard disk.

[0030] Alternatively, the information medium can be an integrated circuit in which the program is incorporated, the circuit being adapted to execute the method in question or to be used in its execution.

[0031] It is intended that combinations of the above-described elements and those within the specification may be made, except where otherwise contradictory.

[0032] It is to be understood that both the foregoing general description and the following detailed description are exemplary and explanatory only and are not restrictive of the disclosure, as claimed.

[0033] The accompanying drawings, which are incorporated in and constitute a part of this specification, illustrate embodiments of the disclosure and together with the description, and serve to explain the principles thereof.

BRIEF DESCRIPTION OF THE DRAWINGS

[0034] Fig. 1 shows a schematic flow chart of the steps of a method for training a model for image generation according to embodiments of the present disclosure;

[0035] Fig. 2 shows a schematic block diagram of a system according to embodiments of the present disclosure; and

[0036] Fig. 3 shows a schematic block diagram of a hybrid VAE-GAN model according to embodiments of the present disclosure.

DESCRIPTION OF THE EMBODIMENTS

[0037] Reference will now be made in detail to exemplary embodiments of the disclosure, examples of which are illustrated in the accompanying drawings. Wherever possible, the same reference numbers will be used throughout the drawings to refer to the same or like parts.

[0038] Fig. 1 shows a schematic flow chart of the steps of a method for training a model for image generation according to embodiments of the present disclosure. The model has a hybrid variational auto-encoder (VAE) – generative adversarial network (GAN) architecture.

[0039] The aim of the training method is to learn generative models for image distributions $\mathrm{x} \sim p(\mathrm{x})$ that transform a latent distribution $\mathrm{z} \sim p(\mathrm{z})$ to a

learned distribution $\hat{x} \sim p_\theta(x)$ approximating $p(x)$. The samples from the learned distribution $\hat{x} \sim p_\theta(x)$ must be sharp and realistic (likely under $p(x)$) and diverse – covering all modes of the distribution $p(x)$.

[0040] In a first step S01 the same input image in inputted multiple times into the VAE which outputs in response respective multiple distinct output image samples. This allows the encoder multiple chances to draw desired samples.

[0041] In a subsequent step S02 the best of the multiple output image samples is determined. Said best output image is referred to in the following as a "best-of-many sample". The best-of-many sample is characterized by having the minimum reconstruction cost compared to the other output samples.

[0042] In a further step S03 the model is trained based on a predefined training objective. Said predefined training objective integrates (or is based on or comprises) the best-of-many sample reconstruction cost and a GAN-based synthetic likelihood term.

[0043] Due to this objective the encoder is enabled to maintain low divergence to the prior while generating realistic images. Further desirable details of the training method are described in the following, also in context of fig. 3.

[0044] Fig. 2 shows a schematic block diagram of a system according to embodiments of the present disclosure.

[0045] In this figure, a system 200 for training a model for image generation has been represented. The model comprises a hybrid variational auto-encoder (VAE) – generative adversarial network (GAN) framework. This system 200, which may be a computer, comprises a processor 201 and a non-volatile memory 202. The system 200 may not only be configured for training the model for image generation. It may also apply the trained model to another algorithm 400. For example the trained model may be applied to a computer vision system 400. In other words, a computer vision system for processing an input image sample 400 may comprise a pre-processor module configured to generate image samples based, the pre-processor module comprising said trained model.

[0046] As an option, the system 200 may further be connected to a (passive) optical sensor 300, in particular a digital camera. The digital camera

300 is configured such that it can take pictures which may be used as input image samples provided to the model.

[0047] In the non-volatile memory 202, a set of instructions is stored and this set of instructions comprises instructions to perform a method for training a model.

[0048] In particular, these instructions and the processor 201 may respectively form a plurality of modules:

a module A configured for a multiple input of an input image into the VAE which outputs in response multiple distinct output image samples,

a module B for determining the best of the multiple output image samples as a best-of-many sample, the best-of-many sample having the minimum reconstruction cost, and

a module C for training the model based on a predefined training objective, the predefined training objective integrating the best-of-many sample reconstruction cost and a GAN-based synthetic likelihood term.

[0049] Fig. 3 shows a schematic block diagram of a hybrid VAE-GAN model according to embodiments of the present disclosure. In particular, fig. 3 shows the model architecture at training time. The model is "hybrid" such that the VAE and the GAN share the same Generator G_θ.

[0050] The model thus leverages the strengths of VAEs and GANs to attain the two goals set out above. The GAN portion (G_θ, D_I) alone cangenerate realistic images, but has trouble covering all modes. The VAE portion $(R_\phi, G_\theta, D_\mathrm{L})$ can cover all modes of the distribution $p(\mathrm{x})$. However, this comes at a cost – it is difficult to maintain both the VAE latent space close to Gaussain and cover all modes of the distribution $p(\mathrm{x})$ at the same time. Therefore, in contrast to previous hybrid VAE-GAN approaches (Rosca et. al. as cited above), a novel objective is employed which leverages "Best-of-Many" samples to cover all modes of the distribution $p(\mathrm{x})$ while generating realistic images and maintaining a latent space as close to Gaussian as possible.

[0051] The following detailed description begins with an explanation of the VAE objective and its shortcomings, followed by the proposed "Best-of-Many" objective for image generation which address its shortcomings.

Shortcomings of the VAE objective

[0052] The VAE objective maximizes the log-likelihood of the data ($x \sim p(x)$). The log-likelihood, assuming the latent space to be distributed according to $p(z)$ is,

$$\log(p_\theta(x)) = \log\left(\int p_\theta(x|z)p(z)dz\right). \quad (1)$$

[0053] Here, $p(z)$ is usually Gaussian and the log-likelihood $p_\theta(x|z)$ is usually the L_1/L_2 norm based reconstruction ($e^{-\lambda\|x-\hat{x}\|_n}$). This requires the generator G_θ to generate samples that reconstruct every training example x for a likely $z \sim p(z)$. This ensures that the decoder θ covers all modes of the data distribution $x \sim p(x)$. In contrast, GANs never directly maximize the (reconstruction based) likelihood and there is no direct incentive to cover all modes.

[0054] However, the integral in (1) is intractable. Variational inference may use an (approximate) variational distribution $q_\phi(z|x)$, which is jointly learned using an encoder,

$$\log(p_\theta(x)) = \log\left(\int p_\theta(x|z)\frac{p(z)}{q_\phi(z|x)}q_\phi(z|x)dz\right). \quad (2)$$

[0055] During training, samples may be drawn instead from a recognition network $q_\phi(z|x)$ (R_ϕ) and the variational auto-encoder based objective may be maximized,

$$\mathcal{L}_{VAE} = \mathbb{E}_{q_\phi(z|x)}\log(p_\theta(x|z)) - KL(p(z)||q_\phi(z|x)). \quad (3)$$

[0056] This objective has two important shortcomings. Firstly, this objective severly constrains the recognition network $q_\phi(z|x)$ (R_ϕ) as high data log-likelihood and low divergence to the prior are at odds. As the expected log-likelihood is considered, the recognition network has to always generate latent samples $\hat{z}$ which are decoded by the generator close to x. Otherwise, the expected data log-likelihood would be low. Thus, the encoder is forced to trade-off between a good estimate of the data log-likelihood and the divergence to the true latent $p(z)$ distribution, which causes the generated latent space (by the recognition network) to be far from a Gaussain. Secondly, it considers only a reconstruction-based log-likelihood which is known to lead to blurry image generations.

[0057] Next, it is described how multiple samples can be effectively leveraged from $q_\phi(\mathrm{z}|\mathrm{x})$ to deal with the first shortcoming. Finally, a synthetic likelihood term is integrated to deal with blurriness.

Leveraging Multiple Samples

[0058] An alternative variational approximation of (1) may be derived, which uses multiple samples to relax the constrains on the recognition network. For example, the Mean-value theorem of Integration may be used, in order to derive a unconditional version of the (conditional) multi-sample objective starting from (2) (full derivation in Suppmat),

$$\mathcal{L}_{\mathrm{MS}} = \log\left(\int p_\theta(\mathrm{x}|\mathrm{z}) q_\phi(\mathrm{z}|\mathrm{x})\, dz\right) - KL\left(p(\mathrm{z}) \,||\, q_\phi(\mathrm{z}|\mathrm{x})\right). \quad (4)$$

[0059] In comparision to the VAE objective (3), in (4) the likelihood is computed considering all the generated samples. The recognition network gets multiple chances to draw samples with high likelihood. This encourages diversity in the generated samples and the recognition network can provide a good estimate of the data log-likelihood while not diverging from the prior $p(\mathrm{z})$ – without trade-off.

[0060] However, also a good estimate of the likelihood $p_\theta(\mathrm{x}|\mathrm{z})$ is desirable. Considering only L_1 or L_2 reconstruction based likelihoods would lead to the generation of blurry images. Therefore, (and because of the intractability of (1)), GANs instead use an adversary that provides indirect information of the likelihood – classifier that is jointly trained to distinguish between generated samples and real data samples.

[0061] Next, it is described how it can be leveraged such a classifier to directly obtain synthetic estimates of the likelihood that lead to the generation of crisp images.

Integrating Synthetic Likelihoods with the "Best-of-Many" Samples

[0062] Synthetic estimates of the likelihood leads to the generation of sharper images by leveraging a classifier which is jointly trained to distinguish

between real and generated images. A generated image which is indistinguishable from a real image is assigned higher likelihood. Starting from (4), a synthetic likelihood term (with weight $1-\alpha$) is integrated to both encourage the generator to generate realistic images and to cover all modes (L_1 reconstruction loss), thus meeting the initial two goals. First the likelihood term is converted to a likelihood ratio form which allows for synthetic estimates,

$$\log\left(\int p_\theta(\mathrm{x}|\mathrm{z})q_\phi(\mathrm{z}|\mathrm{x})\,dz\right) - KL(p(\mathrm{z}) \,||\, q_\phi(\mathrm{z}|\mathrm{x}))$$

$$= (1-\alpha)\log\left(\int p_\theta(\mathrm{x}|\mathrm{z})q_\phi(\mathrm{z}|\mathrm{x})\,dz\right) +$$

$$\alpha\log\left(\int p_\theta(\mathrm{x}|\mathrm{z})q_\phi(\mathrm{z}|\mathrm{x})\,dz\right) - KL(p(\mathrm{z}) \,||\, q_\phi(\mathrm{z}|\mathrm{x}))$$

$$\propto (1-\alpha)\log\left(\int \frac{p_\theta(x|z)}{p(x)}q_\phi(\mathrm{z}|\mathrm{x})\,dz\right) +$$

$$\alpha\log\left(\int p_\theta(\mathrm{x}|\mathrm{z})q_\phi(\mathrm{z}|\mathrm{x})\,dz\right) - KL(p(\mathrm{z}) \,||\, q_\phi(\mathrm{z}|\mathrm{x})). \qquad (5)$$

[0063] Now the likelihood ratio $p_\theta(\mathrm{x}|\mathrm{z}) \,/\, p(\mathrm{x})$ can be estimated using a classifier. To do this, the auxiliary variable y is introduced where, y = 1 denotes that the sample was generated and y = 0 denotes that the sample is from the true distribution. Now (6) can be written as (using Bayes theorem),

$$(1-\alpha)\log\left(\int \frac{p_\theta(x|z,y=1)}{p(x|y=0)}q_\phi(\mathrm{z}|\mathrm{x})\,dz\right) +$$

$$\alpha\log\left(\int p_\theta(\mathrm{x}|\mathrm{z})q_\phi(\mathrm{z}|\mathrm{x})\,dz\right) - KL(p(\mathrm{z})||q_\phi(\mathrm{z}|\mathrm{x})).$$

$$= (1-\alpha)\log\left(\int \frac{p_\theta(y=1|z,x)}{p(y=0|x)}q_\phi(\mathrm{z}|\mathrm{x})\,dz\right) + \qquad (6)$$

$$\alpha\log\left(\int p_\theta(\mathrm{x}|\mathrm{z})q_\phi(\mathrm{z}|\mathrm{x})\,dz\right) - KL(p(\mathrm{z})||q_\phi(\mathrm{z}|\mathrm{x}))$$

$$= (1-\alpha)\log\left(\int \frac{p_\theta(y=1|z,x)}{1-p(y=1|x)}q_\phi(\mathrm{z}|\mathrm{x})\,dz\right) +$$

$$\alpha\log\left(\int p_\theta(\mathrm{x}|\mathrm{z})q_\phi(\mathrm{z}|\mathrm{x})\,dz\right) - KL(p(\mathrm{z})||q_\phi(\mathrm{z}|\mathrm{x})).$$

[0064] The probability $p_\theta(y = 1|z,x)$ may be estimated using a classifier $D_I(x)$ (image discriminator in fig. 3) which is jointly trained, leading to a synthetic estimate of the likelihood ratio,

$$\mathcal{L}_{\text{MS-S}} \propto (1-\alpha)\log\left(\int \frac{D_I(x|z)}{1-D_I(x|z)} q_\phi(z|x)\,dz\right) \qquad (7)$$

$$+\alpha \log\left(\int p_\theta(x|z) q_\phi(z|x)\,dz\right) - KL(p(z)||q_\phi(z|x)).$$

[0065] Note that the synthetic likelihood $D_I(x)$ is usually estimated using a softmax layer and the likelihood $p_\theta(x|z)$ takes the form $e^{-\lambda\|x-\hat{x}\|_n}$ in (7). Both these log-sum-exps are numerically unstable. It can be dealt with the first log-sum-exp using the Jenson-Shannon inequality,

$$\log\left(\int \frac{D_I(x|z)}{1-D_I(x|z)} q_\phi(z|x)\,dz\right) \geq \mathbb{E}_{q_\phi(z|x)} \log\left(\frac{D_I(x|z)}{1-D_I(x|z)}\right)$$

[0066] As stochastic gradient descent is performed, it can be dealt with the second log-sum-exp after stochastic (MC) sampling of the data points. The log-sum-exp can be well estimated using the max – the "Best-of-Many" samples,

$$\log\left(\frac{1}{T}\sum_{i=1}^{i=T} p_\theta(x|\hat{z}^i)\right) \geq \max_i \log\left(p_\theta(x|\hat{z}^i)\right) - \log(T)$$

[0067] The "Best-of-Many" samples objective takes the form (ignoring the constant $\log(T)$ term and $\lambda \leq (1-\alpha)$),

$$\mathcal{L}_{\text{BMS-S}} = \lambda\, \mathbb{E}_{q_\phi(z|x)} \log\left(\frac{D_I(x|z)}{1-D_I(x|z)}\right) \qquad (8)$$

$$+\alpha \max_i \log\left(p_\theta(x|\hat{z}^i)\right) - p(z)q_\phi(z|x).$$

[0068] Furthermore, the generator $G_\theta$may be penalized using only the least realistic sample, and the likelihood ratio be estimated directly using D_I,

$$\mathcal{L}_{\text{BMS-S}} = \lambda \min_i \log\left(D_I(x|\hat{z}^i)\right) \qquad (9)$$

$$+\alpha \max_i \log\left(p_\theta(x|\hat{z}^i)\right) - KL(p(z) \parallel q_\phi(z|x)).$$

[0069] To further ensure smoothness, the Lipschitz constant K of D_I may be directly controlled, by setting it to be equal to 1, using Spectral Normalization,

T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks. ICLR, 2018.

The synthetic likelihood ratio term is namely unstable during training – as it is the ratio of outputs of a classifier, any instability in the output of the classifier is magnified. Therefore it is proposed to directly estimate the ratio using a network with a controlled Lipschitz constant, which leads to significantly improved stability.

[0070] In contrast to prior work (e.g. Rosca et.al.), (8) provides multiple chances to the recognition network to generate samples likely under the reconstruction based likelihood. Furthermore, the synthetic likelihood term ensures that every generated sample is realistic.

[0071] Intuitively, this objective can be seen as a generalization of prior hybrid VAE-GAN based models. If it is set $T = 1$ in (8) the exact objective used in the α-GAN model is recovered. Moreover, in e.g. Rosca et.al. for every sample $\mathrm{x} \sim p(\mathrm{x})$, the recognition network is used to obtain the exact $\hat{\mathrm{z}}$ from latent space. In contrast, the objective (8) only requires the recognition network to only point to the appropriate region in the latent space.

[0072] Next, a detailed description of the optimization of the hybrid VAE-GAN model is provided using the "Best-of-Many" samples objective, which is called BMS-GAN.

Optimization

[0073] As recent works (e.g. Rosca et.al.) have shown, point-wise minimization of the KL-divergence using its analytical form leads to degradation in generatated image quality. The KL-divergence term can also be recast in a likelihood ratio form (similar as (6)) allowing to leverage synthetic likelihoods using a classifier and minimize it globally instead of point-wise. The latent space discriminator D_{L} is used to enforce the KL-divergence constraint $p(\mathrm{z})q_{\phi}(\mathrm{z}|\mathrm{x})$ in (8).

[0074] During optimization, samples from the true data distribution $\mathrm{x} \sim p(\mathrm{x})$ are first sampled. For each x, the recognition network R_{ϕ} gives a region of the latent space $q_{\phi}(\mathrm{z}|\mathrm{x})$. It is assumed $q_{\phi}(\mathrm{z}|\mathrm{x}) = \mathcal{N}(\mu(x), \sigma(x))$. The generator G_{θ} now generates samples in the data (image) space $\hat{\mathrm{x}} \sim p_{\theta}(\mathrm{x}|\mathrm{z})q_{\phi}(\mathrm{z}|\mathrm{x})$ from that region of the latent space. These samples are then given as input to the data

(image) discriminator D_{I}, which provides a synthetic estimate of the likelihood. The latent space discriminator D_{L} uses the latent samples $\hat{\mathrm{z}} \sim q_{\phi}(z|x)$ to provide a synthetic estimate of the divergence $\mathrm{KL}(p(\mathrm{z})||q_{\phi}(\mathrm{z}|\mathrm{x}))$.

[0075] Based on the generated samples and synthetic likelihood estimates, it is now updated: 1. D_{I} and D_{L} using the standard GAN update rule (using true and generated samples x and $\hat{\mathrm{x}}$, z and $\hat{\mathrm{z}}$). 2. R_{ϕ} using synthetic likelihood estimates from D_{I}, D_{L} and the "Best-of-Many" reconstruction cost $\max_i \log(p_{\theta}(\mathrm{x}|\hat{\mathrm{z}}^i))$. 3. G_{θ} using synthetic likelihood estimate from D_{I} and the "Best-of-Many" reconstruction cost.

[0076] Throughout the description, including the claims, the term "comprising a" should be understood as being synonymous with "comprising at least one" unless otherwise stated. In addition, any range set forth in the description, including the claims should be understood as including its end value(s) unless otherwise stated. Specific values for described elements should be understood to be within accepted manufacturing or industry tolerances known to one of skill in the art, and any use of the terms "substantially" and/or "approximately" and/or "generally" should be understood to mean falling within such accepted tolerances.

[0077] Although the present disclosure herein has been described with reference to particular embodiments, it is to be understood that these embodiments are merely illustrative of the principles and applications of the present disclosure.

[0078] It is intended that the specification and examples be considered as exemplary only, with a true scope of the disclosure being indicated by the following claims.

CLAIMS

1. A method of training a model for image generation,
the model comprising a hybrid variational auto-encoder (VAE) – generative adversarial network (GAN) framework,
the method comprising the steps of:
a - multiple input (S01) of an input image into the VAE which outputs in response multiple distinct output image samples,
b – determine (S02) the best of the multiple output image samples as a best-of-many sample, the best-of-many sample having the minimum reconstruction cost,
c – train (S03) the model based on a predefined training objective, the predefined training objective integrating the best-of-many sample reconstruction cost and a GAN-based synthetic likelihood term.

2. The method according to any one of the preceding claims 1, wherein the model is trained by using only the best-of-many sample for training the model and by disregarding the further multiple output image samples.

3. The method according to any one of the preceding claims 1 and 2, wherein
the model is trained based on the best-of-many sample in relation to the input image according to a predefined VAE objective.

4. The method according to any one of the preceding claims, wherein
the model is a deep neural network or comprises at least one deep neural network.

5. The method according to any one of the preceding claims, wherein
the model comprises:
a variational auto-encoder (VAE) including a recognition network and a generator, and
a generative adversarial network (GAN) including a generator and a discriminator.

6. The method according to the preceding claim, wherein
the variational auto-encoder (VAE) and the generative adversarial network (GAN) share a common generator.

7. The method according to any one of the preceding claims, wherein
the model is trained in step c based on the GAN-based synthetic likelihood term to learn generating sharper images by leveraging a discriminator of the GAN which is jointly trained to distinguish between real and generated images.

8. The method according to any one of the preceding claims, wherein
during each training iteration the latent distribution of the input image is sampled by:
multiple input of the input image into a recognition network which outputs in response respective regions in a latent space, and
generation of respective output image samples in the image space by inputting the respective regions in the latent space into a generator.

9. The method according to any one of the preceding claims, wherein
the output image samples are inputted into a discriminator of the GAN which outputs the GAN-based synthetic likelihood term, or
only the worst of the multiple output image samples is inputted into a discriminator of the GAN which outputs the GAN-based synthetic likelihood term.

10. The method according to any one of the preceding claims, wherein
the Lipschitz constant of the GAN-based synthetic likelihood term is constrained to be equal to a predetermined value, in particular equal to 1, using Spectral Normalization.

11. A system for training a model for image generation,
the model comprising a hybrid variational auto-encoder (VAE) – generative adversarial network (GAN) framework, the system comprising:
a module A configured for a multiple input of an input image into the VAE which outputs in response multiple distinct output image samples,

a module B for determining the best of the multiple output image samples as a best-of-many sample, the best-of-many sample having the minimum reconstruction cost, and

a module C for training the model based on a predefined training objective, the predefined training objective integrating the best-of-many sample reconstruction cost and a GAN-based synthetic likelihood term.

12. The system according to the preceding claim, further comprising the model.

13. A system for generating an image sample,

comprising the trained model of step c of any one of claims 1 to 10 or of the trained module C of claim 11 or 12.

14. A computer program comprising instructions for executing the steps of the method according to any one of the preceding method claims 1 to 10, when the program is executed by a computer.

15. A recording medium readable by a computer and having recorded thereon a computer program including instructions for executing the steps of a method according to any one of claims 1 to 10.

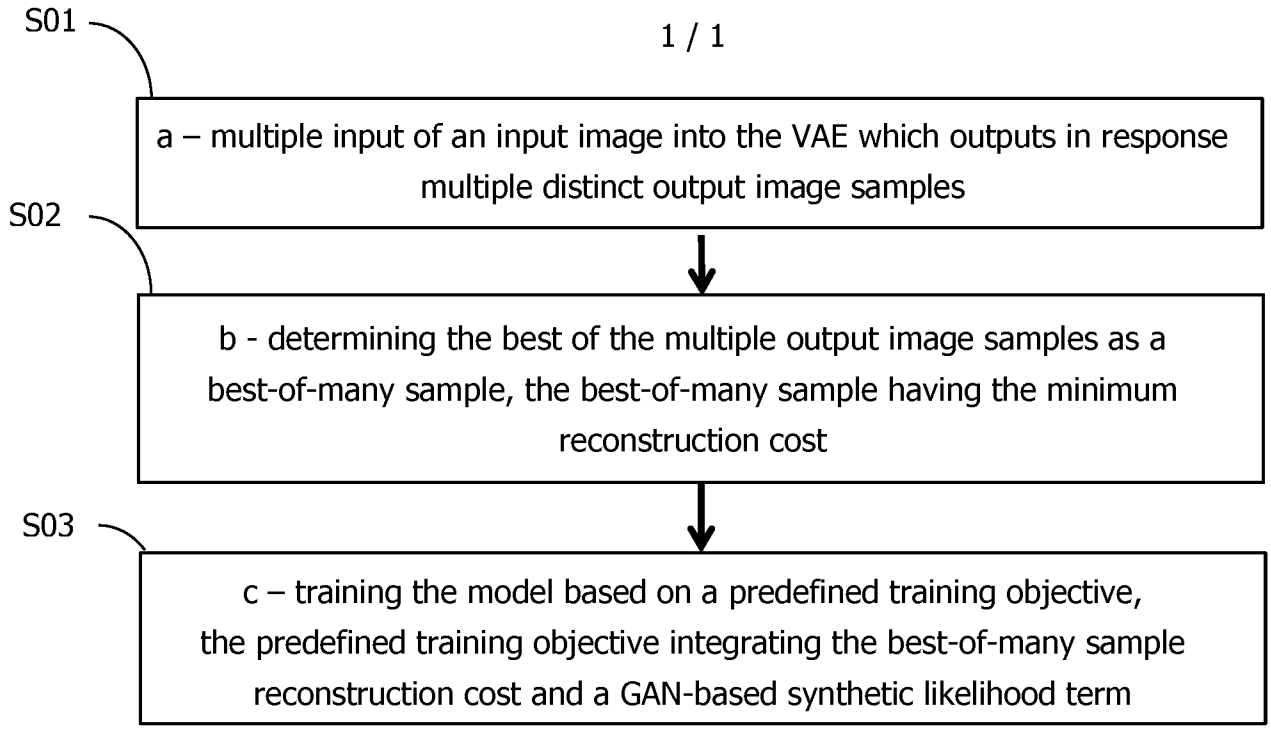

Fig. 1

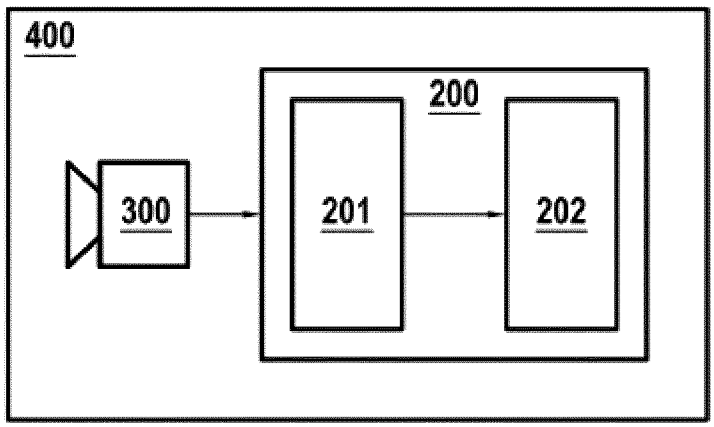

Fig. 2

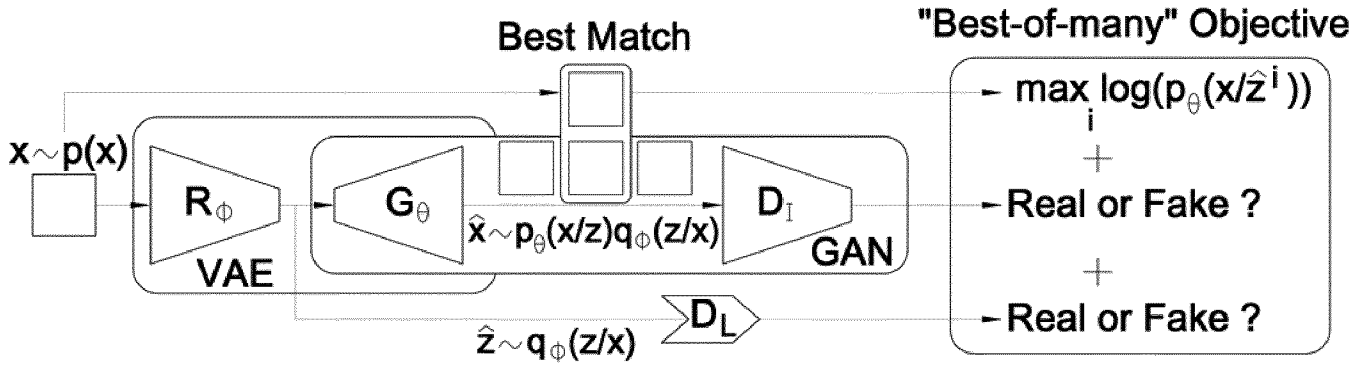

Fig. 3

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2019/063853

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06T11/00
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G06T

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

EPO-Internal, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	LARS MESCHEDER ET AL: "Adversarial Variational Bayes: Unifying Variational Autoencoders and Generative Adversarial Networks", ARXIV.ORG, CORNELL UNIVERSITY LIBRARY, 201 OLIN LIBRARY CORNELL UNIVERSITY ITHACA, NY 14853, 17 January 2017 (2017-01-17), XP081324143, abstract section 1 Introduction and Algorithm 1 section 5.2 Generative Models 5. Experiments ----- -/--	1-15

[X] Further documents are listed in the continuation of Box C. [ ] See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
10 January 2020	21/01/2020

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Blaszczyk, Marek

1

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	BHATTACHARYYA APRATIM ET AL: "Accurate and Diverse Sampling of Sequences Based on a "Best of Many" Sample Objective", 2018 IEEE/CVF CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, IEEE, 18 June 2018 (2018-06-18), pages 8485-8493, XP033473772, DOI: 10.1109/CVPR.2018.00885 [retrieved on 2018-12-14] 1. Introduction 3.2. Best of Many Samples Objective	1-15
A	ALEXEY DOSOVITSKIY ET AL: "Generating Images with Perceptual Similarity Metrics based on Deep Networks", ARXIV.ORG, CORNELL UNIVERSITY LIBRARY, 201 OLIN LIBRARY CORNELL UNIVERSITY ITHACA, NY 14853, 8 February 2016 (2016-02-08), XP080682203, the whole document	1-15
A	Alireza Makhzani ET AL: "Adversarial Autoencoders", , 25 May 2016 (2016-05-25), pages 1-16, XP055532752, Retrieved from the Internet: URL:https://arxiv.org/pdf/1511.05644.pdf [retrieved on 2018-12-11] the whole document	1-15
A	Shuangfei Zhai ET AL: "GENERATIVE ADVERSARIAL NETWORKS AS VARIA- TIONAL TRAINING OF ENERGY BASED MODELS", , 1 January 2017 (2017-01-01), XP055655707, Retrieved from the Internet: URL:https://arxiv.org/pdf/1611.01799.pdf [retrieved on 2020-01-08] the whole document	1-15

1