Quantifying Predictive Uncertainty for Stochastic Video Synthesis from Audio

Moitreya Chatterjee[1] Narendra Ahuja[1] Anoop Cherian[2]
[1]University of Illinois at Urbana-Champaign Champaign, IL 61820, USA
[2]Mitsubishi Electric Research Laboratories, Cambridge, MA 02139
metro.smiles@gmail.com n-ahuja@illinois.edu cherian@merl.com

Abstract

In this paper, we study the problem of synthesizing video frames from the accompanying audio and a few past frames – a task with immense potential, e.g., in occlusion reasoning. Prior methods to solve this problem often train deep learning models that derive their training signal by computing the mean-squared error (MSE) between the generated frame and the ground truth. However, these techniques do not account for the predictive uncertainty of the frame generation model. This frailty might result in sub-optimal training, especially when this uncertainty is high. To address this challenge, we introduce Predictive Uncertainty Quantifier *(PUQ) - a stochastic quantification of the generative model's predictive uncertainty, which is then used to weigh the MSE loss. PUQ is derived from a hierarchical, variational deep net and is easy to implement and incorporate into audio-conditioned stochastic frame generation methods. Experiments demonstrate our method's faster and improved convergence versus competing baselines on two challenging datasets.*

1. Introduction

In this work, we explore the problem of generating a sequence of semantically plausible and contextually consistent video frames while listening to only the accompanying audio. This task holds immense potential in developing a variety of useful applications; including reasoning under occlusions or developing platforms for the hearing impaired.

From a machine learning perspective, generating the frames of a video, given only the audio modality, is severely ill-posed and thus most prior methods explore this research direction in limited settings, such as synthesizing talking heads, given the speech [8]. Only recently has the community started exploring this problem in more unrestricted settings [2]. However, this entails dealing with the challenging problem of capturing the possible stochasticity in the events in the video. To characterize this stochasticity, a stochastic frame generation model, dubbed *Sound2Sight*, was recently proposed [2]. This model uses the audio features and the video context to generate frames. Specifically, it augments a standard encoder-decoder sequence-to-sequence model for predicting a future frame [5] using a stochastic module that characterizes the random details in the generated frames that the sound alone cannot account for. However, Sound2Sight is trained mainly using the mean-squared-error (MSE) loss, by comparing the generated frames against one of a potentially infinite set of plausible ground truths (arising because of the stochasticity). As a result, minimizing the mean-squared error may be difficult and training the generative model could be challenging.

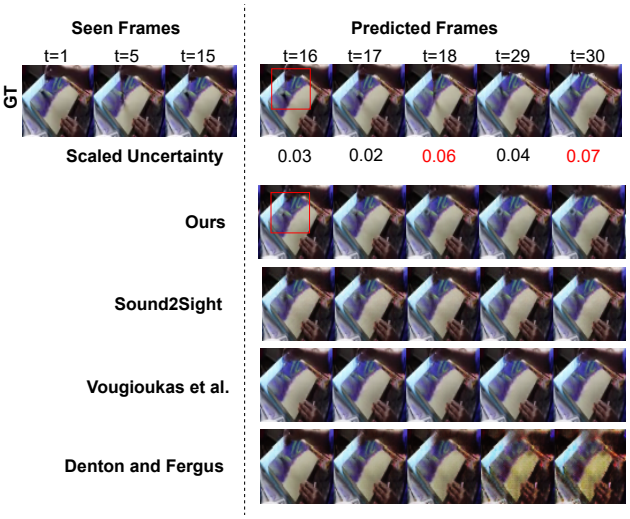


Figure 1: Generation results by our method (PUQ) versus baselines on YouTube Painting [2]. The red square indicates the region of dominant motion. Scaled uncertainty ($\frac{1}{m_t}$) is high (in red), when the painter abruptly switches strokes.

To address the above difficulty, we present *Predictive Uncertainty Quantifier* (PUQ) – a hierarchical, variational setup to estimate the predictive uncertainty of an audio-conditioned visual frame generation model. The key intuition behind PUQ comes from the observation that the predictive uncertainty is directly linked to the variance in the distribution of the latent space embedding of the next frame of, for example, a Sound2Sight [2] model. However, in such models this uncertainty is modeled in a latent space and thus can neither be directly interpreted in the space of generated frames, nor can it be explicitly incorporated in the training objective. Moreover, the space of generated frames is very high dimensional (equal to the number of pixels), and producing a variance for every pixel could be hard. Thus, we design PUQ to be a scalar quantifier that captures a summary of the predictive uncertainty of the generated frame and derive it using a hierarchical, variational framework.

To empirically verify the effectiveness of our model, we present experiments on a synthetic dataset: Multimodal Stochastic Moving MNIST (M3SO) [2], and a challenging real world dataset: YouTube-Painting [2]. Our results show that our framework trains faster than prior methods,

and leads to state-of-the-art generation quality (see Figure 1).

2. Related Works

Video Prediction/Forecasting entails predicting the future frames of a video, given a few frames from the past. Towards this end, *deterministic* techniques typically employ an encoder-decoder architecture to generate video frames autoregressively, without modeling the data stochasticity (e.g., the randomness in an object's motion, compression noise, etc.) [5]. This caveat has been addressed by stochastic frame prediction approaches [4, 1, 3], however they are mostly unimodal . Different from these techniques, our model is both stochastic and multimodal.

Techniques for Video Generation from Audio so far have mostly operated in very restricted settings, such as generating human face animations conditioned on speech [9]. These approaches often make use of additional details to simplify the problem further – such as using the identity of the person, facial landmarks, etc.- hindering their application to more generic settings. Some of them, however, propose to synthesize face motions directly from speech and an initial frame and are thus more scalable [8]. Chatterjee and Cherian [2], attempt to synthesize generic videos from audio, by employing a variational learning scheme coupled with adversarial training. Nonetheless, different from our method, these frameworks do not model the predictive uncertainty of the generated frames.

3. Background

Let $x_{1:T} := [x_1; x_2; \ldots; x_T]$ denote a sequence of T frames of a video and $a_{1:T} := [a_1; a_2; \ldots; a_T]$, the corresponding time-aligned audio samples. Assuming we have access to a few initial frames $x_{1:S}$, to set the visual context (where $1 \leq S < T$), and the audio-stream $a_{1:T}$, our goal is to generate the remaining frames $x_{S+1:T}$ autoregressively. This task requires estimating a prediction model $p_\theta(\cdot)$, parameterized by θ, which generates the frame x_t and minimizes the expected negative log-likelihood. When generating x_t, we use the audio $a_{1:t}$ to peek into the future.

In order to incorporate stochasticity into the generative setup, one may assume that the frame generation module is conditioned on a latent prior model, derived from the provided audio $a_{1:t}$, i.e., $z_t \sim p(z_t|a_{1:t})$. The output frame x_t is then generated by $x_t \sim p_\theta(x_t|x_{1:t-1}; z_t)$. However, one of the key problems that plagues such latent stochastic prior models is the intractability of the estimation of the *evidence*: $p(x_t|x_{1:t-1}; a_{1:t}) = \int_{z_t} p(x_t|x_{1:t-1}; z_t)p(z_t|a_{1:t})dz_t$. A typical solution to this problem [2] is to cast this estimation process in a variational encoder-decoder regime [6], where the encoder comprises a variational posterior $q_\phi(z_t|x_{1:t}; a_{1:t})$ (with parameters ϕ) that approximates the true posterior $p(z_t|x_{1:t}; a_{1:t})$, while the decoder learns a distribution for estimating x_t $p_\theta(x_t|x_{1:t-1}; z_t)$. The likelihood setup of this model is given by $\log p(x_t|x_{1:t-1}; a_{1:t}) \geq L_{\theta,\phi}$, where:

$$L_{t} = \mathbb{E}_{q_\phi(z_t|x_{1:t};a_{1:t})} \log p_\theta(x_t|x_{1:t-1}; z_t) - \mathrm{KL}\left(q_\phi(z_t|x_{1:t}; a_{1:t}) \| p(z_t|a_{1:t})\right); \text{for } t > S. \quad (1)$$

The likelihood loss in Eq. 1 can be efficiently optimized by employing the re-parametrization trick [6]. Prior works [2, 4] recommend using a data-driven stochastic prior model for improved results. Thus, $p(z_t|a_{1:t}) := p_\psi(z_t|x_{1:t-1}; a_{1:t})$, parametrized by ψ. Also note that if $p_\theta(\cdot)$ is assumed to be a Gaussian distribution with an isotropic constant variance, the expectation term in Eq. 1 boils down to a standard MSE loss over all predicted frames $x_{S+1:T}$.

While learning the distribution $q_\phi(\cdot)$ captures the stochasticity of the generation process in the latent space, the model's predictive uncertainty (in the output space) remains unaccounted for. This might unfairly penalize the model for an incorrect prediction if the frame stochasticity is high, at a certain time step. Incidentally, a direct per-pixel uncertainty estimate from the decoder results in a prohibitively large prediction space for successful training (quadratic in the number of frame pixels). Our proposed uncertainty estimation framework therefore, leverages the variance of the posterior distribution over z_t, $q_\phi(\cdot)$, to compute the uncertainty in the decoded (output) space via a two-step hierarchical process.

4. Proposed Method

Our goal is to regulate the importance of the MSE loss in the training objective (Eq. 1) by estimating the uncertainty of predicting the next frame. In our setup, as shown in Figure 2, the prediction model consists of a series of 2d convolution layers, accepting the previous frame x_{t-1} as input, followed by an LSTM - which couples this embedding with the stochasticity vector z_t. Its output is then decoded to generate the next frame, x_t, via a stack of 2D deconvolution layers. Sound2Sight [2] assumes the data likelihood model $p_\theta(x_t|x_{1:t-1}; z_t) \sim \mathcal{N}(x_t; s_t)$, with the variance $s_t > 0; s_t \in \mathbb{R}$ as an isotropic constant. Thus, the negative log-likelihood of the predicted frame $\hat{x}_t$ reduces to computing the ℓ_2-loss. Instead, in this work, we propose to directly condition the data likelihood model with the uncertainty derived from the variance of the latent-space posterior, thereby ensuring that when the uncertainty is high for a predicted frame, the ℓ_2-loss term is weighed down and vice-versa.

We proceed by assuming that the prior $p_\psi(z_t|x_{1:t-1}; a_{1:t})$ is a normal distribution $\mathcal{N}(\mu_t^z; \sigma_t^z)$, with parameters μ_t^z, the mean, and σ_t^z the covariance matrix. We instantiate this prior model with an LSTM, which operates on the concatenated audio-visual embeddings obtained from the respective unimodal Transformers [2]. We assume a similar setup for the posterior

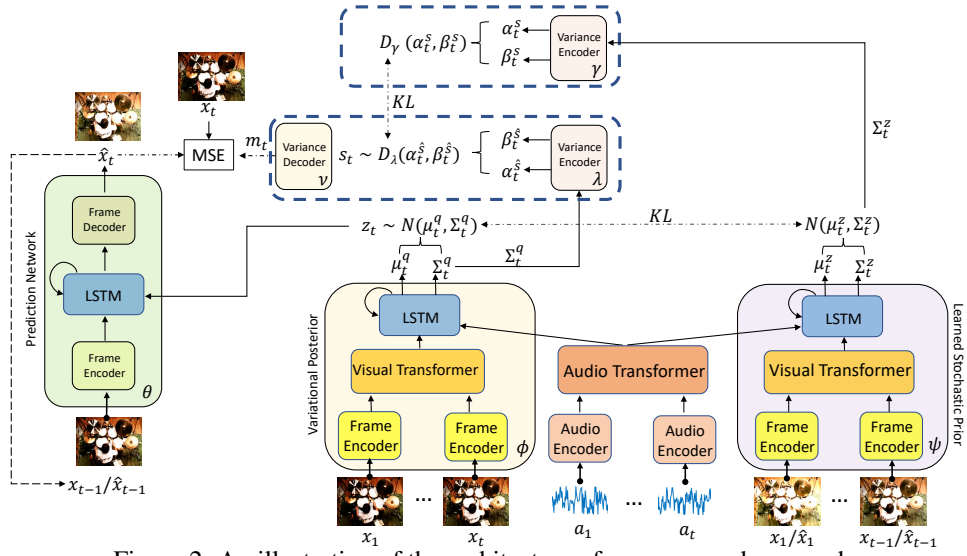


Figure 2: An illustration of the architecture of our proposed approach.

$q_\phi(z_t|x_{1:t}; a_{1:t})$ $N(\ ^q_t; \ ^q_t)$, and instantiate it using another LSTM. Let m_t 2 R; $m_t > 0$, denote the precision (or inverse variance 1=s_t) in our data likelihood model. Further, let the lielihood governing m_t be $p(m_t|x_{1:t\ \ 1}; a_{1:t})$. Then:

$$\log p(x_t|x_{1:t\ \ 1}; a_{1:t}) = \int_{z_t; m_t} \underbrace{\log p(x_t|z_t; m_t; x_{1:t\ \ 1}; a_{1:t})}_{A_1} +$$
$$+ \underbrace{\log p(z_t|x_{1:t\ \ 1}; a_{1:t})}_{A_2} + \underbrace{\log p(m_t|x_{1:t\ \ 1}; a_{1:t})}_{A_3} \, dz_t \, dm_t \tag{2}$$

Computing the above integral exactly, is intractable. So we approximate it by sampling m_t as well as z_t. Note that $A_1 + A_2$ in the above equation (Eq. 2) is lower bounded by the variational lower bound [6] (as in Eq. 1), as follows:

$$A_1 + A_2 \quad \mathsf{E}_{q_\phi(z_t|x_{1:t}; a_{1:t})} \log p\ (x_t|x_{1:t\ \ 1}; z_t; m_t)$$
$$\mathsf{KL}\,(q\ \ (z_t|x_{1:t}; a_{1:t}) k p\ \ (z_t|x_{1:t\ \ 1}; a_{1:t})); \tag{3}$$

where $t > S$. Further, as stated before, we seek to derive m_t from the uncertainty of the posterior $\ ^q_t$ in the latent space. We accomplish this via a variational encoder-decoder network, nested into the larger prediction model, thereby making our model hierarchical. For increased adaptability, PUQ permits the parameters of the posterior and prior distributions of the nested network to be learnt from data. In particular, during training, the encoder component of this network, $_\lambda(\ )$, with parameters , embeds $\ ^q_t$, to produce the sufficient statistics of the distribution governing the latent space, $q_\lambda(s_t|x_{1:t}; a_{1:t})$, while the decoder $_\nu(\ )$, with parameters , draws a sample from this distribution, s_t, and decodes it to generate m_t, where m_t $p_\nu(m_t|s_t)$.

In order to appropriately regularize the latent space distribution, $q_\lambda(\ )$, we assume a prior distribution, $p_\gamma(s_t|x_{1:t\ \ 1}; a_{1:t})$ - the parameters of which are inferred by embedding $\ ^z_t$ via a network $_\gamma(\ )$, with parameters . We assume $p_\gamma(\ )$ $D_\gamma(\ ^s_t; \ ^s_t)$ with parameters ($^s_t; \ ^s_t$) is estimated by the network $_\gamma(\ )$, while $q_\lambda(\ )$ $D_\lambda(\ ^{\hat{s}}_t; \ ^{\hat{s}}_t)$, with parameters ($^{\hat{s}}_t; \ ^{\hat{s}}_t$) is estimated by $_\lambda(\ )$, respectively.

Table 1: Human preference score: Ours vs Sound2sight [2]

Datasets	Prefer ours
M3SO	**67%**
YouTube Painting	**78%**

Given this setup, analogous to Eq. 1, for $t > S$, we get the variational lower bound [6] on the likelihood of m_t:

$$A_3 \quad \mathsf{E}_{q_\lambda(s_t|x_{1:t}, a_{1:t})} \log p_\nu(m_t|s_t)$$
$$\mathsf{KL}\,(q_\lambda(s_t|x_{1:t}; a_{1:t}) k p_\gamma(s_t|x_{1:t\ \ 1}; a_{1:t})):$$

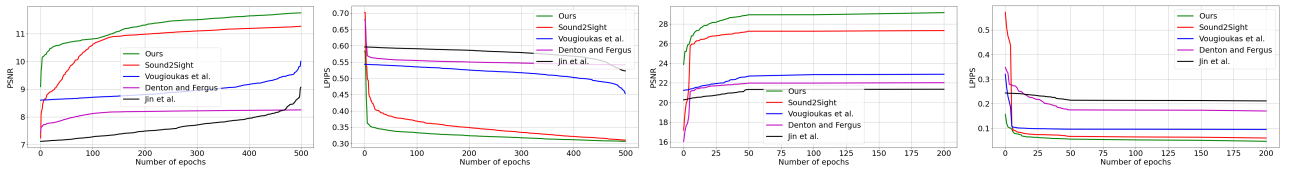
Assuming that $p_\theta(x_t|x_{1:t\ \ 1}; z_t; m_t)$ follows a Gaussian distribution $N(x_t; \frac{1}{m_t})$ and $_{1;\ 2}$ 0, $_{1;\ 2}$ 2 R, leads us to our final objective, which we minimize using the reparametrization trick [6]:

$$L^P_{\theta,\phi,\psi,\lambda} = \sum_{t=S+1}^{T} \frac{1}{2} \Big[m_t \| \hat{x}_t \quad x_t \|^2 \quad \log m_t \Big] \tag{4}$$
$$\mathsf{E}_{q_\lambda(s_t|x_{1:t}, a_{1:t})} \log p_\nu(m_t|s_t) +$$
$$_1 \mathsf{KL}(q_\phi(z_t|x_{1:t}; a_{1:t}) k p_\psi(z_t|x_{1:t\ \ 1}; a_{1:t})) +$$
$$_2 \mathsf{KL}\,(q_\lambda(s_t|x_{1:t}; a_{1:t}) k p_\gamma(s_t|x_{1:t\ \ 1}; a_{1:t})) ;$$

As is standard in variational settings, we instantiate $D_\gamma(\ ;\ )$ to be a Gaussian, however since m_t 0, $D_\lambda(\ ;\ )$ assumes a truncated Normal form. Additionally, we observed improved performance when PUQ was trained in conjunction with the multimodal discriminator of Sound2Sight [2].

5. Experiments and Results

Multimodal MovingMNIST with a Surprise Obstacle (M3SO): M3SO is a challenging, synthetic, publicly available dataset [2], which consists of digits from the MNIST dataset [7] moving along rectilinear paths in a 48 48 box, bouncing off in random directions when a collision occurs with the box boundaries or a fixed-size block introduced at a random location in the box. M3SO equips each of the ten digits with a unique Dual Tone Multi-Frequency (DTMF)



(a) PSNR - M3SO (b) LPIPS - M3SO (c) PSNR - YouTube Painting (d) LPIPS - YouTube Painting

Figure 3: Quantitative evaluation of PUQ versus competing baselines on the M3SO and YouTube Painting test sets.

(a) M3SO

(b) YouTube Painting

Figure 4: Generation results by our method on M3SO and YouTube Painting. The red square indicates the region of dominant motion. High scaled uncertainty ($\frac{1}{m_t}$) is in red.

tone as well, making the dataset multimodal. The dataset has 8,000 training, 1,000 validation, and 1,000 test videos. During training, algorithms are shown the first 30 frames of a video and are required to predict the next 30, while test time prediction is performed over 40 unseen frames.

YouTube Painting: This is a challenging, publicly available dataset [2], containing 64 64 YouTube videos of a painter painting on a canvas in an indoor environment, with the audio channel having sounds of the painter's brush strokes. There are 4.8K videos are available for training, 500 for validation, and 500 for test. Here, the training regime permits 15 seen and 15 unseen frames, while the test requires a prediction of 30 unseen frames.

Baselines and Evaluation Metrics: We compare our method against both multimodal ([2, 8]) and unimodal ([4, 5]) baselines, using standard image quality assessment metrics: (i) Peak Signal to Noise Ratio (PSNR) and (ii) Learned Perceptual Image Patch Similarity (LPIPS) [10].

Results: Figure 3 presents a quantitative evaluation of PUQ versus competing baselines on M3SO and YouTube Painting over different training epochs, as measured by PSNR and LPIPS. The plots reveal the superiority of our approach over competing baselines, consistently across both measures, especially in the early epochs. Moreover, we see that the multimodal baselines outperform their unimodal counterparts, in terms of generation quality. However, failure to incorporate the uncertainty associated with predicting stochastic frames, poses an insurmountable challenge even for multimodal methods. Qualitative generation results, as shown in Figures 1,4, concur with these observations as well. Furthermore, the scaled uncertainty ($\frac{1}{m_t}$), shown in the figures is seen to correlate well with the predictability of the direction of the brush stroke or the direction of digit motion. Additional results are in the supplementary. Further, we also evaluated a randomly selected subset of the generated videos in order to subjectively assess the video generation quality by conducting human preference evaluation. Annotators were presented with generated results by PUQ and its closest competitor Sound2Sight [2] and asked which one resembled the ground-truth more closely. Table 1 clearly shows that the annotators overwhelmingly preferred generations by our method, over 67% of the time.

6. Conclusions

In this work we introduce a novel technique for generating videos from audio and a few context frames – PUQ, which explicitly incorporates the frame prediction uncertainty into the learning objective. Empirical evaluations on two challenging audio-visual datasets, show that using PUQ results in faster training of the generative model while also outperforming competing approaches.

References

[1] M. Babaeizadeh *et al*. Stochastic variational video prediction. In *ICLR*, 2018. 2

[2] M. Chatterjee and A. Cherian. Sound2sight: Generating visual dynamics from sound and context. In *ECCV*, 2020. 1, 2, 3, 4

[3] M. Chatterjee *et al*. Hierarchical variational neural uncertainty model for stochastic video prediction. In *ICCV*, 2021. 2

[4] E. Denton and R. Fergus. Stochastic video generation with a learned prior. In *ICML*, 2018. 2, 4

[5] B. Jin *et al*. Exploring spatial-temporal multi-frequency analysis for high-fidelity and temporal-consistency video prediction. In *CVPR*, 2020. 1, 2, 4

[6] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *ICLR*, 2014. 2, 3

[7] Y. LeCun *et al*. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998. 3

[8] K. Vougioukas *et al*. Realistic speech-driven facial animation with gans. *IJCV*, 2019. 1, 2, 4

[9] W. Wang *et al*. Speech driven talking head generation via attentional landmarks. In *INTERSPEECH*, 2020. 2

[10] R. Zhang *et al*. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 4