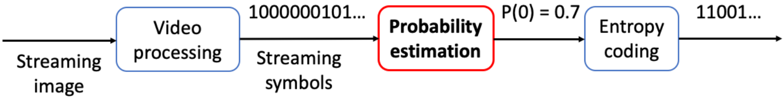
Online Probability Model Estimation For Video Compression

Yue Sun [†‡] Jingning Han[‡] Yaowu Xu[‡]

[†] University of Washington, Seattle
[‡] Google Inc.

March 2020

Probability estimation in video coding system



The minimum codelength coincides with the entropy

$$H(s_1, s_2, ..., s_t) = -\sum_{i=1}^{t} p^*(s_i|s_{i-1}...s_1) \log_2 p^*(s_i|s_{i-1}...s_1).$$

A good estimation will give us a small code length.

Baselines

- Suppose $p^*(s_t|s_{t-1}...s_1) = p^*(s_i)$, i.e., the symbols are i.i.d. At time t, symbol 0 appears k times, then the estimation[1] $p(0) = \frac{k+1}{t+2}$.
- CABAC[2] and AV1[3] use the update rule

$$
\begin{aligned}
p(1)_t &= ap(1)_{t-1} + (1-a)s_t \\
&= (1-a)(s_t + as_{t-1} + a^2 s_{t-2} + ... + a^{t-1}s_1) + a^t p(1)_0.
\end{aligned}
$$

[1] Suppose the prior of 0 and 1 are both 0.5.
[2] Wiegand et al., Overview of the h. 264/avc video coding standard.
[3] Chen et al., An overview of core coding tools in the av1 video codec.

Principle of new algorithms

1. Better entropy.
2. Robustness to noise.
3. Computational efficiency for real time application.
4. * Adaptivity with respect to data.

Due to the interest of 2. and 3., we do not use complex models such as neural networks or combinatorial methods.

Warmup: Second order system[4]

If the update rule is

$$\begin{aligned} q^{+} &= aq + (1-a)u, \\ r^{+} &= br + (1-b)u, \\ p &= wq + (1-w)r, \;\; w \in (0,1). \end{aligned}$$

then the probability estimation update is the following second order linear system

$$\begin{aligned} p_{t+1} = (a+b)p_t - abp_{t-1} \\ + (w(1-a) + (1-w)(1-b))u_t + (ab - (1-w)a - wb)u_{t-1}. \end{aligned}$$

[4] Alshin et al., High precision probability estimation for cabac," in 2013 Visual Communications and Image Processing (VCIP).

Idea: Fixed or adaptive aggregation of algorithms

We can average more than 2 update rules.

- Many algorithms work, for example, $p_+ = ap + (1-a)u$ with different a all roughly do the job.
- Denote $p \in \mathbb{R}^{n_p \times 2}$, each row of p corresponds to a "good" update rule. Denote coefficients $w \in \mathbb{R}^{n_p}$, such that $w \geq 0$, $\mathbf{1}^T w = 1$.
- Then we use $w^T p$ – weighted average of baseline estimations as final estimation.

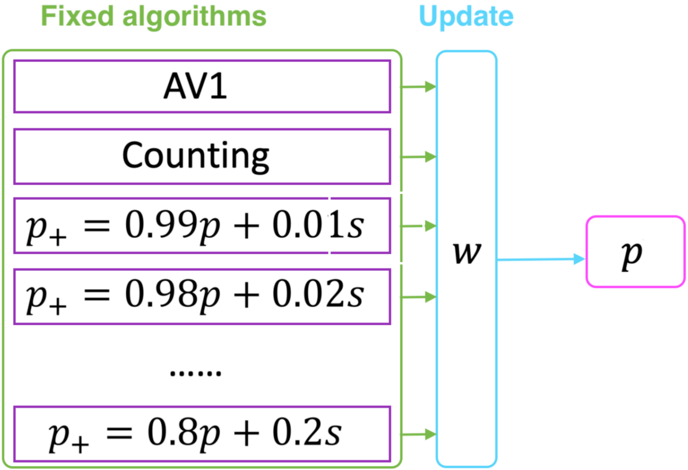
Idea: Fixed or adaptive aggregation of algorithms

We can average more than 2 update rules.

$$
\begin{aligned}
p_i^+ &= ap_i + (1-a)u, \\
p &= \sum_i w_i p_i.
\end{aligned}
$$

The problem is that we do not know w. We can either pick a fixed w, or even update w online.

Adaptive aggregation of algorithms



Adaptive aggregation of algorithms

For each symbol s_t, we incur the entropy

$$f(w, p; s_t) = -\log_2((w^T p)(s_t)),$$

and we take gradient with respect to w,

$$\nabla_w f(w, p; s_t) = -\frac{1}{(w^T p)(s_t)} p(:, s_t).$$

At each time, we run a gradient step[5]

$$\begin{aligned} w \leftarrow & \operatorname*{argmin}_{w_+ > 0, \mathbf{1}^T w_+ = 1} \|w_+ - w\|_2^2 + 2\eta_t (w_+ - w)^T \nabla_w f(w, p; s_t) \\ = & \operatorname*{argmin}_{w_+ > 0, \mathbf{1}^T w_+ = 1} \|w_+ - (w - \eta_t \nabla_w f(w, p; s_t))\|^2. \end{aligned}$$

$$\eta_t = 1, 1/2, 1/3, 1/4, 1/5, 1/6, 1/7, \ldots$$

$$\eta_t = 1, 1/4, 1/4, 1/4, 1/4, 1/9, 1/9, \ldots$$

[5]We can do a batch algorithm, where we take average of gradients with batch size $1, 4, 9, \ldots$ and update with fixed step size at the end of batches. $O(t^{1/3})$ updates until time t.

Fast projection

The dual of

$$\min_{x \geq 0, \mathbf{1}^T x = 1} \frac{1}{2} \|x - y\|^2.$$

is

$$\max_{\mu} \frac{1}{2} \| \max(y - \mu \mathbf{1}, \mathbf{0}) - y\|^2 + \mu(\mathbf{1}^T \max(y - \mu \mathbf{1}, \mathbf{0}) - 1),$$

and correspondingly,

$$x_i = \max(y_i - \mu, 0).$$

1 dimensional convex optimization, solve by binary search. To reach ϵ accuracy for μ^*, need $-\log_2(\epsilon)$ function evaluations.

Experiments

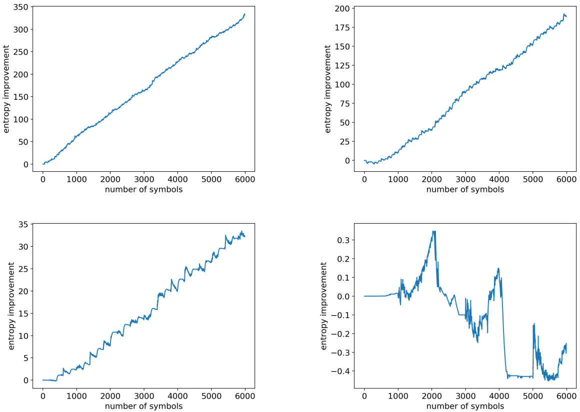


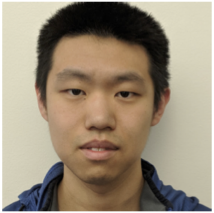
Figure 1: Codeword length reduction on synthetic data. Generate symbols with Bernoulli 0.01,0.3 alternatively with chunk size $50, 100, 200, 1000$. We plot the improvement of entropy of the proposed Multimodal SGD as compared to CABAC.

Experiments

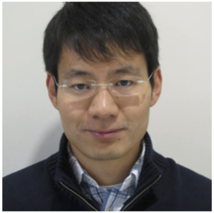
algorithm/dataset	200	400	800	1200	2000	2800	3600	5200
Multimodal Fixed	1368	2587	2940	3860	3770	3709	3465	2388
Multimodal SGD	**1364**	**2571**	**2930**	**3822**	**3733**	**3671**	**3433**	2363
Multimodal Batch	1375	2577	**2930**	3827	3734	3673	3437	**2358**
CABAC	1375	2592	2951	3873	3789	3727	3476	2401
AV1	1382	2580	2939	3843	3760	3698	3455	2380

algorithm/dataset	200	400	800	1200	2000	2800	3600	5200
Multimodal Fixed	215	267	477	726	623	472	322	**184**
Multimodal SGD	214	**265**	474	719	613	**468**	320	185
Multimodal Batch	**213**	**265**	**470**	**717**	**612**	469	**319**	186
CABAC	217	269	481	732	627	473	323	186
AV1	215	267	475	726	619	472	321	**184**

Figure 2: Cheer and harbour test clips in CIF. Entropy of probability estimation algorithms.



Yue Sun



Jingning Han

Yaowu Xu

Thank you for listening!