Source Coding

Joe Louis Paul. I

Outline

- Coding Theory
- Two Aspects of Coding Theory
- Source Encoding Vs Channel Encoding
- Need of Source Coding
- Source Codes
- Source Encoder
- Types of Source Coding
- Fixed Length and Variable Length Codes
- Development of an Efficient Source Encoder
- A Typical Source Coding Scheme
- Average Codeword Length
- Coding Efficiency
- Shannon’s Source Coding Theorem
- Determining the Value of L_{min}
- Coding Efficiency using H(S)

Coding Theory

- The purpose is to design an efficient and reliable data transmission methods.
- This typically involves the removal of redundancy and the correction (or detection) of errors in the transmitted data.
- It also includes the study of the properties of codes and their fitness for a specific application.

Two Aspects of Coding Theory

- **Data compression** (or, *source coding*)
- **Error correction** (or, *channel coding*)

Source Encoding Vs Channel Encoding

- *Source encoding*, attempts to compress the data from a source in order to transmit it more efficiently.
  - This practice is found every day on the Internet where the common "Zip" data compression is used to reduce the network load and make files smaller.
- The second, *channel encoding*, adds extra data bits to make the transmission of data more robust to disturbances present on the transmission channel.

Need of Source Coding

- The important problem in communication is that **<u>how the efficient representation of data produced by discrete source can be achieved?</u>**
- The process by which this representation is achieved called as **source coding** (i.e. **representation of symbols in binary**).
- *<u>The aim of source coding is to take the source data and make it smaller.</u>*

Source Codes

- Basically source codes *try to reduce the redundancy present in the source, and represent the source with fewer bits that carry more information.*

Source Encoder

- The **device or an element which involves with source coding in a communication system**.

Types of Source Coding

- **Fixed Length Coding (FLC)**
- **Variable Length Coding (VLC)**

Fixed Length Coding (FLC)

- Modern codes
  - used for the interchange and processing of information
  - encode each character by one or more octets, an octet being a sequence of 8 bits.
- *A code is fixed length* if each character of the code is represented by the same number of octets.

Example: Fixed Length Coding

- ASCII
  - One of the most well known coded character sets is ASCII, the *American Standard Code for Information Interchange*.
  - This represents *128* characters and uses all possible combinations of 7 bits.

Table I: 8-bit Extended ASCII

65	01000001	A	Uppercase A
66	01000010	B	Uppercase B
67	01000011	C	Uppercase C
68	01000100	D	Uppercase D
69	01000101	E	Uppercase E
70	01000110	F	Uppercase F
71	01000111	G	Uppercase G
72	01001000	H	Uppercase H
73	01001001	I	Uppercase I
74	01001010	J	Uppercase J
75	01001011	K	Uppercase K
76	01001100	L	Uppercase L
77	01001101	M	Uppercase M
78	01001110	N	Uppercase N
79	01001111	O	Uppercase O
80	01010000	P	Uppercase P
81	01010001	Q	Uppercase Q
82	01010010	R	Uppercase R
83	01010011	S	Uppercase S
84	01010100	T	Uppercase T
85	01010101	U	Uppercase U
86	01010110	V	Uppercase V
87	01010111	W	Uppercase W
88	01011000	X	Uppercase X
89	01011001	Y	Uppercase Y
90	01011010	Z	Uppercase Z

Variable Length Coding (VLC)

- In *coding theory* a **variable-length code** is a code which maps source symbols to a *variable* number of bits.

Example: Variable Length Coding

- Morse Code
  - Letters of the alphabet and numbers are encoded into sequence of marks and spaces denoted as "." and "-" respectively.
  - Example
    - Letter E denoted as "."
    - Letter Q denoted as "--.-"

Table II: Morse Code-A Variable Length Code

A	.-	N	-.	1	.----	.	.-.-.-
B	-...	O	---	2	..---	,	--..--
C	-.-.	P	.--.	3	...--	?	..--..
D	-..	Q	--.-	4	....-	(	-.--.
E	.	R	.-.	5	.....	)	-.--.-
F	..-.	S	...	6	-....	-	-....-
G	--.	T	-	7	--...	"	.-..-.
H	....	U	..-	8	---..	_	..--.-
I	..	V	...-	9	----.	'	.----.
J	.---	W	.--	0	-----	:	---...
K	-.-	X	-..-	/	-..-.	;	-.-.-.
L	.-..	Y	-.--	+	.-.-.	$	...-..-
M	--	Z	--..	=	-...-		

Fixed Length and Variable Length Codes (1/3)

- A **fixed length code**, or **block code** is a code whose codewords all have the same length n.
- In this case, the number n is also called as the *length of the code*.
- If a code C contains codewords of varying lengths then it is called as a *variable length code*.

Fixed Length and Variable Length Codes (2/3)

- Fixed length codes have advantages and disadvantages over variable length codes.
- One advantage is that <u>**they never require a special symbol to separate the characters in the message being coded**</u>.
- For example, consider the ***encoded ASCII message***,

  010000110100111101000100010000101

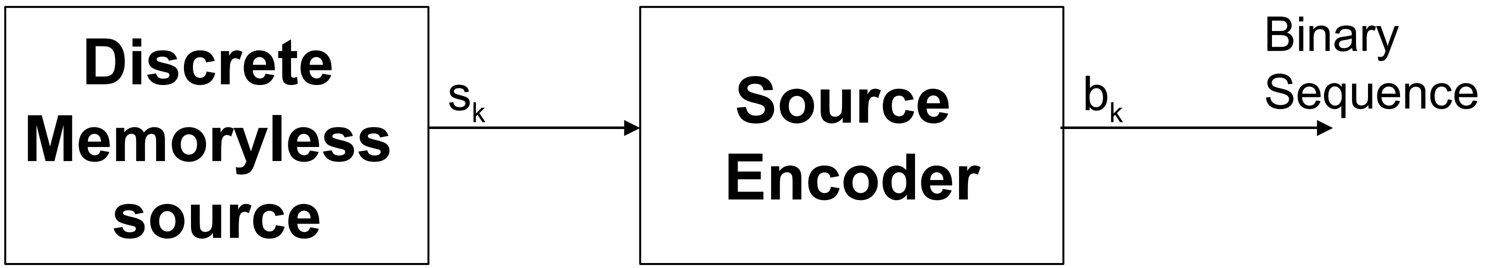
Fixed Length and Variable Length Codes (3/3)

- Binary (ASCII) code is a fixed length code whose codewords have length 8, we know that the first 8 bit represents the first character of the original message. Similarly the second set of 8 bits represents the second character in that message.
- Perhaps, the main disadvantage of fixed length codes, such as the ASCII code, is that <u>*characters that are used frequently such as the letter e, have codes as long as characters that are used in frequently, such as the space character*</u>.
- On the other hand, variable length codes, which can encode frequently used characters using <u>shorter codewords</u> can save a great deal of space and time.

Development of an Efficient Source Encoder

- The *statistics of the source is required*.
- **Two Fundamental Requirements**
  - The codewords produced by the source encoder are in binary form.
  - The source coding scheme should be *uniquely decodable* so that the original source sequence can be reconstructed exactly from the encoded binary sequence.

A Typical Source Coding Scheme



- Let us assume that
  - Source sequence s_k
  - Binary sequence b_k
  - Length of the binary sequence l_k

Source Coding

- Consider a source that has an alphabet of **K** different symbols such that $\mathbf{s_k}$ is the $\mathbf{k^{th}}$ symbol in the alphabet.
- Also let $\mathbf{s_k}$ occurs with probability $\mathbf{p_k}$ and let the binary codeword assigned to $\mathbf{s_k}$ by the encoder have length l_k measured in bits.

Average Codeword Length $\overline{L}$

- The average codeword length of the source encoder is defined as

$$\overline{L} = \sum_{k=0}^{K} p_k l_k$$

- The unit of average codeword length is **bits per symbol**.

Coding Efficiency **η**

- $\bar{L}$ represents the average number of bits per source symbol used in the source encoding process.

- If $L_{\min}$ denotes the minimum possible codeword length then, the coding efficiency of the source encoder is defined as

$$\eta = \frac{L_{\min}}{\bar{L}}$$

- The coding efficiency will be equal to one if $\bar{L} = L_{\min}$ i.e. efficient source input.

Shannon's Source Coding Theorem

- With $\overline{L} \geq L_{\min}$, then $\eta \leq 1$.
- The source encoder is said to be efficient when η approaches unity.
- **Source coding Theorem**
  - Given a discrete memoryless source of entropy H(S), the average code-word $\overline{L}$ length for any source encoding is bounded as

$$\overline{L} \geq H(S)$$

Determining the Value of L_{min}

- From Coding efficiency formula,

$$\bar{L} \geq L_{\min}$$

- From Shannon's Source Coding Theorem,

$$\bar{L} \geq H(S)$$

- From these above equations,

$$L_{\min} = H(S)$$

Coding Efficiency using H(S)

- If $H(S)$ denotes the minimum possible codeword length then, the coding efficiency of the source encoder is defined as

$$\eta = \frac{H(S)}{\overline{L}}$$