Contents

Computer Science Fundamentals

Unit 1: Digital Information

Date: June 29, 2024

Lossless Text Compression

Compression Algorithm: Finding repeated sequence and replace them with shorter representation.

E.g.

“to be or not to be, that is the question”

We can rewrite the above sentence, by replace repeated words with shorted representation. First identify the repeated words and make representation of it and **created table of replacements**.

Replacement	**Original**
!	To
@	Be

So, the above sentence can be represented such as:

! @ or not ! @, that is the question.

We can also replace part certain characters of a word, as we can see “th” is repeated, so lets replace it with #, we have updated **table of replacements.**

Replacement	**Original**
!	To
@	Be
#	th

And our sentence is now:

! @ or not ! @, #at is #e question.

Exercise: Try replacing it.

I am Sam, Sam I am. That Sam-I-am! That Sam-I-am! I do not like that Sam-I-am! Do you like green eggs and ham? I do not like them, Sam-I-am. I do not like green eggs and ham.

Solution: Lets first identify repeated word and characters and make a table of replacements:

Replacement	**Original**
~	I am

@	Sam-I-am
#	I do not like
%	That
&	Sam
*	eggs
(	ham
)	green

Compressed version of the above text:

~ &, & ~. % @! % @! # % @! Do you like) * and (? # them, @. #) * and (.

Note: Some text can't be compressed at all. For example, a text which has no repeated words or character pattern.

Date: June 30, 2024

Lossless Image Compression

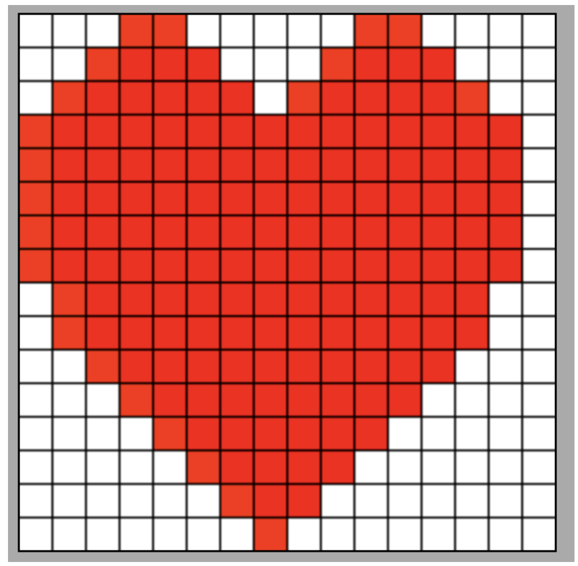
RLE (Run-Length Encoding):

RLE is used for Simplest images.

Bitmaps

Bitmap represents image into binary.

For example: Here we have an icon, which shows heart in white and red color. So, we can represent red with 1's and white with 0's and create a 1's and 0's grid. Hence each pixel either 1 or 0, it is called bitmap. Since it is mapping pixels to bits.



So, we can represent heart icon like this:

```
0001100000110000
0011110001111000
0111111011111100
1111111111111110
1111111111111110
1111111111111110
1111111111111110
1111111111111110
0111111111111100
0111111111111100
0011111111111000
0001111111110000
0000111111100000
0000011111000000
0000001110000000
0000000100000000
```

White using run length encoding, we first assume 0's and then 1's, for example

Actual Bit Representation	Bit Counts
0001100000110000	3,2,5,2,5
0011110001111000	2,4,3,4,3
0111111011111100	1,6,1,6,2
1111111111111110	0,15,1
1111111111111110	0,15,1
1111111111111110	0,15,1
1111111111111110	0,15,1
1111111111111110	0,15,1
0111111111111100	1,13,2
0111111111111100	1,13,2
0011111111111000	2,11,3
0001111111110000	3,9,4
0000111111100000	4,7,5
0000011111000000	5,5,6

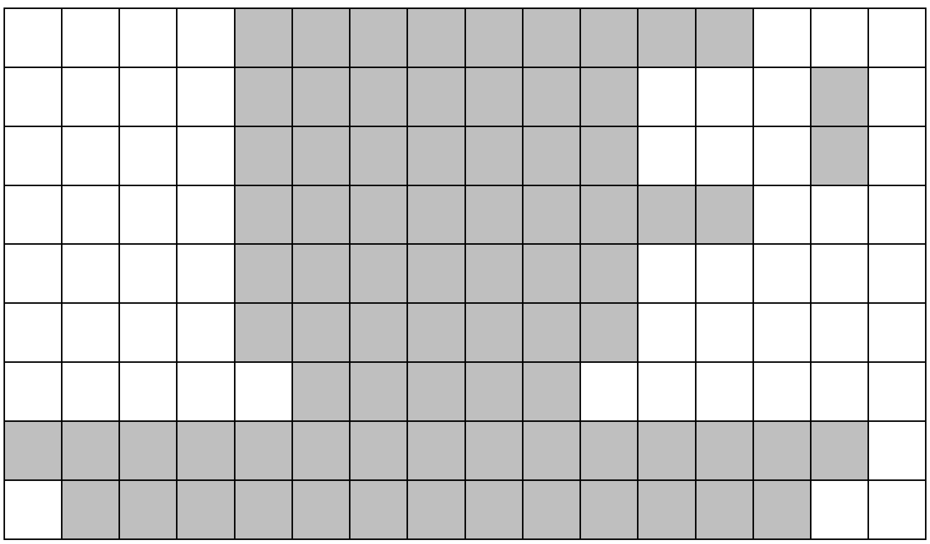
0000001110000000	6,3,7
0000000100000000	7,1,8

As we compressed the image so, should we be able to decompress it as well:

E.g. Decompress the below RLE

4, 9, 3
4, 7, 2, 1, 2
4, 7, 2, 1, 2
4, 9, 3
4, 7, 5
4, 7, 5
5, 5, 6
0, 15, 1
1, 13, 2

Solution:

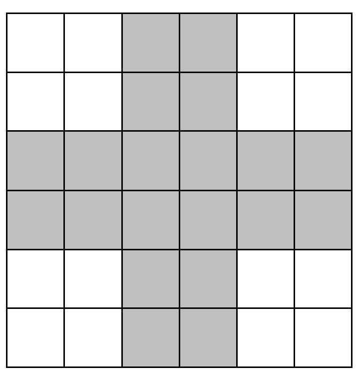


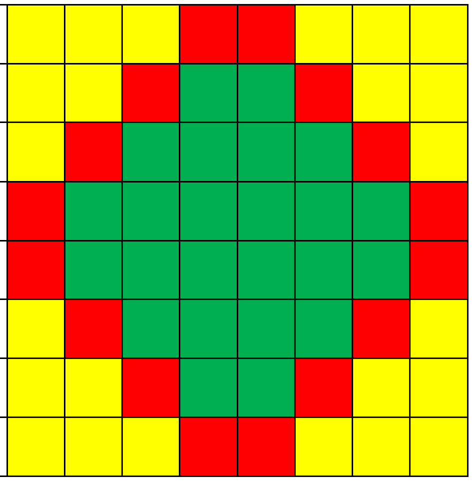
Exercise:

2,2,2
2,2,2
0,6
0,6
2,2,2
2,2,2

Solution:

It shows a plus sign icon.



3Y,2R,3Y	
2Y,1R,2G,1R,2Y	
1Y,1R,4G,1R,1Y	
1R,6G,1R	
1R,6G,1R	
1Y,1R,4G,1R,1Y	
2Y,1R,2G,1R,2Y	
3Y,2R,3Y	

Compression Ratio

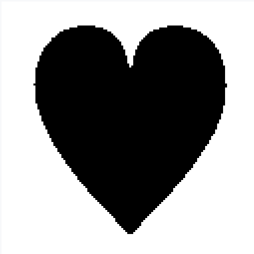
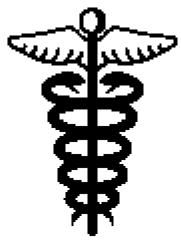
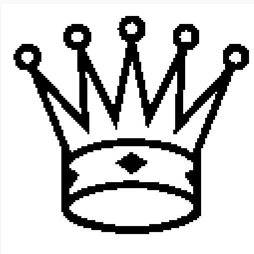
We've claimed that run-length encoding can save us space when storing simple images—but *how much* space?

To find out, I wrote a program to encode black & white bitmaps with RLE. This table summarizes the results on three heart icons of increasing size:

Image	Dimensions	Uncompressed	After RLE	Space savings
	16x16	256	228	10.9%
	32x32	1024	532	48.0%
	128x128	16384	2898	82.3%

Take a look at that final column, the space savings. Notice a pattern? We save much more space as the size increases since the runs are much longer.

What about images of the same size? This table summarizes the results of compressing three large icons with RLE:

Image	Dimensions	Uncompressed	After RLE	Space savings
	128x128	16384	2898	82.3%
	128x128	16384	8298	49.4%
	128x128	16384	8730	46.7%

Now you can see why I picked a heart icon as my example: it compresses very well, thanks to its many runs of black or white. RLE compression still halves the size of the other icons, but it doesn't save nearly as much space.

In fact, sometimes RLE can't save any space at all...

RLE Limits

Now images are way more complex than old times, now each may contain not same but similar color, which does not allow us to use RLE.

RLE Uses

RLE is used to compress fax documents, which are since they are only black and white dots.

Conclusion: So basically, RLE works with images with limited color pallets, but images with continuously changing color or similar color such as photographic images require more complex method of complex, correct me if I am wrong.

Lossless Bit Compression

It is used to reduce the file size without losing any information, every single bit must be preserved when decompressed.

Huffman Coding Algorithm

It is a compression algorithm, that assigns shorter code (binary code) to more frequent symbols and longer code for rare ones.

How it works:

1. Count the frequency of each symbol (Bit patterns or byte)
2. Build a binary tree by combining repeatedly combining the two least frequent nodes.
3. Assign a shorter code to more frequent symbols and longer code for rare ones.

Example: for text "This is an example…" a Huffman tree produces code like e:000, t:0110, space:111, reducing the total bits from 288 to 135.

Date: July 1, 2024

Lossy Compression

Lossy data compression reduces file(s) size by discarding less important information. In very clever ways to remove detail without human noticing (too much).

Let's explore how to do it:

Images: Keep the brightness, average the color

Human eye is good at differentiating between different brightness, than color itself. So, computer uses it to its advantage, by keeping the brightness the same, while reducing the amount of color information, a process known as **chroma subsampling**.

Steps to do it:

1. Separate the brightness information from the chroma(colors).
2. Zoom in blocks
3. Take average of 2x2 blocks and set that each block. There you have it, you have reduced 4 times the size already.

Audio: Drop the inaudible sounds

A human a limit to what it can hear, so we can analyze the file and discard or represent the inaudible sounds with fewer bits that are outside the hearing capacity, a process known as **perceptual audio coding**.

One limitation known as **temporal masking**. That's when a sudden sound can mask(hide) other sounds for a period after it occurs – or even a bit before. Computer can see those hidden sounds but human can hardly hear it.

Note: It was just one limitation and human being a many other limitations, which can be used to reduce the file size if possible.

Compression Quality

Using lossy compression, we lose certain percentage (%) of data, may be 10%, 20, or 70%, etc. It all depends on use case.

Date: July 5, 2024

Copyright, DRM, and DMCA

Copyright

Copyright introduced in 1700's in England, due to unregulated use of printed press.

DRM (Digital Rights Management):

it was introduced to protect owner work right, by mediating the sales process, which make sures only authorized user has the access to the product.

DMCA (Digital Millennium Copyright Act):

Introduced in 1998, in united states, to act again circumvent of DRMS.

Creative Commons and Open Source

Creative Common

Creative common is a non-profit organization that offers 6 licenses for sharing creative work.

Date: July 7, 2024

The Internet

What is Internet?

Internet consist of incredibly large collection of independently operated networks. It is fully distribution, has no central control.

From what I know, it starts with connecting devices to share data and do communication like Bluetooth is used to share data when devices are closer.
As distance increase new names of networks are creates, like network connecting device in building or local devices is called "Local Area Network" and go higher if multiple such building or towns are to communicate with each other "Metropolitan Area Network" was created, and if you go even higher where whole each can communicate that network is named Internet.
So basically, Internet is a network that connect, communicates, and share data among devices at small and large distances.

- Internet is an open network.
- Protocols are rules that define how devices can communicate with each other. The internet is powered by many layers of protocols:
  - **Wires & wireless:** Physical connection between devices, plus protocols for converting electromagnetic signal into binary data.
    - **Wired Internet** uses physical cables for connection, while **Wireless Internet** transmit data through radio waves.
  - **IP:** Internet Protocol is a unique address that identifies a device on a network, allowing it to send and receive data.
  - **TCP/UDP (Transmission Control Protocol/User Datagram Protocol):** Protocols that can transport packets of data from device to another and checking errors along the way.
  - **TLS (Transport Layer Security):** A secure protocol for sending encrypted data so that attackers can’t view private information.
  - **HTTP & DNS:** The protocols powering the World Wide Web, what browser uses every time you load a webpage.

Date: July 8, 2024

Computer Networks

- Internet is world's largest network.

Types of Networks:

It is defined by the size and characteristics of a network.

Local Area Network (LAN):

A network that covers limited area like a house or school, has a cable to connectivity called backbone.

Wide Area Network (WAN):

It is the largest of the network that extends over a large geographic area, and is composed of many, many LANS.

Network Protocols:

Devices need protocols to communicate, protocols defined common format to exchange data to avoid any conflicts.

Physical Network Connection

Coper Cable

Above is CAT5, twisted pair cable, that is designed for use in computer networks.

It follows Ethernet Standards for data transmission, that is why it is also called **Ethernet cable**.

Fiber Optic Cable

It contains optic fiber that can carry light (instead of electricity). Fiber is coated with plastic layers and sheathed in a protective tube to protect it from the environment.

It typically follows Ethernet Standards to make sure that communication is doable.

Fiber Optic Cable can send data at higher data per second than coper cables.

Wireless

- It doesn't involve wires at all.

- A wireless card inside the computer turns binary data into radio waves and transmits them through the air:
- It can't travel very far.
    - In a closed space 70 – 100 feet.
    - In an opened space up to 1000 feet.

Date: July 9, 2024

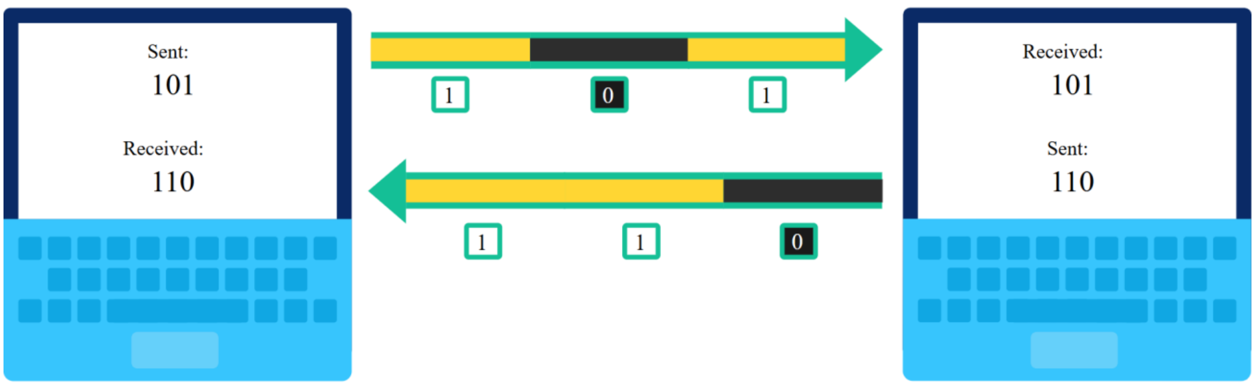
Bit rate, Bandwidth, and Latency

Internally computers can use as many wires required to represent the number. But that is not the case when send data to another computer. In fact, there is one wire when sending data.

What is the solution?

Basically, the stream of bits can be sent over time period. Where time period is equal to number of bits to represent the data.

Let's say we have 5(101 in binary) data stored in computer. When sending data, we may send 1,0,1 series vise and is received by the receiver series vise. Like shown below:



Note:

- In electrical connection (such as Ethernet), signal would be voltage or current.
- In optical connection (such as fiber-optic-cable), the signal would be the intensity of the light.
- **Line coding** is the process of turning binary data into a time-based signal.

Bit Rate

The number of bits of data that are sent each second also known as Bits Per Seconds(bps).

Here below are some common units to measure **Bit Rate**.

UNIT	NUMBER OF BITS	
KILOBIT	1000	1 thousand
MEGABIT	1000^2	1 million
GIGABIT	1000^3	1 billion
TERABIT	1000^4	1 trillion
PETABIT	1000^5	1 quadrillion

Bandwidth

It is used to describe maximum bit rate of the system.

For example, is a network has 100Mbps, it means it can't transfer data more than 100Mbps.

Latency

It is another way to measure the speed of network. It measures how late the bit arrives. Basically, time between sending of data and receiving of data, measured in milliseconds.

Typically, "round-trip" latency is measured. For example:

We are sending www.google.com as request to the servers which takes about 30ms, and our computer receives acknowledgement after 40ms. That is total round-trip 30ms + 40ms = 70ms.

Congestion in network mean my request has to wait in line for its turn.

We can decrease latency speed by resolving congestion and improving physical connection.

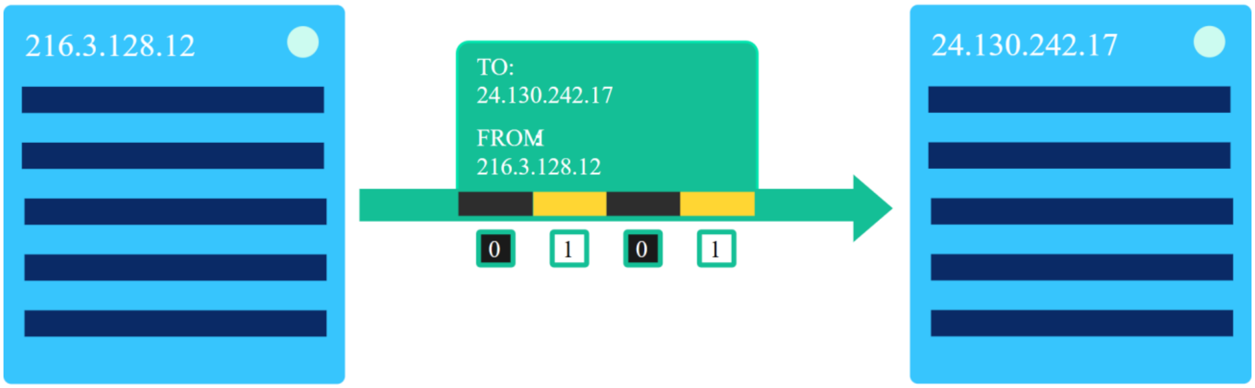
Internet Speed

- Speed is combination of Bandwidth and Latency (Bandwidth + Latency).
- Messages (data) are split into packets.
- Packets are sent in order; second packet can't go before the first packet.

Lesson 3: Addressing the Internet

IP Addresses

- Internet Protocol (IP) is used to handle both addressing and routing.
- IP address uniquely identifies the Internet connected device.



IPv4 Addresses

There are actually two versions of Internet Protocol:

- IPv4: the first version ever used on the Internet.
  - In IPv4 Protocol, IP Address looks like this:
  - 74.125.20.103
  - We write these numbers in decimals, but computers store them in binary.
  - Each IP address is split into 4 numbers, and each number range from 0 to 255.
  - Each number can be describe using 8bits, that is why it is also called “octets”.
  - Overall, that’s 2^{32} possible values: 4,249, 967,296
  - But there are more than 4 billion devices available now, this IP Address is not sufficient.
- IPv6: a backwards-compatible successor.
  - 2001:0db8:0000:0042:8a2e:0370:7334
  - This IP Addressing Protocol is addressed using Hexa-decimal numbers.
  - There are 8 hexa-decimal numbers
  - Each number is 4 digits longer
  - The highest value in a number is FFFF with equals to 65,535 in decimals.
  - F need 4 bits to be represented so FFFF need 16 bits to be represented.
  - Hence 16 * 8 = 128 bits are used in IPv4 to represent any IPv6 address. That is 340 undecillions.
  - Computer using WiFi or Ethernet at home or schools has dynamic IP Address, meaning each a computer connect to the Internet it is assigned a different IP address based on IP Address ranges that is provides by Internet Service Providers (ISPs).
  - But computers that act as servers, often have static IP addresses.

IP Address Hierarchy

Both IPv4 and IPv6 are hierarchical, for now let's examine IPv4, and consider IPv4 address: 24.147.242.217

The first sequence of bits (first two octets) identifies the network: 24.147

And final identifies the individual node in the network: 242.217

Subnets

Network administrator can further break IP addresses into subnetworks (subnets) as needed.

Starting with this IP address:

141.213.127.13

That could be break into 3 parts:

141.213	127	13
UMich network	Medicine Department	Lab Computer

Splitting Octets

In actuality, IP addresses are often split in the middle of the octets.

For example: 141.213.127.13 and represent it in binary

141	213	127	13
10001101	11010101	01111111	00001101

All together that translate into binary:

10001101110101010111111100001101

The first 16 bits could route to all of UMich, the next 2 bits could route to a specifi UMich Department, and final 14 bits could route to individual computers. Like shown below:

1000110111010101	01	11111100001101
UMich network	Medicine Department	Lab Computer

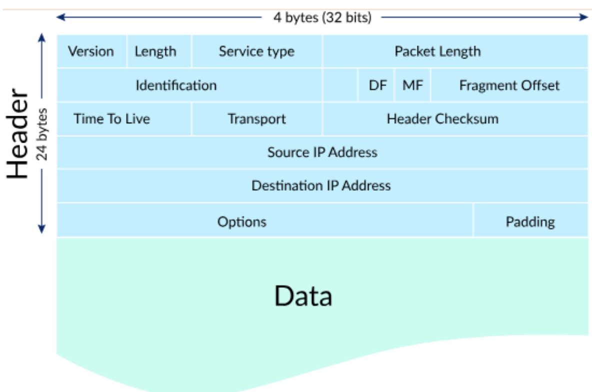
This hierarchy give UMich give capability to differentiate between 2^2 departments and 2^{14} Computers in each department.

Date: July 10, 2024

Lesson 4: Routing with redundancy

IP Packets

- Ping is a tiny message used to check if another device is online.
- As there is limit to how much data can be transmitted on network.
- That is why data is split into multiple small packets and IP describes the structure of packet what whizz around the Internet.
- Each IP Packet contains both a header (20 or 24 bytes long) and data (variable length).
- The **header** contains the IP addresses of both source and destination, plus other fields that help to route the packet.
- The **data** is actual content such as string, letter, or part of webpage.



Internet Routing Protocol

- The Internet Protocol (IP) describe how to route messages from one computer to another computer on the network.
- Each message is split up into packets, and the packets hop from router to router on the way to their destination.
- Note: **Router** is a device that help used in computer networks that help move the packets along.

Let's see in steps:

- Step 1: Send packet to router
    - Computer sends first packet to the nearest router.
- Step 2: Router receives the packet
    - When router receiver the packet, it looks at destination IP address, which tell router where packet has to go.
- Step 3: Router forwards the packet
    - Router has multiple paths it could send a packet along
    - But its goal is to send packet to the router that is closer to the destination.
    - How does it decide?
        - It has **forwarding table** that helps it pick the next path based on the destination IP Address.

    - That table does not have row for every possible IP Address; as there are 2^{32} possible IP Addresses, and that is far too much to store. Instead, table has rows for IP address prefixes.
    - IP Addresses are hierarchical, when two IP addresses have same prefix, that often means they are on the same large network.

IP address prefix	path
91.112	#1
91.198	#2
192.92	#3

- Step 4: Final router forwards the message
    - If everything goes well, final router will send the packet to its recipient.

Redundancy and Fault Tolerance

Redundancy

In routing, redundancy means having backup paths or extra routes so that data can still reach destination even if one route fails.

Fault Tolerance

Fault tolerance in routing means a network can still send data from one place to another even if one or more routers, link or path fail.

Note: we try to find single points of failure and find ways to add redundancy at those points.